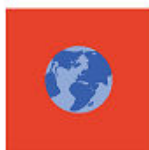


Études de démonstration de l'intérêt de la géostatistique dans le domaine des sites et sols pollués

- Rapport final -

ADEME



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

Table des matières

TABLE DES MATIERES.....	2
PREAMBULE.....	3
CONTRIBUTIONS A LA REALISATION DES ETUDES DE DEMONSTRATION.....	5
OBJECTIFS DES ETUDES DE DEMONSTRATION	6
RESUME DES TRAVAUX REALISES ET DES RESULTATS OBTENUS.....	7
ÉTUDE GEOSTATISTIQUE DE LA CONTAMINATION D’UN DEPOT D’HYDROCARBURES : LE SITE « TOTAL ».....	11
ÉTUDE GEOSTATISTIQUE D’UN ANCIEN SITE DE STOCKAGE DE PRODUITS CHIMIQUES : LE SITE 1BIS ATTENA.	45
ÉTUDE GEOSTATISTIQUE DE LA CONTAMINATION D’UNE FRICHE INDUSTRIELLE : LE SITE T.	127

Préambule

Le groupe de travail GeoSiPol a été mis en place fin 2003 afin de promouvoir l'application de la géostatistique lors de l'évaluation et de la réhabilitation des sites et sols pollués. Outre les membres fondateurs (membres experts) à l'origine du projet (KIDOVA¹, Geovariances, Mines ParisTech²), le groupe de travail rassemble aujourd'hui plus de 20 membres adhérents concernés par la gestion des sites et sols pollués : institutionnels (ADEME, INERIS), bureaux d'études et industriels.

Afin de permettre aux organismes publics et aux professionnels des sites pollués de mieux se positionner par rapport à ce type de démarche et de la proposer, une phase préliminaire a consisté à faire la revue, à partir de l'expérience acquise par les membres fondateurs, des applications possibles de la démarche géostatistique à l'étude des sites pollués.

Cette phase, rendue en partie possible grâce à une subvention de l'ADEME et à un soutien financier de l'INERIS, a donné lieu à la rédaction d'un manuel méthodologique : « Géostatistique appliquée aux sites et sols pollués – Manuel méthodologique et exemples d'applications », Version 1, novembre 2005, 139 p.

Suite à ces travaux méthodologiques et pédagogiques, il est ressorti une demande des membres de GeoSiPol pour que des travaux de mise en œuvre pratique de la géostatistique soient réalisés pour mieux en appréhender la portée. L'idée de réaliser des études de démonstration à partir de cas d'étude réels a été proposée et acceptée par l'ensemble des membres de GeoSiPol.

Les études de démonstration qui ont été réalisées ne sont cependant pas et ne doivent pas être considérées comme des exemples d'études géostatistiques. Si la finalité est semblable, en s'intéressant à l'estimation et à la délimitation de volumes de sols pollués et aux incertitudes associées, les travaux présentés ici sont plus variés que ceux réalisés habituellement dans les études et adressent des objectifs plus larges. Au-delà de l'analyse et de la modélisation géostatistiques menées habituellement, il s'agissait de :

- trouver des moyens pour confronter les modèles géostatistiques à des données de validation qui n'ont pas été définies ou choisies dans ce but,
- comprendre et expliciter les différences observées en les ramenant à des choix de modélisation fondés nécessairement sur les jeux de données disponibles pour construire ces modèles,
- dans la mesure du possible, démontrer comment des améliorations des modèles géostatistiques, qui auraient pu être anticipées en disposant initialement de données plus complètes ou mieux adaptées, peuvent corriger en partie au moins les écarts avec les données de validation.

La présentation des études de démonstration se veut également pédagogique en s'intéressant à l'importance des données dans le bon aboutissement des études et aux précautions à prendre pour les mener à bien.

Quatre sites ont finalement été retenus parce qu'ils reflètent la variété des situations de contamination rencontrées usuellement par les praticiens et devaient permettre une validation des modélisations géostatistiques (sites ayant fait l'objet d'une dépollution par excavation des sols ou pour lesquels des données de contrôle sont envisageables). Ces sites, désignés suivant les organismes qui les ont proposés, sont les suivants :

1. Site TOTAL France. Le site de Saint Herblain est un ancien dépôt pétrolier dont les sols et les eaux souterraines présentent des anomalies de teneurs en hydrocarbures. Le site a déjà fait l'objet de diagnostics, d'une évaluation des risques sanitaires et d'une étude technico-économique. Il est aujourd'hui dépollué. Un total de 80 sondages a été réalisé en diagnostic pour analyses chimiques. Les sols ont également été échantillonnés le long de 17 sondages avant dépollution selon un

¹ Anciennement FSS International r&d.

² Anciennement Ecole des Mines de Paris.

maillage de 15mx15m au droit des zones à risque pour analyse des hydrocarbures totaux et des BTEX.

2. Site Ibis ATTENA (Burgeap). Le site Ibis ATTENA est un ancien site de stockage, conditionnement et expédition de produits chimiques. Les sols y sont pollués par des COHV, des BTEX, des HCT et des composés organo-solubles (alcools, cétones). Une ESR a déjà été réalisée et un suivi annuel de la qualité de la nappe existe depuis 2001. Les données de pollution disponibles proviennent de sondages, piezair, piézomètres et mesures géophysiques. Ce site a préalablement été retenu pour l'étude de démonstration de l'évaluation de l'Atténuation Naturelle dans la gestion des pollutions par des organo-halogénés volatils dans le cadre du projet r&d ATTENA (2009-2012), cofinancé par l'ADEME. De nombreuses investigations complémentaires y sont réalisées. Elles incluent des sondages à 5 et 30 m avec des prélèvements de sols multi-niveaux, des piézomètres, des piezairs et des analyses en laboratoire sur les matrices sol, air et eau. L'accord a été donné à GeoSiPol par les participants à ATTENA pour utiliser ces données dans les études de démonstration.
3. Site T (ArcelorMittal Real Estate). Le site T est celui d'une ancienne cokerie dont les terrains présentent des HAP et des métaux. Plusieurs campagnes de reconnaissance et plusieurs études de la qualité des sols ont été réalisées depuis le début des années 1990. Les études les plus anciennes ont porté sur l'ensemble du site (≈15 ha) jusqu'à une profondeur de 5 m (profondeur de la nappe). Plus récemment, des EDR se sont intéressées aux premiers 30 cm de sols et à la reconnaissance d'une zone ciblée, jusqu'à 3 m de profondeur, pour la délimiter. De nombreuses données sont ainsi disponibles sur les sols. Elles incluent des analyses en laboratoire sur échantillons mais aussi des données géophysiques (profils de résistivité électrique).
4. Site Château (ICF Environnement). Le site de Château est une ancienne station service pour poids lourds. La pollution concerne essentiellement les sols et est constituée d'hydrocarbures. Les investigations du site ont été réalisées, ainsi qu'une étude technico-économique des alternatives de réhabilitation. 40 sondages ont été réalisés sur le site pour échantillonnage des sols et analyse des fractions d'hydrocarbures. La dépollution des sols par excavation était prévue pour la fin 2010 et avait été la raison première pour prolonger la durée d'exécution des études de démonstration de 9 mois. Cette dépollution n'a finalement pas été entreprise et ne le sera pas avant 2012. Sans disposer de données de dépollution, ce cas d'étude a finalement été abandonné après l'analyse des premières données disponibles. En contrepartie, des travaux complémentaires ont été réalisés sur le cas d'étude Ibis ATTENA fourni par Burgeap.

Les sites TOTAL France et Site T sont considérés comme des sites de grande taille du fait de leur extension géographique et du nombre de données disponibles. Les sites Ibis ATTENA et Château sont considérés comme des sites de petite taille.

Le site Ibis ATTENA est un cas particulier pour deux raisons : d'une part, la pollution y a été essentiellement étudiée par rapport aux eaux souterraines, d'autre part, il ne fera pas l'objet d'une dépollution avec excavation des sols. Ce site a cependant été retenu du fait du très grand nombre de données disponibles (nombreux échantillons de sols mais aussi d'eau et air analysés) et qui devaient bien se prêter à une étude de démonstration en exploitant certaines de ces données comme données de vérification de modèles géostatistiques établis à partir des autres données. Les objectifs de l'étude de démonstration menée sur le site Ibis ATTENA ont été définis avec BURGEAP. Il est à noter que l'évaluation de la distribution spatiale des solvants chlorés dans les aquifères reste une difficulté importante de la caractérisation et de la dépollution des sites pollués que les méthodes géostatistiques peuvent aider à lever. Ces aspects ne rentraient pas dans le cadre des études de démonstration et n'ont pas été abordés.

Ces études de démonstration bénéficient d'une subvention de l'ADEME dans le cadre d'une convention passée entre l'ADEME et GeoSiPol. Ce rapport final présente l'ensemble des travaux réalisés et des résultats obtenus.

Contributions à la réalisation des études de démonstration

Ce rapport final a été co-écrit par les membres experts du groupe de travail GeoSiPol qui ont réalisé les études de démonstration :

Jean-Paul Chilès (Mines ParisTech)

Claire Fauchoux (Geovariances)

Michel Garcia (KIDOVA)

Nicolas Jeannée (Geovariances)

Gaëlle Le Loc'h (Mines ParisTech)

Jean-Baptiste Mathieu (KIDOVA)

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

Thierry Blondel (Cabinet Blondel)

Olivier Bour (INERIS)

Véronique Croze (ICF Environnement)

Sébastien Kaskassian (Burgeap)

Jean-François Lafon (CSD INGENIEURS)

Jean-Michel Métivier (IRNS)

Hélène Roussel (ADEME)

Guillaume Suzanne (Antea Group)

Objectifs des études de démonstration

Les études de démonstration visent à démontrer les performances d'une approche géostatistique, par rapport à des approches plus classiques, sur plusieurs cas d'étude et à en illustrer concrètement les étapes de mise en œuvre. Sur chaque cas d'étude, en fonction des données disponibles, il était envisagé d'aborder tout ou partie des aspects suivants.

- Modélisation de la pollution par une approche géostatistique adaptée. Relation entre nombre de données disponibles et précision de l'estimation des teneurs en place.
- En cas d'existence de données de dépollution : vérification des estimations, localement (sur les teneurs) et globalement (sur les volumes). Si de telles données n'existent pas, d'autres modes de vérification pourront être établis à partir de données de contrôle qui seront définies au cas par cas.
- Dans la mesure du possible, modélisation de la pollution par une ou des approches déterministes classiques telles que couramment employées lors de la caractérisation de la pollution et de l'étude technico-économique.
- Comparaison des résultats obtenus avec les différentes approches.

Les études de démonstration proviennent de jeux de données fournis par des industriels et des bureaux d'étude. Les analyses géostatistiques ont été réalisées par les membres experts du groupe de travail.

Comme explicité dans le préambule, **les études de démonstration ne sont pas des exemples d'études géostatistiques** et la présentation de ces études s'étend au-delà des rapports d'étude habituels. Les orientations ont largement été dictées ou adaptées aux données disponibles qui ont servi à réaliser les études géostatistiques ou à les valider.

Résumé des travaux réalisés et des résultats obtenus

Les études de démonstration ont été réalisées selon l'état de l'art en géostatistique (voir Manuel Méthodologique GeoSiPol). Elles ont conduit à des conclusions propres à chacune et qui illustrent l'apport de la géostatistique, ses limites et les précautions à prendre quant à son application aux sites et sols pollués. Les travaux réalisés et les résultats obtenus pour chaque étude sont résumés ci-après.

1. CAS D'ETUDE TOTAL FRANCE

Le site est un ancien dépôt d'hydrocarbures démantelé suite à la fin de son activité et ayant fait l'objet de plusieurs campagnes d'échantillonnage, concernant l'analyse des hydrocarbures dans les sols. L'Évaluation Détaillée des Risques (EDR) a été menée afin d'évaluer le niveau de risque résiduel lié à l'ancienne activité du dépôt, en fonction de l'usage futur envisagé sur le site et a fixé un seuil de dépollution global, vis à vis des hydrocarbures totaux (HCT), de 2500 ppm (sur matière sèche).

Enfin, la dépollution du site ayant eu lieu, il est possible aujourd'hui de comparer les volumes estimés par l'approche géostatistique aux volumes réels rencontrés sur le site lors de cette phase.

L'étude géostatistique avait pour objectifs la cartographie de la pollution moyenne en HCT sur le site et de la probabilité de dépassement du seuil de 2500 ppm ainsi que l'évaluation des volumes de terres contaminées et à excaver en vue d'un chiffrage du coût de la dépollution du site.

L'intérêt d'une telle approche est de baser l'évaluation des coûts liés aux volumes contaminés sur les données disponibles, en intégrant cependant l'incertitude liée à l'hétérogénéité spatiale de la contamination et à la reconnaissance uniquement parcellaire de celle-ci.

Les conclusions ont montré que la première étude géostatistique réalisée à maille fine d'1 m de côté était adaptée pour prévenir d'éventuels contrôles ponctuels, pour délimiter les zones polluées et pour conseiller la réalisation de nouveaux sondages dans les zones les plus incertaines. En revanche la maille d'estimation étant très différente de la maille retenue pour l'excavation, le calcul des volumes pollués n'est pas pertinent. En effet, en l'absence de connaissance de la maille retenue lors de l'excavation, aucune quantification des volumes ne peut être réalisée sans biais, l'estimation des volumes pollués au-delà d'un certain seuil dépendant du support d'estimation choisi.

Pour cela une deuxième étude a été réalisée. Intégrant de plus les nouveaux sondages réalisés à la suite de la première étude géostatistique, elle a montré la pertinence de la géostatistique pour l'estimation des volumes pollués. Le volume pollué réel de 13 171 m³ est en effet proche de la valeur estimée comme étant la plus probable (12 059 m³), correspond au quantile 25% des valeurs estimées et est donc largement inclus dans l'intervalle de confiance à 90% construit à partir des estimations. Il est également important de noter que l'excavation ne s'est pas déroulée exactement à maille 15 m mais avec des irrégularités et des passes successives horizontales en bordure de zone contaminée. Verticalement, la maille utilisée est inconnue et probablement au minimum de 50 cm alors que l'estimation géostatistique de l'épaisseur est faite de manière beaucoup plus fine.

Après l'estimation du volume pollué, il est également nécessaire de calculer un volume qu'il conviendra d'excaver pour atteindre toute la pollution. En effet, en raison de la variabilité spatiale des teneurs et de l'incertitude liée à la profondeur réelle de la couche contaminée en HCT, il est recommandé de ne pas baser l'excavation sur l'estimation du volume pollué uniquement, en raison du risque important qu'il présente de laisser en place des sols contaminés.

En considérant le quantile 25% des estimations pour le volume pollué ainsi que le scénario le plus probable et le scénario sécuritaire (qui considère les quantiles 25% pour la profondeur du toit de la couche et 75% pour la base) pour la géométrie de la couche potentiellement contaminée, un encadrement du volume à excaver entre 17 750 m³ et 31 239 m³ est réalisé. Cet intervalle comprend le volume réel excavé de 22 348 m³.

Il est évident que procéder de la sorte conduira à l'excavation de volumes non négligeables de terres saines, qui pourraient éventuellement être triées en cours de dépollution. Prévoir la réhabilitation de

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

tels volumes permet cependant de s'affranchir du risque de laisser en place des terres contaminées et de beaucoup mieux prévoir les coûts de dépollution.

Enfin, la comparaison maille à maille entre la zone réellement excavée et les prédictions a été réalisée en superposant les cartes. Elle montre une très bonne concordance entre l'étude géostatistique et les zones réellement excavées, confirmant la qualité des résultats globaux précédents et donc la pertinence de l'approche géostatistique mise en œuvre sur ce site.

2. CAS D'ETUDE 1BIS ATTENA (BURGEAP)

Ce site de recherche n'a pas fait l'objet d'une dépollution avec excavation des sols. Par conséquent, les résultats de l'étude géostatistique n'ont pas pu être validés de façon classique en les confrontant à des volumes de sols pollués excavés. En rapport avec les travaux de recherche menés dans le cadre du projet ATTENA, d'autres objectifs d'étude tout aussi intéressants ont cependant pu être définis en s'intéressant à la délimitation de la « zone source » constituée des sols dont la somme des teneurs en PCE, TCE et DCE est suffisamment élevée pour qu'une phase organique y soit présente.

Pour cette étude de démonstration, trois modèles géostatistiques ont été générés. Ces modèles ont consisté à simuler conjointement les teneurs en PCE, TCE et DCE en exploitant progressivement les nombreuses données disponibles (287 échantillons de sols analysés).

1. Modèle géostatistique de base. Le jeu de données complet a été décomposé en deux sous-ensembles de 49 et 238 échantillons, le premier devant servir à construire le modèle géostatistique de la pollution des sols, le second à vérifier les bonnes prédictions du modèle. Le choix des données du modèle a été effectué par Burgeap en se déterminant par rapport à des conditions d'étude habituelles sur le nombre de données et leur répartition (en particulier verticale).
2. Modèle géostatistique amélioré. Une première amélioration a porté sur les modèles de distribution et variographiques (distributions, corrélations entre composants et variogrammes). Le jeu de données complet (287 échantillons) a ainsi été utilisé pour estimer plus justement ces modèles et étudier l'impact de cette amélioration sur la simulation géostatistique des teneurs en utilisant toujours le jeu de données restreint (49 échantillons) pour conditionner les simulations. Comme précédemment, les 238 données de validation, non prises en compte dans les simulations, ont servi à analyser les prédictions du modèle géostatistique amélioré.
3. Modèle géostatistique complet. En s'appuyant sur les modèles statistiques améliorés établis précédemment, les teneurs ont été simulées à nouveau mais en les conditionnant sur le jeu de données complet (287 données). Le modèle géostatistique ainsi obtenu montre comment les 238 données nouvelles contribuent à réduire amplement l'incertitude spatiale. Ce modèle peut être considéré comme un modèle de référence de la zone source par rapport aux modèles précédents.
4. Ce cas d'étude présente deux particularités intéressantes. D'une part, il nécessite de simuler conjointement les trois composants, qui sont corrélés entre eux, pour modéliser la zone source. D'autre part, il permet d'aborder le problème des données sous les limites de détection analytiques, qui représentent de 15 à 50% des données selon les composants et dont la prise en compte dans les modèles géostatistiques nécessite des précautions particulières.

En tant que démonstration de l'application de la géostatistique aux sites et sols pollués, cette étude apporte des résultats sur les aspects suivants.

- Pertinence de la géostatistique pour estimer le volume et la localisation d'une zone source ou plus généralement de sols pollués (teneurs supérieures à un seuil défini de manière indépendante).
- Validation des prédictions des modèles géostatistiques et interprétation des écarts.
- Importance des analyses statistiques et variographiques sur la qualité des modèles géostatistiques.
- Importance des données complémentaires pour réduire les incertitudes spatiales.
- Complexité de ces études et façon de les aborder.

3. CAS D'ETUDE T (ARCELORMITTAL REAL ESTATE)

L'ancienne cokerie du site T a fait l'objet de plusieurs campagnes d'échantillonnage dans le but de délimiter et quantifier la contamination du sol. La présente étude reprend les données de deux campagnes, LECES et GEOSTOCK. L'examen des données met en évidence une difficulté majeure pour l'étude de la pollution : alors que le sol est extrêmement hétérogène, aucune information n'a été transmise concernant les méthodes d'échantillonnage et le volume des échantillons prélevés. Or, dans un milieu hétérogène comme celui-ci, ces informations sont fondamentales car de nombreuses décisions méthodologiques en dépendent.

En l'absence d'informations précises sur l'échantillonnage, nous nous sommes attachés à présenter à chaque étape les divers choix qui s'offraient à nous, et nous avons investigué les conséquences de chacun d'eux.

In fine, nous avons proposé une méthode de cartographie des contaminations en HAP, plomb et zinc qui regroupe tous les échantillons, sans que nous sachions si les hypothèses sous-jacentes étaient vérifiées. L'analyse de la carte associée d'écart-type de krigeage montre que la moyenne de cet écart-type sur toute l'épaisseur étudiée est toujours supérieure au seuil de contamination. Les cartes de contamination ponctuelle ne permettent donc pas de délimiter les zones contaminées, sauf à accepter un risque élevé de laisser de la contamination en place.

Pour l'estimation de la concentration sur des mailles de 10m x 10m x 3m, le calcul de l'intervalle de confiance conventionnel permet de séparer des zones très probablement contaminées de zones où la contamination est possible mais non certaine. Il n'y a pas de zone très probablement non contaminée. Mais la mauvaise qualité des données et de leur modélisation ne nous permet d'avoir qu'une confiance limitée dans des résultats.

L'intérêt de ce site pour une étude de démonstration n'est donc pas de montrer les méthodes géostatistiques à utiliser, mais de permettre d'expliquer en quoi les informations sur les méthodes d'échantillonnage sont fondamentales pour une étude de géostatistique de qualité.

Les données supplémentaires que nous espérions pour valider la méthodologie nous ont été transmises tardivement. La mauvaise qualité de la modélisation étant déjà connue, les résultats d'une validation, qu'elle soit favorable ou non, ne peuvent pas être significatifs et cette validation n'a pas été effectuée.

4. ANALYSE EXPLORATOIRE DES DONNEES DU SITE « CHATEAU » (ICF ENVIRONNEMENT)

Le site de Château est une ancienne station-service pour poids lourds. La pollution concerne essentiellement les sols et porte sur des hydrocarbures. Les investigations du site ont été réalisées, ainsi qu'une étude technico-économique des alternatives de réhabilitation. Quarante sondages ont été réalisés sur le site avec prélèvement d'échantillons de sols qui ont été analysés en laboratoire (mesure de fractions d'hydrocarbures).

La dépollution des sols, initialement prévue en 2009, puis début 2010, a finalement dû être repoussée à 2012. Cette dépollution était et continue d'être prévue par excavation des sols. Lors de l'excavation, les terres seront triées en s'appuyant sur des indices de terrain lié à la présence d'hydrocarbures.

Sans pouvoir disposer des données de dépollution indispensables à la démonstration, la modélisation géostatistique de la pollution des sols de ce site n'a finalement pas été réalisée. En contrepartie, des travaux complémentaires ont pu être menés sur le cas d'étude ATTENA qui porte également sur un petit site.

Le site de Château demeure cependant un cas d'étude intéressant et complémentaire des autres cas d'étude présentés dans ce rapport. La possibilité d'en faire une étude de démonstration sera étudiée en 2012 lorsque les données de dépollution seront enfin disponibles.

Étude géostatistique de la contamination d'un dépôt d'hydrocarbures : Le site « TOTAL ».

Fournisseur du cas d'étude :	TOTAL
Réalisation de l'étude :	GEOVARIANCES
Contrôle qualité :	KIDOVA
Rédacteurs :	Claire Fauchaux, Nicolas Jeannée
Outil logiciel utilisé :	Isatis® version 2011 (GEOVARIANCES)
Relecteur :	Michel Garcia, Jean-Baptiste Mathieu

TABLE DES MATIERES

1. PRESENTATION DU SITE ET DES DONNEES	15
1.1 PRESENTATION DU SITE.....	15
1.2 OBJECTIFS.....	16
1.3 DONNEES	16
1.3.1 Campagne de 2002	16
1.3.2 Campagne de 2005	16
1.3.3 Échantillonnage complémentaire de juin 2006	17
1.3.4 Données d'excavation	18
2. METHODOLOGIE ET ANALYSE STATISTIQUE DES DONNEES	19
2.1 METHODOLOGIE.....	19
2.2 DEMARCHE ITERATIVE.....	19
2.3 ANALYSE EXPLORATOIRE DES TENEURS EN HCT	20
3. ÉVALUATION DE LA COUCHE POTENTIELLEMENT CONTAMINEE	21
3.1 MODELISATION DU TOIT DE LA COUCHE POTENTIELLEMENT CONTAMINEE	21
3.2 MODELISATION DE LA BASE DE LA COUCHE POTENTIELLEMENT CONTAMINEE.....	22
3.3 CONCLUSION.....	24
4. MODELISATION DES TENEURS EN HCT AU SEIN DE LA COUCHE CONTAMINEE REALISEE DEBUT 2006.....	25
4.1 SIMULATIONS GEOSTATISTIQUES	25
4.2 CARTOGRAPHIE DES TENEURS ET PROBABILITE DE DEPASSEMENT DU SEUIL DE 2500 PPM	25
4.3 ESTIMATION DES VOLUMES CONTAMINES ET DES VOLUMES A EXCAVER ET COMPARAISON AVEC LES RESULTATS DE DE POLLUTION	27
4.3.1 Volume de sol contaminé	27
4.3.2 Volume de sol à excaver	28
5. MODELISATION DES TENEURS EN HCT AU SEIN DE LA COUCHE CONTAMINEE EN INTEGRANT LES SONDAGES DE JUIN 2006 ET EN PRENANT EN COMPTE LA MAILLE D'EXCAVATION.....	31
5.1 PRISE EN COMPTE DU MAILLAGE D'EXCAVATION	31
5.1.1 Effet de support	31
5.1.2 Probabilité de dépassement de seuil	32
5.1.3 Estimation des volumes et comparaison avec les résultats de dépollution	33
5.2 PRISE EN COMPTE DES DIX-SEPT SONDAGES COMPLEMENTAIRES	33
5.2.1 Probabilité de dépassement de seuil	33
5.2.2 Estimation des volumes pollués et comparaison aux résultats de dépollution	35
5.2.3 Estimation des volumes à excaver et comparaison aux résultats de dépollution	36
5.2.4 Comparaison des mailles polluées estimées et excavées	37
6. CONCLUSION	39
ANNEXE A : DONNEES DES CAMPAGNES DE 2005 ET 2006.....	41
ANNEXE B : DONNEES D'EXCAVATION	44

TABLE DES FIGURES

FIGURE 1 : CONTOUR INDICATIF DU SITE. LOCALISATION EN JAUNE DE LA ZONE SUSPECTEE COMME ETANT CONTAMINEE AU TERME DE LA CAMPAGNE DE 2002 ET DE L'EDR. LOCALISATION EN ROUGE DES MAILLES ET SONDAGES REALISES LORS DE LA CAMPAGNE DE DECEMBRE 2005.	15
FIGURE 2 : LOCALISATION DES 17 SONDAGES COMPLEMENTAIRES REALISES EN JUIN 2006.	17
FIGURE 3 : CARTE D'IMPLANTATION DES 82 SONDAGES DE LA CAMPAGNE DE 2005 (CROIX VERTES) ET DES 17 SONDAGES COMPLEMENTAIRES DE JUIN 2006 (ETOILES BLEUES).	20
FIGURE 4 : HISTOGRAMME DES TENEURS MOYENNES EN HCT DANS LA COUCHE POTENTIELLEMENT CONTAMINEE (EN VERT : DONNEES 2005, EN BLEU : SONDAGES COMPLEMENTAIRES).	20
FIGURE 5 : VARIOGRAMME EXPERIMENTAL ET MODELE AJUSTE POUR LA TRANSFORMEE GAUSSIENNE DU TOIT DE LA COUCHE POTENTIELLEMENT CONTAMINEE. INDICATION DU NOMBRE DE PAIRES POUR CHAQUE POINT DU VARIOGRAMME EXPERIMENTAL ET DE LA VARIANCE STATISTIQUE (TRAIT NOIR).	21
FIGURE 6 : QUANTILES A 10, 25 ET 50% POUR LA PROFONDEUR DU TOIT DE LA COUCHE POTENTIELLEMENT CONTAMINEE (EN M). INDICATION DES LIGNES D'ISO-VALEURS TOUTS LES 0,5 M.	22
FIGURE 7 : VARIOGRAMME EXPERIMENTAL ET MODELE AJUSTE POUR LA TRANSFORMEE GAUSSIENNE DE L'EPAISSEUR DE LA COUCHE POTENTIELLEMENT CONTAMINEE. INDICATION DU NOMBRE DE PAIRES POUR CHAQUE POINT DU VARIOGRAMME EXPERIMENTAL ET DE LA VARIANCE STATISTIQUE (TRAIT NOIR).	23
FIGURE 8 : QUANTILES A 50, 75 ET 90% POUR L'EPAISSEUR DE LA COUCHE POTENTIELLEMENT CONTAMINEE (EN M). INDICATION DES LIGNES D'ISO-VALEURS TOUTS LES 0,5 M.	23
FIGURE 9 : QUANTILES A 50, 75 ET 90% POUR LA PROFONDEUR DE LA BASE DE LA COUCHE POTENTIELLEMENT CONTAMINEE (EN M). INDICATION DES LIGNES D'ISO-VALEURS TOUTS LES 0,5 M.	24
FIGURE 10 : VARIOGRAMME EXPERIMENTAL ET MODELE AJUSTE POUR LA TRANSFORMEE GAUSSIENNE DES TENEURS EN HCT (DEBUT 2006) AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE. INDICATION DU NOMBRE DE PAIRES POUR CHAQUE POINT DU VARIOGRAMME EXPERIMENTAL ET DE LA VARIANCE STATISTIQUE (TRAIT NOIR).	25
FIGURE 11 : CARTOGRAPHIE DES TENEURS EN HCT (DEBUT 2006) AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE (EN PPM). ISO-CONTOUR A 2 500 PPM.	26
FIGURE 12 : CARTOGRAPHIE DE LA PROBABILITE DE DEPASSEMENT DE 2 500 PPM EN HCT AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE. CONTOURS D'ISOPROBABILITE A 5% (PROBABILITE DE 0,05 : TRAIT FIN NOIR), 10% (0,1 : TRAIT EPAIS NOIR), 25% (TRAIT FIN ROUGE) ET 50% (TRAIT EPAIS ROUGE), CONTOURS DES ZONES CIBLES POUR LES CALCULS DE VOLUMES (EN GRIS).	26
FIGURE 13 : DISTRIBUTION (HISTOGRAMME CUMULE INVERSE) DES VOLUMES CONTAMINES EN HCT AU SEIN DES TROIS ZONES CIBLES INCLUANT LA CONTAMINATION DE 25 CM DU TOIT DES ARGILES ; INDICATION DES VOLUMES CORRESPONDANTS AUX QUANTILES A 5, 10, 50, 90 ET 95 (EN M3).	27
FIGURE 14 : ILLUSTRATION DE L'EFFET DE SUPPORT LORS DU CALCUL DE VOLUMES DE TERRES POLLUEES DEPASSANT UN SEUIL.	32
FIGURE 15 : CARTOGRAPHIE DE LA PROBABILITE DE DEPASSEMENT DE 2 500 PPM EN HCT AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE A MAILLE 15 M. CONTOURS D'ISOPROBABILITE A 10% ET A 25 %.	32
FIGURE 16 : DISTRIBUTION (HISTOGRAMME CUMULE INVERSE) DES VOLUMES CONTAMINES EN HCT (ESTIMES A MAILLE 15 M) AU SEIN DES TROIS ZONES CIBLES INCLUANT LA CONTAMINATION DE 25 CM DU TOIT DES ARGILES ; INDICATION DES VOLUMES CORRESPONDANTS AUX QUANTILES A 5, 10, 50, 90 ET 95%.	33
FIGURE 17 : SUPERPOSITION DE LA CARTOGRAPHIE DE LA PROBABILITE DE DEPASSEMENT DE 2500 PPM EN HCT AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE A MAILLE 15 M ET DES TENEURS DES 17 NOUVEAUX SONDAGES (LA TAILLE DES CARRES NOIRS EST PROPORTIONNELLE A LA TENEUR).	34
FIGURE 18 : CARTOGRAPHIE DE LA PROBABILITE DE DEPASSEMENT DE 2 500 PPM EN HCT AU SEIN DE LA COUCHE POTENTIELLEMENT CONTAMINEE A MAILLE 15 M ET EN INCLUANT LES NOUVEAUX SONDAGES. CONTOURS D'ISOPROBABILITE A 10% ET A 25 %.	34
FIGURE 19 : NUAGE DE CORRELATION ENTRE PROBABILITES DE DEPASSEMENT DE SEUIL AVANT ET APRES INTEGRATION DES NOUVEAUX SONDAGES ET CARTE CORRESPONDANTES. LES COULEURS DES MAILLES MISES EN EVIDENCE SONT LES MEMES SUR LE NUAGE DE CORRELATION ET SUR LES CARTES.	35
FIGURE 20 : DISTRIBUTION (HISTOGRAMME CUMULE INVERSE) DES VOLUMES CONTAMINES EN HCT (ESTIMES A MAILLE 15 M ET INCLUANT LES NOUVEAUX SONDAGES) AU SEIN DES TROIS ZONES CIBLES INCLUANT LA CONTAMINATION DE 25 CM DU TOIT DES ARGILES ; INDICATION DES VOLUMES CORRESPONDANTS AUX QUANTILES A 5, 10, 50, 90 ET 95%.	35
FIGURE 21 : SUPERPOSITION DE LA ZONE REHABILITEE (EN ROUGE) AUX CONTOURS DE PROBABILITES DE DEPASSEMENT DE SEUIL A 25 ET 50%.	37
FIGURE 22 : SUPERPOSITION DE LA CARTE DE PROBABILITE DE DEPASSEMENT DU SEUIL DE 2500 PPM EN HCT ET DES CONTOURS DE LA ZONE EXCAVEE AINSI QUE DES MAILLES IRRÉGULIÈRES UTILISÉES.	38

TABLE DES TABLEAUX

TABLEAU 1 : VOLUMES A EXCAVER POUR DIFFERENTS SCENARII DE RISQUE LIES A LA GEOMETRIE DE LA COUCHE POTENTIELLEMENT CONTAMINEE AINSI QU'AU DEPASSEMENT DE 2 500 PPM EN HCT (RESULTATS BRUTS PRENANT EN COMPTE UNE CONTAMINATION PROBABLE DU TOIT DES ARGILES SUR 25 CM).	30
TABLEAU 2 : VOLUMES A EXCAVER POUR DIFFERENTS SCENARIOS DE RISQUE LIES A LA GEOMETRIE DE LA COUCHE POTENTIELLEMENT CONTAMINEE AINSI QU'AU DEPASSEMENT DE 2500 PPM EN HCT (RESULTATS BRUTS PRENANT EN COMPTE UNE CONTAMINATION PROBABLE DU TOIT DES ARGILES SUR 25 CM, LA MAILLE DE DE POLLUTION ET LES NOUVEAUX SONDAGES).	36
TABLEAU 3 : PRINCIPALES ETAPES REALISEES POUR L'ETUDE ET LEUR PERTINENCE DANS LE CADRE D'UNE ETUDE GEOSTATISTIQUE.....	40
TABLEAU 4 : DONNEES DE 2005 UTILISEES POUR L'ETUDE GEOSTATISTIQUE : NUMERO DE SONDAGE, COORDONNEES X, Y EN LAMBERT II, COTE DU SOL ZSOL EN IGN69, PROFONDEURS DE LA BASE DES REMBLAIS ET DU TOIT DES ARGILES, NIVEAUX ECHANTILLONNES (MIN ET MAX) ET TENEURS EN HCT CORRESPONDANTES (EN PPM).	41
TABLEAU 5 : SONDAGES COMPLEMENTAIRES DE JUIN 2006 : NUMERO DE SONDAGE, COORDONNEES X, Y EN LAMBERT II, PROFONDEURS DE LA BASE DES REMBLAIS ET DU TOIT DES ARGILES, NIVEAUX ECHANTILLONNES (MIN ET MAX) ET TENEURS EN HCT CORRESPONDANTES (EN PPM).....	43
TABLEAU 6 : DONNEES DE DE POLLUTION : VOLUMES DE TERRES EXTRAITS ET POLLUES.	44

1. PRESENTATION DU SITE ET DES DONNEES

1.1 Présentation du site

Le site concerné est un ancien dépôt d'hydrocarbures démantelé en 2002 suite à la fin de son activité. Entre 1996 et 2002, il a fait l'objet de plusieurs campagnes d'échantillonnage, notamment concernant l'analyse des hydrocarbures dans les sols. Ces campagnes ont mis en évidence plusieurs zones de contamination significative, cohérentes avec l'activité industrielle du site.

Une Évaluation Détaillée des Risques (EDR) a été menée en 2004 afin d'évaluer le niveau de risque résiduel lié à l'ancienne activité du dépôt, en fonction de l'usage futur envisagé sur le site. Cette EDR est basée exclusivement sur la campagne menée de mai à juin 2002, réalisée lors du diagnostic du site après démantèlement des installations, en raison de sa cohérence avec l'état actuel du site. Au terme de cette EDR, **un seuil de dépollution global pour le site** a été proposé, vis à vis des **hydrocarbures totaux (HCT)**, de **2 500 ppm** (sur matière sèche), valeur intégrant l'ensemble des composés organiques toxiques.

Concernant les volumes contaminés, le mémoire de cessation d'activité de 2002 avait fourni une première évaluation d'approximativement 15 000 m³ contaminés à plus de 1 000 ppm en HCT. La surface contaminée suspectée au terme de cette campagne et suite aux valeurs établies par l'EDR (reprise en jaune sur le plan du site, Figure 1) est approximativement égale à 7 775 m².

En 2004, l'EDR a également conclu à une contamination exclusive des remblais, la couche d'argiles saine sous-jacente agissant comme une barrière peu perméable vis-à-vis de la contamination. Sous l'hypothèse d'une profondeur de remblais en moyenne comprise entre 1,5 et 2 m, cette zone suspectée comme étant contaminée conduit à un volume de remblais compris entre 11 650 et 15 550 m³.

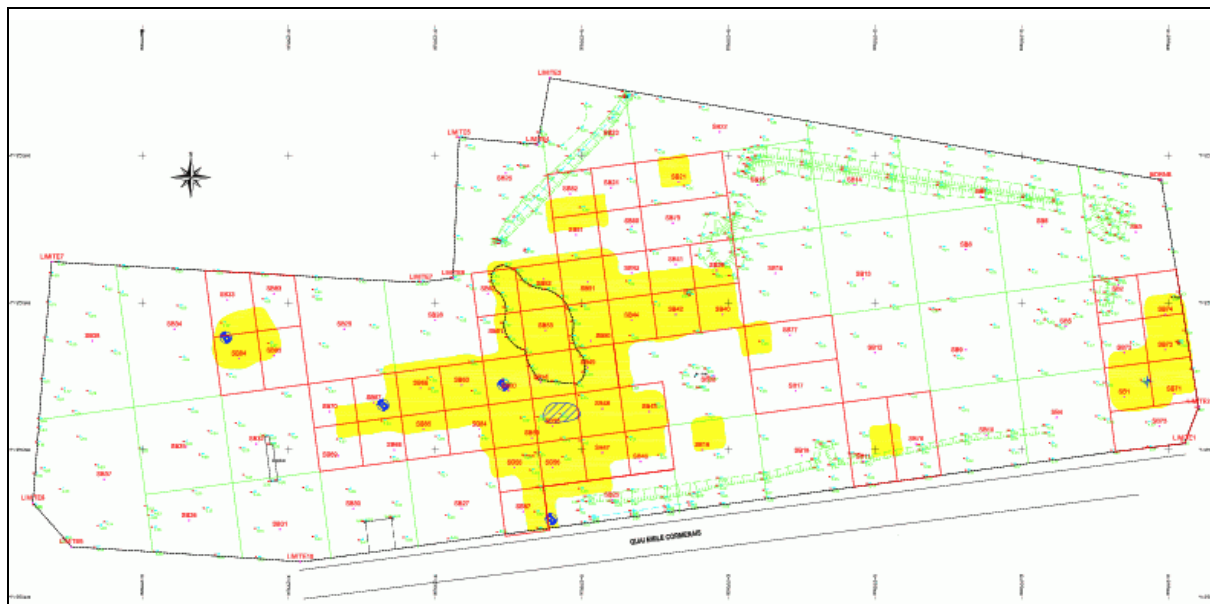


Figure 1 : Contour indicatif du site. Localisation en jaune de la zone suspectée comme étant contaminée au terme de la campagne de 2002 et de l'EDR. Localisation en rouge des mailles et sondages réalisés lors de la campagne de décembre 2005.

1.2 Objectifs

Les objectifs de l'étude géostatistique sont multiples :

- cartographie des teneurs en HCT et des probabilités de dépassement du seuil de 2 500 ppm ;
- estimation des volumes contaminés ;
- estimation des volumes de terre à excaver, compte tenu de la maille utilisée lors de la dépollution et du volume d'accessibilité.

La dépollution du site ayant eu lieu en 2006 et 2007, il est possible aujourd'hui de comparer les volumes estimés par l'approche géostatistique aux volumes réels rencontrés sur le site lors de cette phase.

1.3 Données

1.3.1 Campagne de 2002

Les données analytiques issues de la campagne de 2002 n'ont pas été explicitement intégrées à cette étude, et ce pour plusieurs raisons :

- les protocoles d'échantillonnage mis en œuvre en 2002 puis en 2005 diffèrent notablement, sans que la première campagne n'apporte d'information sur des zones suspectes non couvertes par les investigations de 2005 ;
- elles n'étaient pas disponibles au format numérique.

1.3.2 Campagne de 2005

Néanmoins, les résultats issus des données de 2002 sont bien intégrés à la démarche d'investigation complémentaire de 2005 et aux estimations de volumes qui en découlent. En effet, la campagne de 2005 a été conçue dans un objectif de validation des zones suspectées lors de la campagne de 2002 :

- vérification de l'extension des zones suspectes par un échantillonnage systématique de ces zones (mailles de 15 x 15 m) ;
- investigations sur une maille moins fine de 30 x 30 m des zones non privilégiées lors de la campagne de 2002.

La campagne de 2005 (tableau de données fourni en Annexe A) consiste donc en un échantillonnage systématique à 15 m des zones potentiellement contaminées, complété par quelques points répartis de façon homogène sur le reste du site (voir Figure 1). Deux échantillons ont été prélevés et analysés en chacun des 82 sondages réalisés. Ces échantillons sont prélevés à des profondeurs comprises en moyenne :

- entre 0 et 1 m pour les premiers échantillons (notés A, quelques uns descendant néanmoins jusqu'à 2 m), tous situés dans les remblais,
- entre 1 et 2 m pour les seconds échantillons (notés B, cinq échantillons descendant à 3 m voire 3.5 m), situés majoritairement dans les remblais. Une vingtaine de ces échantillons inclut 10 à 20 cm des terrains sous-jacents aux remblais ; 19 autres échantillons sont situés exclusivement sous les remblais et concernent soit des sables, soit le toit des argiles sous-jacentes.

La distinction entre échantillons A et B n'est qu'en partie valide, la stratégie conduisant au choix des hauteurs prélevées ayant varié sur le site afin d'être cohérente avec les niveaux potentiellement contaminés. Cette stratégie a été élaborée comme suit, sachant que deux échantillons devaient être prélevés par sondage :

- en cas d'absence d'indication visuelle de contamination, les niveaux 0-1 m et 1-2 m ont été prélevés ;
- en cas d'indication visuelle de contamination dans la partie basse des remblais, sachant que les hydrocarbures ont majoritairement migré vers le bas : prélèvement d'un échantillon dans la partie supérieure non contaminée et d'un second dans la partie visiblement contaminée ;
- en cas d'indication visuelle de contamination homogène sur la hauteur des remblais : prélèvement d'un échantillon moyen dans les remblais et d'un second échantillon en début de couche sous-jacente (en raison d'une contamination vraisemblable de 20 à 30 cm du toit des argiles).

Cette stratégie présente l'avantage, à moindre coût, de laisser espérer une bonne délimitation de la zone contaminée, à condition que l'indication visuelle soit licite, ce qui est en général le cas pour les hydrocarbures. Par contre, cette stratégie oriente complètement les calculs de volumes, qui ne pourront être menés que sur base de ces hauteurs échantillonnées, sans espoir de sélectionner plus finement les remblais contaminés. Néanmoins, la contamination paraissant clairement identifiée dans le cas présent, cela peut ne pas être forcément gênant. Une approche à 3D menée directement sur les teneurs en HCT n'est par ailleurs pas envisageable, en raison de l'hétérogénéité des hauteurs prélevées pour les différents sondages.

1.3.3 Échantillonnage complémentaire de juin 2006

Avant le démarrage des travaux d'excavation un diagnostic complémentaire a été réalisé afin :

- d'affiner les zones reconnues comme impactées dans les diagnostics antérieurs au droit des mailles d'extension 30 x 30 m,
- de déterminer les limites d'extension des terres polluées à excaver, localisées sous la zone sud de la plateforme de traitement.

Au total, 17 sondages complémentaires ont été réalisés selon un maillage 15 x 15 m (Figure 2 et tableau de données fourni en Annexe A) et en accord avec le premier travail géostatistique effectué en 2006.

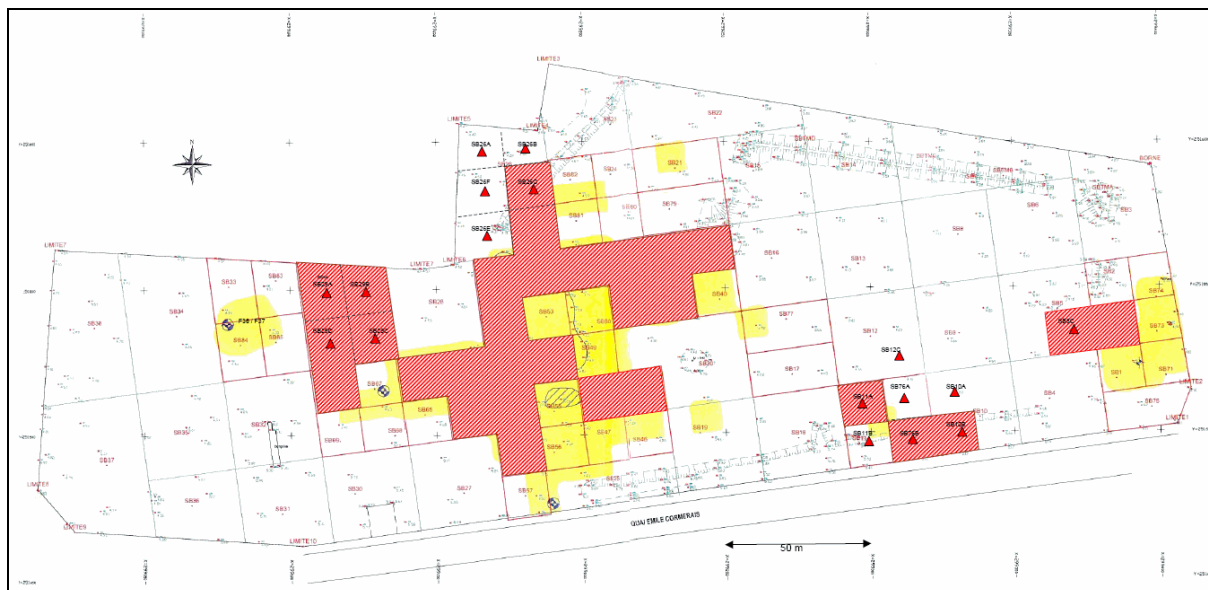


Figure 2 : Localisation des 17 sondages complémentaires réalisés en juin 2006.

1.3.4 Données d'excavation

L'excavation s'est déroulée en plusieurs phases entre juillet 2006 et novembre 2007. Les volumes excavés et pollués rencontrés au cours des différentes phases dans 3 zones cibles définies à l'aide des différents diagnostics réalisés sur le terrain sont repris ci-dessous :

- juillet 2006 : excavation et tri de **16 231 m³** de matériaux parmi lesquels **8 770 m³** étaient pollués (une tolérance de pénétration dans les argiles de 30 cm a été appliquée) ;
- août 2006 : reprise de la phase de tri et d'excavation avec un volume excavé de **6 117 m³** et un volume pollué de **4 401 m³** ;
- novembre 2007 : suite aux investigations précédentes et aux études de risques complémentaires, les fonds des mailles les plus souillées sont excavés sur 1 m de profondeur dans les sédiments fins alluvionnaires. Cette seconde phase conduit à un volume supplémentaire de **1 554 m³**.

L'étude que l'on réalise ici vise à localiser les teneurs et à caractériser les volumes en vue de la première phase d'excavation. La deuxième phase réalisée en novembre 2007 n'est donc pas prise en compte. Le volume pollué total que l'on considère est donc de **13 171 m³** et le volume excavé est de **22 348 m³**. Les données maille par maille sont reportées en Annexe B, sans toutefois pouvoir disposer d'informations supplémentaires telles que le protocole exact d'excavation par maille ou la prise en compte ou non d'informations organoleptiques au cours de l'excavation.

2. METHODOLOGIE ET ANALYSE STATISTIQUE DES DONNEES

2.1 Méthodologie

Les premières observations relatives aux teneurs en HCT et à la stratégie d'échantillonnage vont fortement orienter la méthodologie à privilégier pour l'évaluation des volumes contaminés.

En raison d'une incertitude sur la localisation de la couche contaminée due à l'hétérogénéité du milieu et de la variabilité spatiale des teneurs, la méthodologie retenue comporte deux étapes successives. Elle peut être résumée ainsi :

1. Délimitation géométrique de la couche de remblais potentiellement contaminée :

Cette couche correspond aux niveaux échantillonnés en raison de leur indication visuelle de contamination. Néanmoins, il ne faudrait pas accorder trop de poids à cette indication visuelle, au risque de rater des contaminations significatives. Par conséquent, lorsqu'un prélèvement présente une concentration en HCT supérieure à une valeur de présomption de contamination fixée à 1 000 ppm, le niveau échantillonné sera potentiellement inclus dans les calculs de volumes. Cela signifie que nous considérons qu'il est possible que les remblais soient significativement contaminés à proximité de ce point. La base de cette couche est constituée par la base des prélèvements indiquant une contamination significative.

Si en un sondage aucun des deux prélèvements ne conduit à une teneur supérieure à 1 000 ppm en HCT, alors, la couche potentiellement contaminée est d'épaisseur nulle.

2. Analyse géostatistique des teneurs en HCT au sein de la couche contaminée :

Au sein de cette couche, certains prélèvements présentent des teneurs en HCT inférieures au seuil de 2 500 ppm. Une analyse géostatistique sera donc réalisée pour les teneurs en HCT issues de cette couche, afin de sélectionner les zones qui semblent effectivement contaminées au-delà du seuil préconisé au terme de l'EDR de 2004.

Au sein de la couche contaminée, lorsque elle inclut les deux prélèvements réalisés en un point, la teneur en HCT considérée pour la modélisation moyenne au sein de la couche est obtenue en moyennant les teneurs en HCT des deux prélèvements A et B, pondérées par la hauteur respective de chaque prélèvement, soit :

Teneur moyenne = (Teneur A x Hauteur A + Teneur B x Hauteur B) / (Hauteur A + Hauteur B)

2.2 Démarche itérative

Début 2006, les données de la campagne de 2005 ont été analysées selon cette méthodologie afin de parvenir à une estimation des concentrations sur une grille de maille 1 m. Ce premier travail ne prenait donc pas en compte les résultats des 17 sondages complémentaires réalisés en juin 2006. Le choix de la maille avait été réalisé en vue de contrôles possibles lors de la dépollution et afin de réaliser une délimitation fine de la pollution.

Les résultats de cette première étape ayant conduit à une première estimation sont d'abord présentés et comparés aux données de dépollution.

Dans un second temps deux modifications ont été apportées et l'étude a été reconduite, de manière à :

- prendre en compte le maillage de 15 m utilisé lors de la dépollution du site ayant eu lieu au cours des années 2006 et 2007 ;
- intégrer les résultats des 17 sondages complémentaires réalisés en juin 2006.

Ces nouveaux résultats sont alors eux aussi comparés aux données de dépollution.

2.3 Analyse exploratoire des teneurs en HCT

Les deux figures suivantes présentent la carte d'implantation des 82 sondages initiaux de la campagne 2005 et des 17 sondages complémentaires de juin 2006 (Figure 3) ainsi que l'histogramme des teneurs moyennes en HCT au sein de la couche contaminée (Figure 4). La distribution des teneurs est très dissymétrique et présente peu de valeurs fortes mais beaucoup de valeurs faibles (49 échantillons présentent une valeur inférieure à 1 000 ppm et 69 une valeur inférieure à 2 500 ppm). Afin de réaliser tous les traitements géostatistiques permettant de répondre aux objectifs il est nécessaire de transformer ces données afin d'obtenir une distribution gaussienne. Par la suite, les différents traitements sont donc appliqués sur la transformée gaussienne des teneurs dans la couche contaminée ; la transformation est réalisée par anamorphose gaussienne.

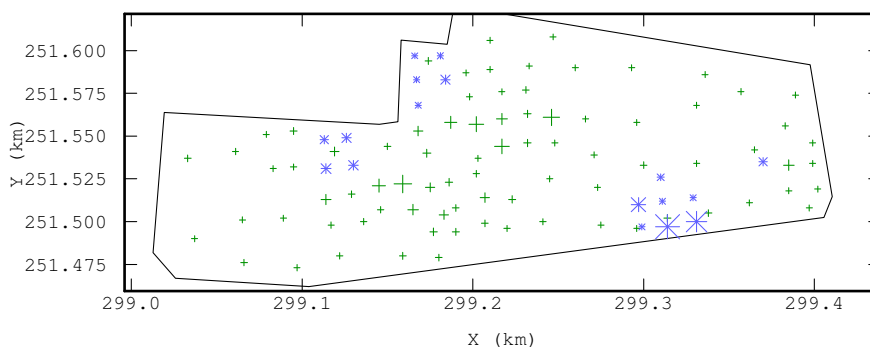


Figure 3 : Carte d'implantation des 82 sondages de la campagne de 2005 (croix vertes) et des 17 sondages complémentaires de juin 2006 (étoiles bleues).

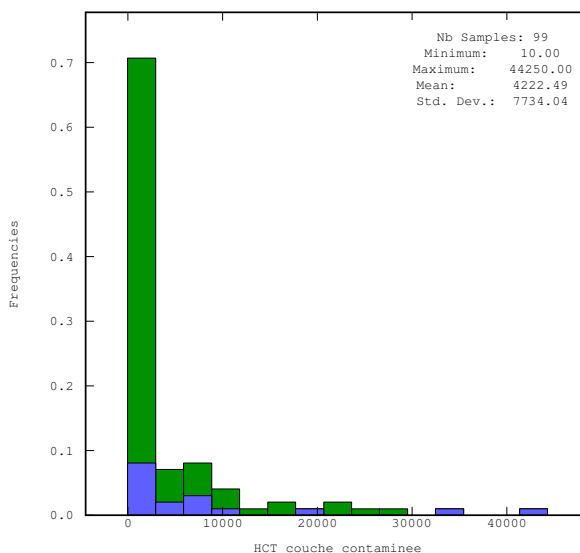


Figure 4 : Histogramme des teneurs moyennes en HCT dans la couche potentiellement contaminée (En vert : données 2005, en bleu : sondages complémentaires).

3. ÉVALUATION DE LA COUCHE POTENTIELLEMENT CONTAMINÉE

Le toit et la base de la couche potentiellement contaminée sont déduits des hauteurs des prélèvements ayant conduit à des teneurs en HCT supérieures à la valeur de présomption égale à 1 000 ppm.

3.1 Modélisation du toit de la couche potentiellement contaminée

Une estimation la plus réaliste possible du toit de la couche potentiellement contaminée n'est pas suffisante, car elle néglige l'incertitude potentiellement forte liée à l'estimation du phénomène en tout point du site à partir de quelques sondages. Ainsi, afin de pouvoir caractériser cette incertitude, une approche par simulations géostatistiques est réalisée. Chaque simulation correspond à un scénario possible pour le toit de la couche, qui respecte :

- les valeurs du toit de la couche observées aux points de sondages,
- la distribution statistique de ces valeurs sur le site,
- la continuité spatiale du toit de la couche, appréhendée par le calcul du variogramme des données de toit de la couche.

L'approche par simulations requiert une hypothèse de normalité pour la variable étudiée. Le toit de la couche ne suivant pas une telle distribution, une transformation par anamorphose est réalisée. L'analyse de la structure spatiale est alors menée pour la transformée gaussienne du toit de la couche potentiellement contaminée.

La Figure 5 présente le variogramme expérimental de la transformée gaussienne du toit de la couche contaminée ainsi que la modélisation associée (les 17 sondages complémentaires ne sont ici pas pris en considération). Il est ensuite utilisé pour réaliser 200 simulations géostatistiques desquelles il est possible de déduire différentes statistiques. La Figure 6 présente ainsi les cartes de quantiles à 10, 25 et 50%. La carte de quantile à 10% représente en chaque maille de 1 x 1 m la valeur du toit telle que l'on a seulement 10% de risques de lui être inférieure. En matière de caractérisation de la couche potentiellement contaminée, cette carte représente donc un choix sécuritaire. Le quantile à 50% peut quant à lui, tout comme la moyenne des simulations, être interprété comme le scénario le plus probable. La différence entre ces cartographies illustre la variabilité spatiale existant pour le toit de la couche potentiellement contaminée.

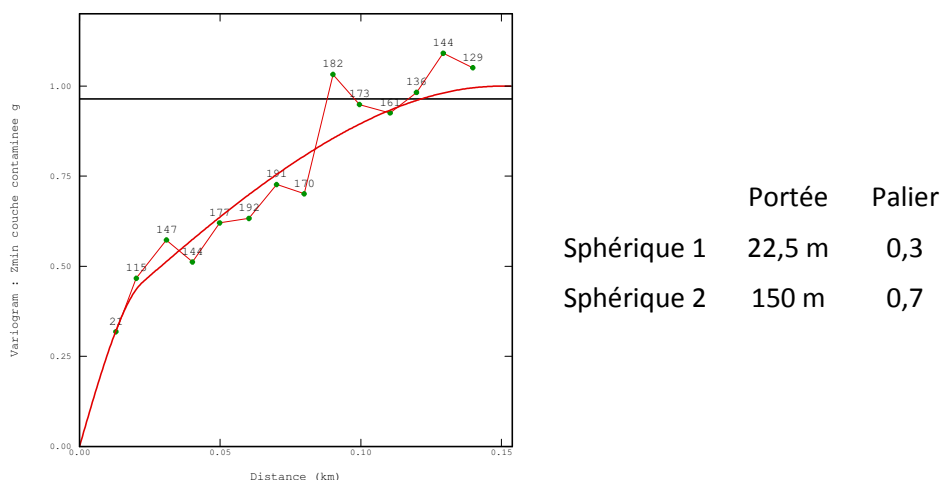


Figure 5 : Variogramme expérimental et modèle ajusté pour la transformée gaussienne du toit de la couche potentiellement contaminée. Indication du nombre de paires pour chaque point du variogramme expérimental et de la variance statistique (trait noir).

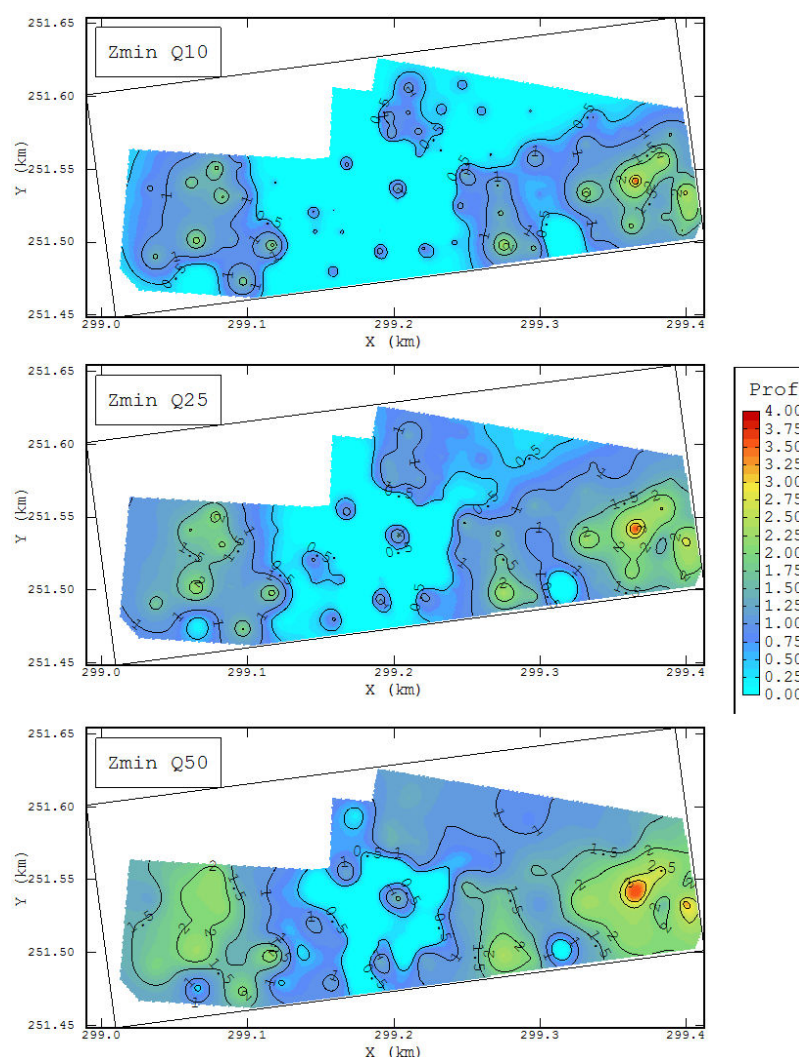


Figure 6 : Quantiles à 10, 25 et 50% pour la profondeur du toit de la couche potentiellement contaminée (en m). Indication des lignes d'iso-valeurs tous les 0,5 m.

3.2 Modélisation de la base de la couche potentiellement contaminée

La base de la couche potentiellement contaminée est obtenue en modélisant l'épaisseur et en lui ajoutant le toit de la couche précédemment modélisé, en considérant que toit et base sont indépendants. De plus, une épaisseur supplémentaire de 25 cm en-dessous des échantillons arrêtés à la base des remblais et ayant conduit à des teneurs supérieures à la valeur de présomption de 1 000 ppm en HCT a été ajoutée, afin d'intégrer la contamination probable du toit des argiles. Les figures ci-dessous illustrent le variogramme de la transformée gaussienne de l'épaisseur de la couche, puis les quantiles 50%, 75% et 90% des estimations par simulations géostatistiques de l'épaisseur et de la base de la couche. Ici, le quantile le plus sécuritaire, qui conduit à une couche plus épaisse, est le quantile 90%.

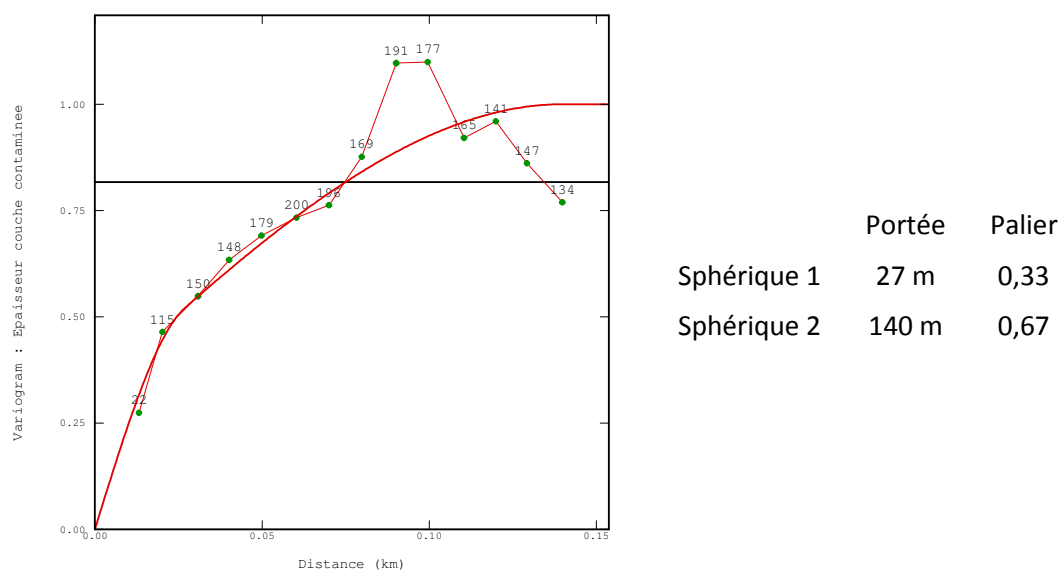


Figure 7 : Variogramme expérimental et modèle ajusté pour la transformée gaussienne de l'épaisseur de la couche potentiellement contaminée. Indication du nombre de paires pour chaque point du variogramme expérimental et de la variance statistique (trait noir).

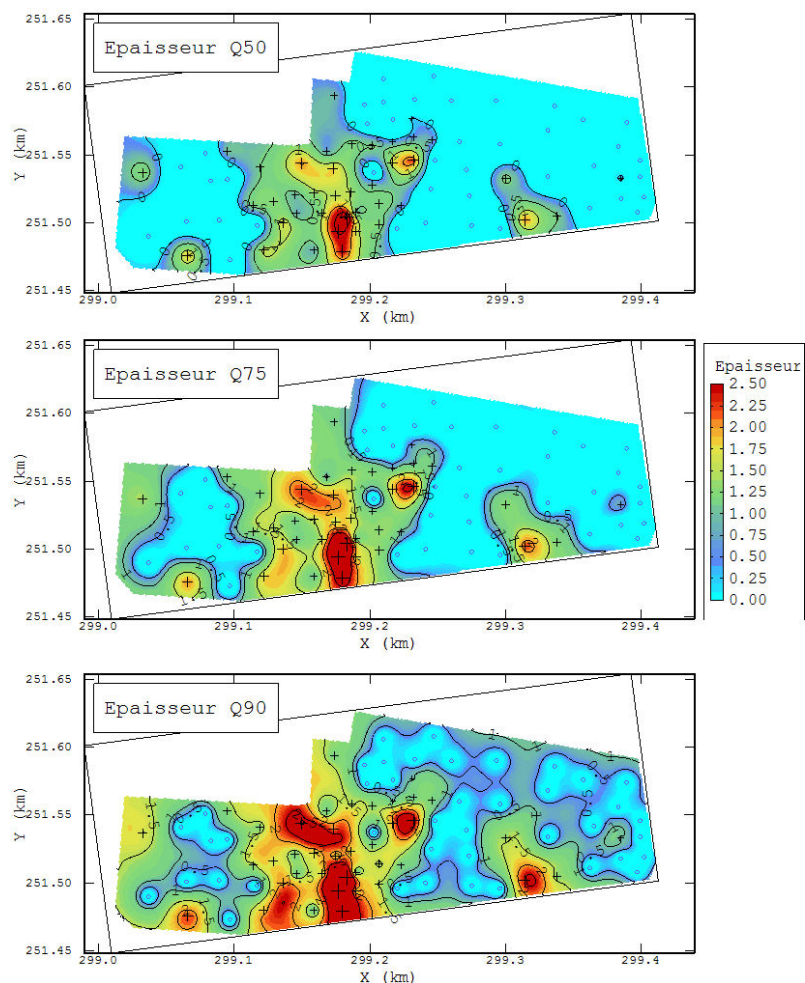


Figure 8 : Quantiles à 50, 75 et 90% pour l'épaisseur de la couche potentiellement contaminée (en m). Indication des lignes d'iso-valeurs tous les 0,5 m.

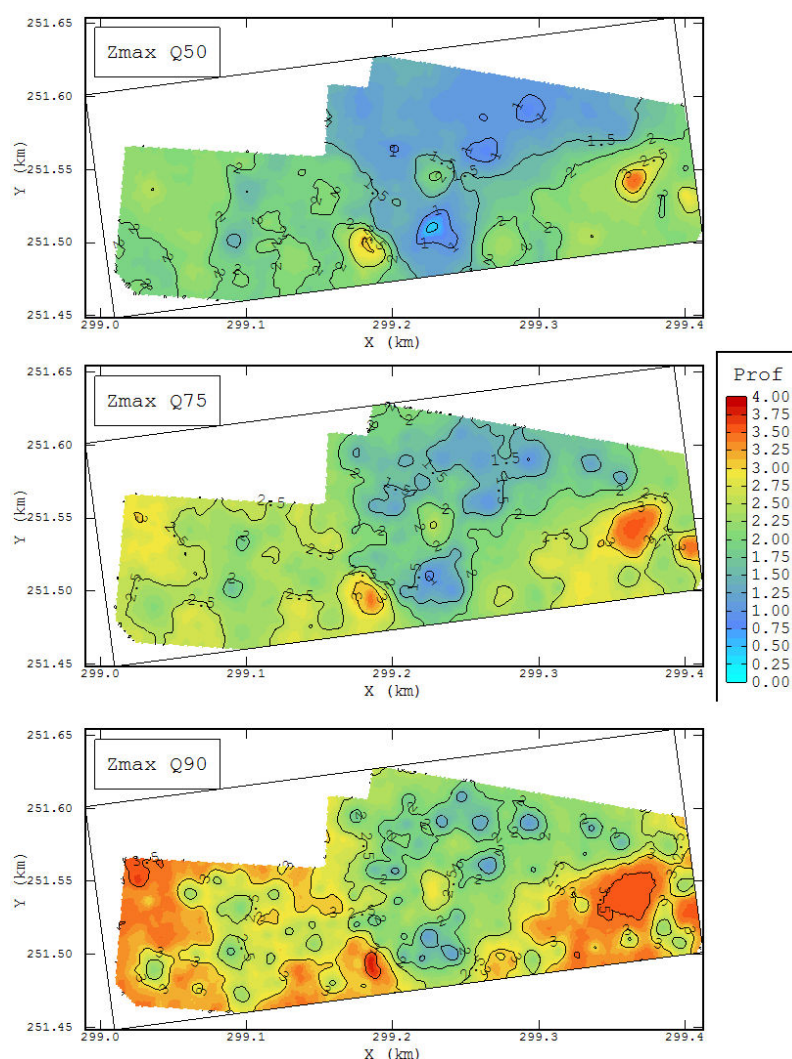


Figure 9 : Quantiles à 50, 75 et 90% pour la profondeur de la base de la couche potentiellement contaminée (en m). Indication des lignes d'iso-valeurs tous les 0,5 m.

3.3 Conclusion

En raison de l'incertitude sur la géométrie de la couche, un scénario, plus ou moins sécuritaire, devra être choisi, à la fois pour le toit et pour la base de la couche. Les volumes que l'on devra excaver pour atteindre les terres polluées seront en conséquence différents.

4. MODELISATION DES TENEURS EN HCT AU SEIN DE LA COUCHE CONTAMINEE REALISEE DEBUT 2006

Cette première modélisation des teneurs en HCT a été réalisée au début de l'année 2006, avant que les 17 sondages complémentaires ne soient effectués. Comme dit précédemment, les estimations ont été effectuées à une résolution de maille 1 m de façon à préparer d'éventuels contrôles ponctuels lors de la dépollution et à délimiter finement les zones polluées.

4.1 Simulations géostatistiques

200 simulations de la teneur moyenne en HCT sont réalisées, après modélisation de la structure spatiale de la transformée gaussienne des teneurs (Figure 10).

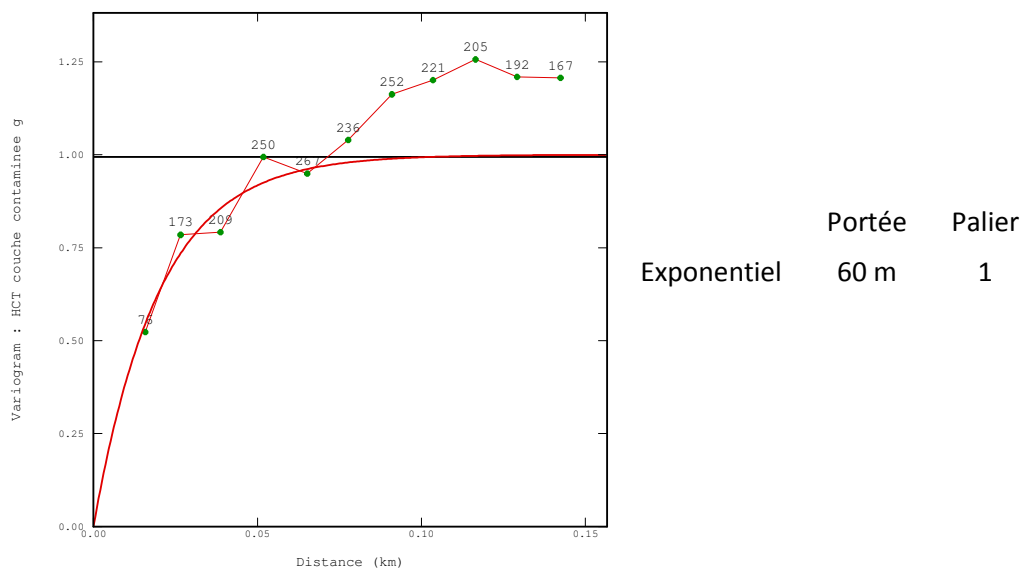


Figure 10 : Variogramme expérimental et modèle ajusté pour la transformée gaussienne des teneurs en HCT (début 2006) au sein de la couche potentiellement contaminée. Indication du nombre de paires pour chaque point du variogramme expérimental et de la variance statistique (trait noir).

4.2 Cartographie des teneurs et probabilité de dépassement du seuil de 2500 ppm

Il est possible de déduire des simulations le scénario le plus probable pour la contamination, qui est simplement en chaque maille la moyenne des valeurs simulées³. Cette cartographie fait clairement ressortir les zones présentant des fortes teneurs en HCT (Figure 11).

³ Les simulations ne sont pas nécessaires pour l'obtention de cette cartographie, qui peut directement être obtenue par krigeage. Néanmoins, nous intéressant surtout à l'incertitude liée au phénomène, les simulations sont directement réalisées, leur moyenne n'étant qu'une représentation probable du phénomène.

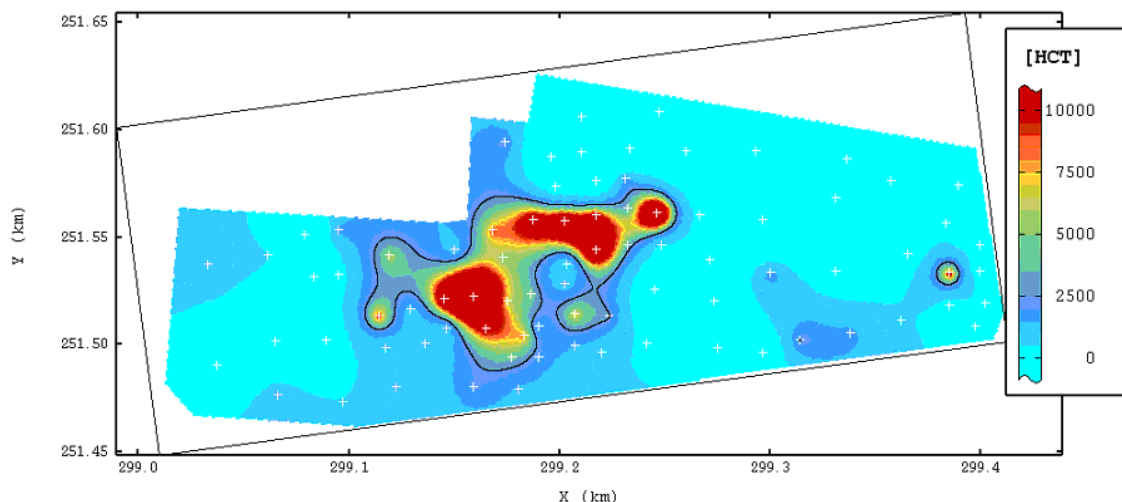


Figure 11 : Cartographie des teneurs en HCT (début 2006) au sein de la couche potentiellement contaminée (en ppm). Iso-contour à 2 500 ppm.

Se baser uniquement sur la cartographie la plus probable néglige l'incertitude associée à l'estimation. Il est donc nécessaire de raisonner en matière de risque de dépassement du seuil de 2 500 ppm. Cette probabilité, illustrée Figure 12, est obtenue en calculant en chaque maille la proportion de valeurs simulées (parmi les 200 simulations géostatistiques effectuées) qui dépasse 2 500 ppm.

Cette figure fait également apparaître les trois zones cibles (en gris) qui seront effectivement excavées. Les calculs de volumes menés par la suite concernent uniquement ces zones cibles.

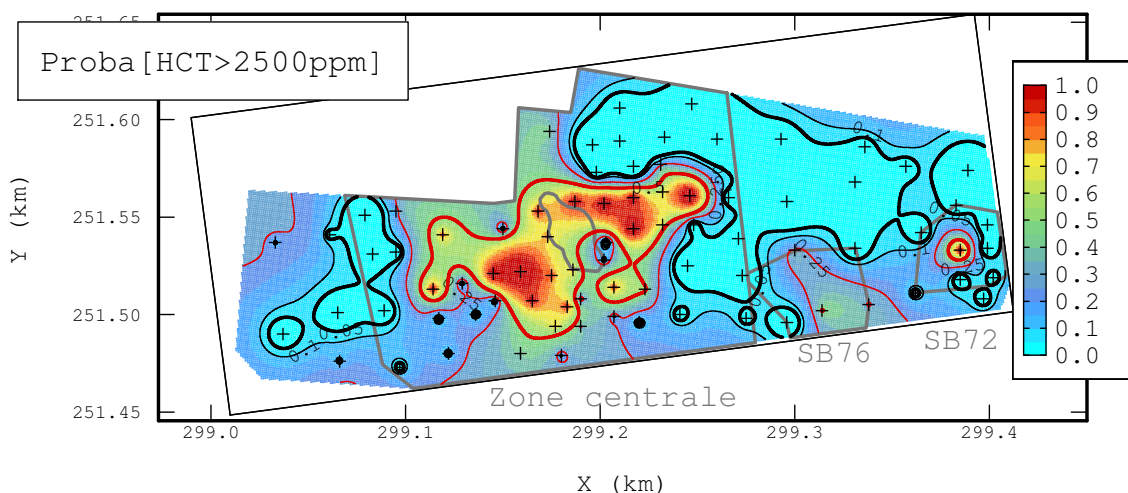


Figure 12 : Cartographie de la probabilité de dépassement de 2 500 ppm en HCT au sein de la couche potentiellement contaminée. Contours d'isoprobabilité à 5% (probabilité de 0,05 : trait fin noir), 10% (0,1 : trait épais noir), 25% (trait fin rouge) et 50% (trait épais rouge), contours des zones cibles pour les calculs de volumes (en gris).

Ces probabilités peuvent s'interpréter ainsi : par exemple, en réhabilitant toutes les mailles présentant un risque de dépassement de 2 500 ppm supérieur ou égal à 5 % (probabilité égale à 0,05), on accepte un risque au plus égal à 5 % de laisser en place une maille qui présenterait une teneur en HCT supérieure à 2 500 ppm. Selon le risque sanitaire accepté, la quantité de mailles à réhabiliter sera modifiée, une probabilité de 5 % correspondant à un choix sécuritaire.

4.3 Estimation des volumes contaminés et des volumes à excaver et comparaison avec les résultats de dépollution

Il est important de bien distinguer volumes pollués et volumes à excaver nécessaires pour accéder aux terres polluées. L'estimation des volumes contaminés doit donc répondre à deux questions :

- Quel est le volume de sol contaminé ? Cette question permet d'estimer globalement le niveau de pollution du site.
- Quel volume doit-on excaver si l'on veut être sûr d'extraire l'ensemble de la contamination ? Cette question concerne la mise en œuvre pratique de la réhabilitation.

4.3.1 Volume de sol contaminé

En croisant les simulations obtenues pour l'épaisseur de la couche potentiellement contaminée avec celles obtenues pour les teneurs en HCT au sein de cette couche, il est possible d'estimer la distribution des volumes de sols présentant des teneurs supérieures à 2 500 ppm : en effet, pour chacun des 200 couples épaisseur / teneur, on somme les épaisseurs obtenues aux mailles présentant une teneur supérieure à 2 500 ppm et on multiplie par la surface d'une maille. En procédant ainsi pour l'ensemble des simulations, on obtient la distribution des volumes contaminés de la Figure 13, prenant en compte une contamination probable des 25 premiers centimètres d'argile.

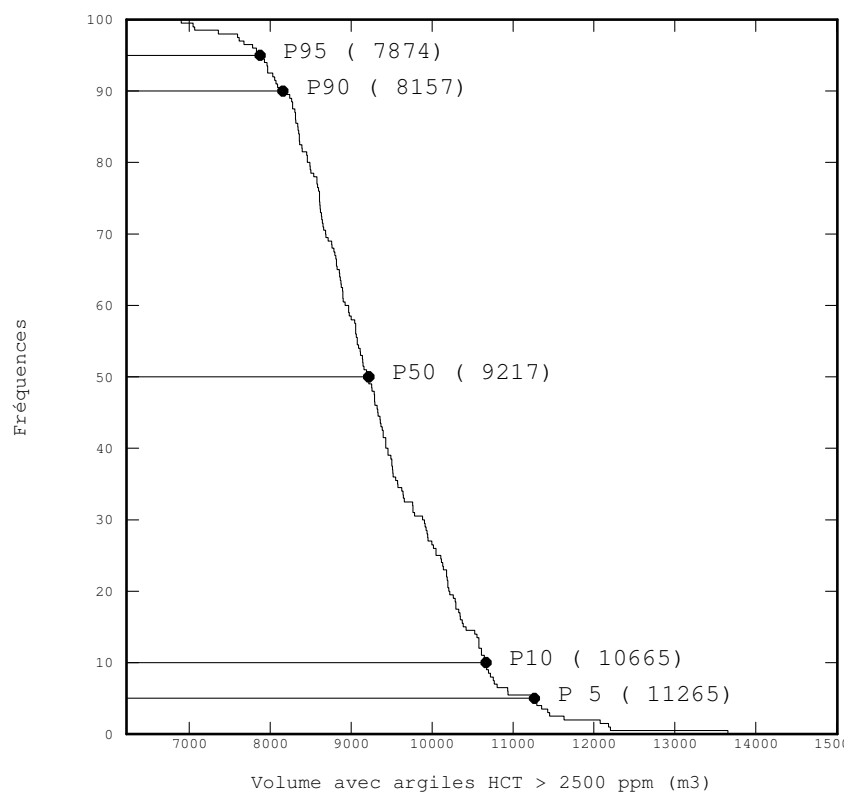


Figure 13 : Distribution (histogramme cumulé inverse) des volumes contaminés en HCT au sein des trois zones cibles incluant la contamination de 25 cm du toit des argiles ; indication des volumes correspondants aux quantiles à 5, 10, 50, 90 et 95 (en m³)

Ainsi le volume contaminé probable que l'on rencontrera sur le site est de 9 217 m³ et il est compris avec une probabilité de 90% dans l'intervalle de confiance [7 874 ; 11 265] m³. Cette première estimation est donc à comparer aux 13 171 m³ de terres réellement polluées excavées lors de la dépollution du site. On observe une grande différence entre ces deux volumes, le volume réel ne faisant même pas partie de l'intervalle de confiance à 90% des estimations. Le volume probable quant à lui sous-estime de 30% le volume réel.

La partie suivante revient sur les résultats d'estimation de volumes pollués présentés ici et détaille les deux raisons principales y conduisant :

- estimations réalisées à maille 1 m alors que l'excavation est réalisée à maille 15 m ;
- non intégration des 17 sondages réalisés en juin 2006 (après cette première étude).

Sachant cela, une nouvelle étude sera menée, intégrant ces éléments et conduisant à de nouveaux résultats.

4.3.2 Volume de sol à excaver

L'estimation du volume de sol à excaver doit donc répondre à la question suivante : quel volume de terres doit-on excaver afin de garantir (avec un niveau de confiance donné) que l'on a réhabilité les volumes contaminés en HCT précédemment estimés ? En raison :

- de l'incertitude sur la profondeur et l'épaisseur de la couche contaminée,
- de la variabilité spatiale des teneurs au sein de cette couche,

le volume à réhabiliter sera vraisemblablement notablement supérieur au volume effectivement contaminé, afin de s'affranchir d'un risque trop élevé de laisser en place des sols contaminés.

Ce raisonnement explique l'approche probabiliste proposée. Plutôt que de se baser sur le scénario le plus probable qui conduirait de toute évidence à laisser en place des terres souillées, il est préférable d'excaver dès que le risque que l'on soit dans la couche contaminée dépasse un certain seuil, par exemple 10 ou 25%.

Ce risque est appréhendé par :

- les quantiles à 10, 25 et 50% pour le toit de la couche (de sécuritaire à médian). Le quantile à 10% correspond à un choix où, pour chaque maille, on débute la réhabilitation dès que le risque d'être au sein de la couche potentiellement contaminée devient supérieur à 10% ;
- les quantiles à 50, 75 et 90% pour la base de la couche (de médian à sécuritaire). Contrairement au toit, c'est cette fois-ci considérer « profonde » une base de la couche potentiellement contaminée qui est sécuritaire ; cette position sécuritaire est décrite par le scénario Q90.

De même, le risque de dépassement du seuil de 2 500 ppm en HCT au sein de cette couche est appréhendé par plusieurs probabilités de 5% (sécuritaire) à 50% (choix relativement risqué, car on s'accorde une chance sur deux de laisser en place des terres contaminées).

Ces quantiles conduisent à de nombreux scénarii qui dépendent du risque que l'on est prêt à accepter. Trois scénarii conduisant à trois volumes à excaver sont ici commentés (

Tableau 1) :

- Le scénario probable correspond à un choix de Q50 pour le toit et la base de la couche contaminée, et un risque de 50% de laisser en place des mailles supérieures à 2 500ppm. Ce scénario conduit à un volume à excaver pour l'ensemble des trois zones d'environ 8 674 m³.
- En conservant les scénarii médians Q50 pour la géométrie de la couche mais en prenant un risque beaucoup plus faible de laisser en place des sols présentant des teneurs en HCT supérieures à 2500 ppm, par exemple de 10%, le volume à excaver passe à 18 108 m³.
- Il est recommandé de considérer des scénarii plus sécuritaires concernant l'extension de la couche, qui est relativement variable. En considérant ainsi un scénario Q25 pour le toit, Q75 pour la base et un risque de dépassement de 10% du seuil de 2500 ppm, le volume à excaver si l'on veut notablement réduire le risque de laisser en place des terres contaminées est égal à 33 755 m³.

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

Tableau 1 : Volumes à excaver pour différents scénarii de risque liés à la géométrie de la couche potentiellement contaminée ainsi qu'au dépassement de 2 500 ppm en HCT (résultats bruts prenant en compte une contamination probable du toit des argiles sur 25 cm).

Quantile considéré pour le toit de la couche	Quantile considéré pour la base de la couche	Quantile considéré pour l'estimation de la teneur en HCT	Volume à excaver
Q50 (probable)	Q50 (probable)	Q50 (probable)	8 674 m ³
Q50 (probable)	Q50 (probable)	Q10 (sécuritaire)	18 108 m ³
Q25 (sécuritaire)	Q75 (sécuritaire)	Q10 (sécuritaire)	33 755 m ³

Note : Étant donnée la forte sous-estimation des volumes pollués obtenue à l'issue de cette première étude, pour les deux raisons présentées précédemment (et détaillées dans la partie suivante), les volumes à excaver ne seront ici pas commentés par rapport au volume excavé réel de 22 348 m³.

Une fois un scénario choisi, par exemple Q25-Q75-Q10, les travaux pourront être menés de la façon suivante :

- L'extension horizontale de la zone à réhabiliter est fournie par la cartographie du risque de dépassement de 2500 ppm : c'est l'isocontour à 10% (traits fins noirs).
- Au sein de cette zone, les profondeurs à réhabiliter sont obtenues en considérant :
 - le scénario Q25 pour le toit de la couche potentiellement contaminée,
 - le scénario Q75 pour la base de la couche potentiellement contaminée.

5. MODELISATION DES TENEURS EN HCT AU SEIN DE LA COUCHE CONTAMINEE EN INTEGRANT LES SONDAGES DE JUIN 2006 ET EN PRENANT EN COMPTE LA MAILLE D'EXCAVATION

Suite à l'étude menée en 2006, des sondages complémentaires ont été réalisés, en accord avec l'étude géostatistique précédente, dans les zones les plus incertaines. Ces nouvelles investigations ont conduit à 17 sondages supplémentaires.

Les résultats obtenus en 2006 étaient calculés à partir de simulations réalisées sur des grilles à maille 1 m. Cependant la dépollution a utilisé une maille d'excavation de 15 m ; il est donc important aujourd'hui de considérer ce maillage de dépollution pour la réalisation de l'étude géostatistique.

La prise en compte de cet effet de support est détaillée en premier. Dans un second temps on considère en plus les 17 nouveaux sondages réalisés à l'issue de la première étude et avant la phase de dépollution. Les simulations et calculs de volumes sont réalisés après chaque étape de manière à mettre en évidence ce que modifie la prise en compte de chacun de ces deux éléments. La méthodologie employée est globalement la même que précédemment, seuls ces deux éléments étant modifiés.

5.1 Prise en compte du maillage d'excavation

5.1.1 Effet de support

Si la première étude géostatistique était adaptée pour se prévenir de contrôles ponctuels lors de la dépollution, pour délimiter finement la zone polluée et pour conseiller des sondages complémentaires, elle n'était en revanche pas adaptée au calcul de volumes pollués et de volumes à excaver pour prévoir la phase de dépollution qui est réalisée à maille 15 m.

La différence de volumes estimés que l'on fait selon la taille de la maille que l'on utilise provient de l'effet de support. Celui-ci est illustré sur la Figure 14 : elle présente les distributions d'une même variable estimée sur une grille à maille très fine (Échantillon) ou sur une maille plus large (Maille de dépollution). Les distributions diffèrent en raison de l'effet de moyenne ou régularisation qui s'opère sur la maille la plus large : la distribution correspondante est beaucoup plus resserrée autour de la moyenne et les valeurs extrêmes moins nombreuses (variance plus faible). On observe alors que lorsque l'on applique un seuil, les volumes calculés (parties hachurées) ne sont pas les mêmes selon la distribution que l'on considère. Selon la position du seuil relativement à la moyenne et selon les différences entre distribution il peut arriver que les volumes soient plus petits ou plus grands avec une maille fine qu'avec une maille large.

Cette figure met ainsi en évidence l'importance du choix de la maille d'estimation lors de la réalisation d'études géostatistiques. En effet, chaque maille d'excavation conduit à une distribution statistique différente et donc à des volumes différents une fois un seuil appliqué.

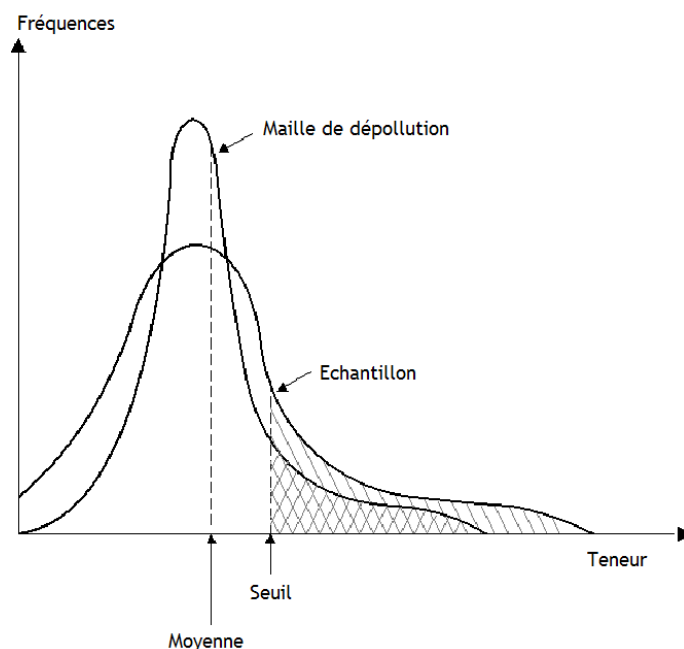


Figure 14 : Illustration de l'effet de support lors du calcul de volumes de terres polluées dépassant un seuil.

5.1.2 Probabilité de dépassement de seuil

En intégrant le maillage utilisé pour la dépollution et en utilisant la même méthodologie que pour la première étude géostatistique, une nouvelle carte de probabilité de dépassement du seuil de 2 500 ppm est obtenue (Figure 15).

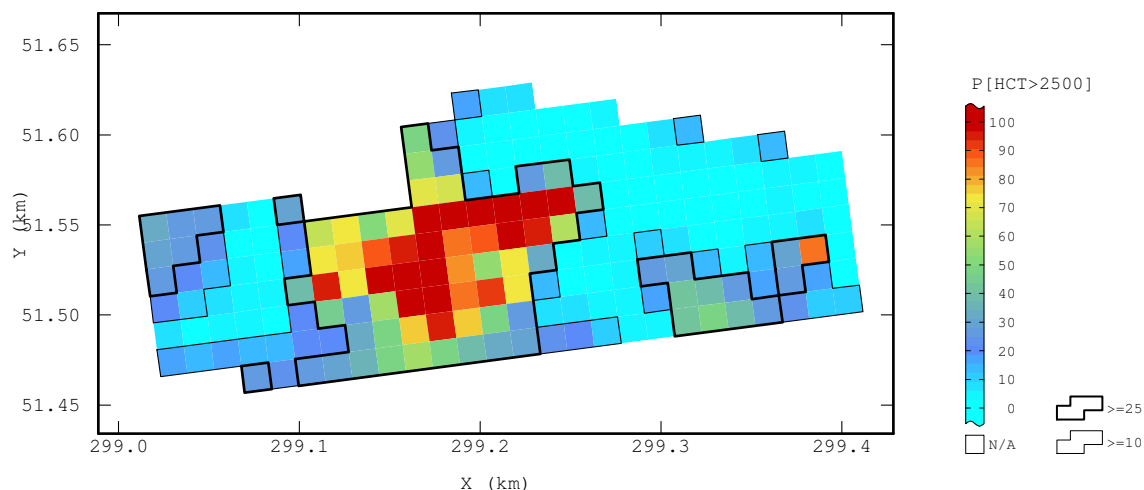


Figure 15 : Cartographie de la probabilité de dépassement de 2 500 ppm en HCT au sein de la couche potentiellement contaminée à maille 15 m. Contours d'isoprobabilité à 10% et à 25 %.

Si les principales zones contaminées sont les mêmes que sur la carte à maille 1 m, on observe par contre bien que la sélectivité est moins importante. Certaines mailles seront excavées entièrement alors que seule une partie sera polluée, d'autres seront délaissées entièrement alors qu'il y a une très faible probabilité qu'une partie de la maille soit contaminée. L'impact de la taille de maille utilisée lors de la phase d'excavation est clairement mis en évidence.

5.1.3 Estimation des volumes et comparaison avec les résultats de dépollution

Les volumes pollués calculés sont eux aussi modifiés, toujours en raison de l'effet de support. Le nouveau graphique présentant la distribution des volumes obtenus pour les 200 simulations géostatistiques effectuées (Figure 16) fait apparaître un volume pollué probable de 11 773 m³. L'intervalle de confiance à 90% est situé entre 9 498 et 14 726 m³. La différence avec le volume estimé précédemment avec la maille 1 m est donc de 2 556 m³ et la sous-estimation par rapport au volume pollué réel passe à 10,6 % alors qu'elle était auparavant de 30%. De plus, l'intervalle de confiance comprend maintenant la valeur réelle de 13 171 m³.

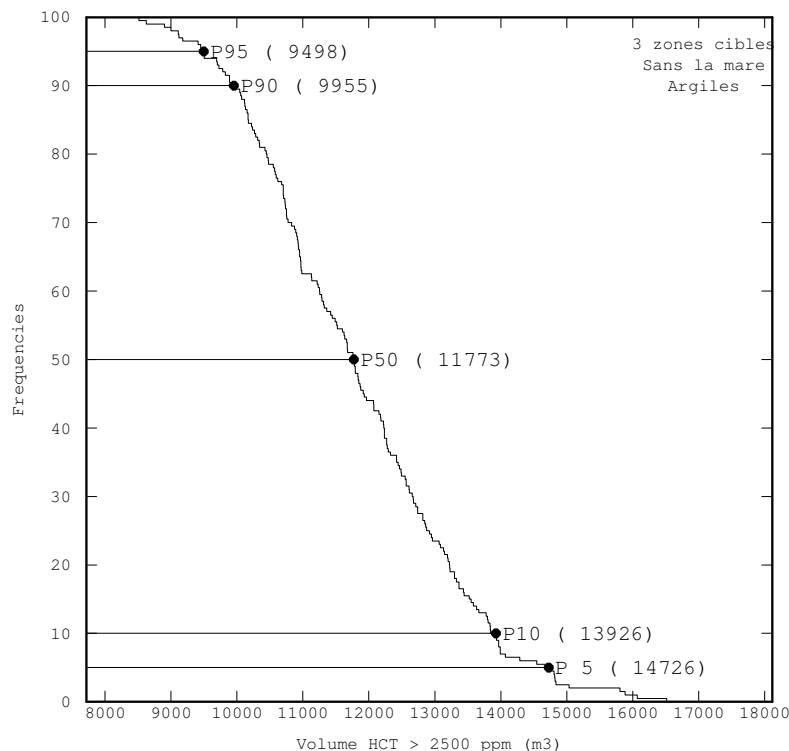


Figure 16 : Distribution (histogramme cumulé inverse) des volumes contaminés en HCT (estimés à maille 15 m) au sein des trois zones cibles incluant la contamination de 25 cm du toit des argiles ; indication des volumes correspondants aux quantiles à 5, 10, 50, 90 et 95%.

5.2 Prise en compte des dix-sept sondages complémentaires

La dernière étape consiste à non seulement effectuer les estimations sur une grille de maille 15 m mais aussi à intégrer les teneurs des 17 sondages effectués à l'issue de la première étude géostatistique. Les simulations sont de nouveau effectuées et les volumes recalculés, en utilisant les variogrammes de l'épaisseur de la couche contaminée et des teneurs en HCT mis à jour avec ces nouveaux sondages.

5.2.1 Probabilité de dépassement de seuil

Les 17 nouveaux sondages ont été réalisés dans les zones les plus incertaines. Il s'agit des zones pour lesquelles la probabilité de dépassement du seuil de 2 500 ppm n'est ni très faible (zones de non contamination quasi certaine) ni très élevée (zones de contamination quasi certaine). La Figure 17 superpose ainsi la carte de probabilité de dépassement de seuil obtenue précédemment aux nouveaux échantillons. Il est aisé d'observer qu'ils se situent sur des mailles pour lesquelles la probabilité de dépassement du seuil est située entre 30 et 60 %.

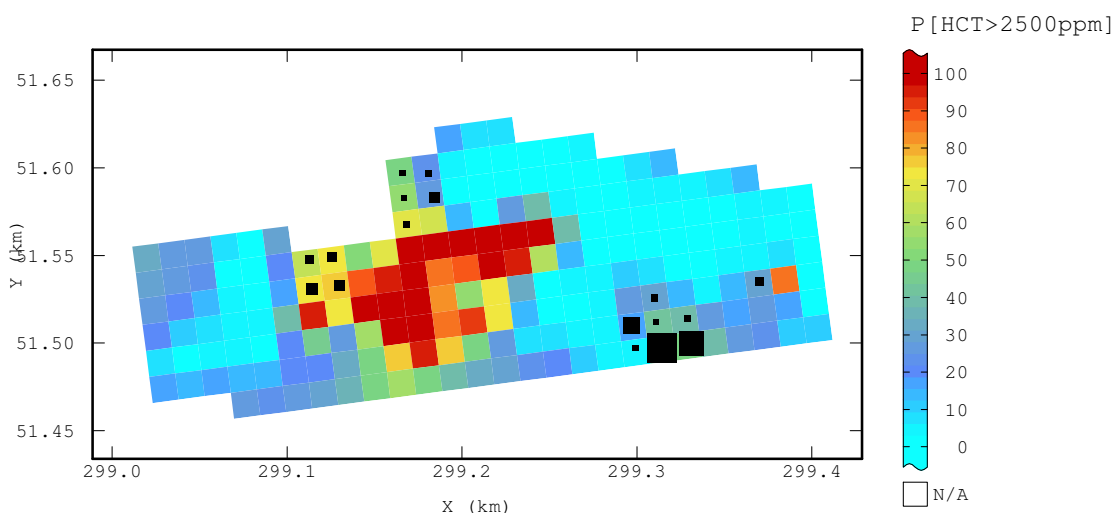


Figure 17 : Superposition de la cartographie de la probabilité de dépassement de 2500 ppm en HCT au sein de la couche potentiellement contaminée à maille 15 m et des teneurs des 17 nouveaux sondages (la taille des carrés noirs est proportionnelle à la teneur).

En incluant les nouveaux sondages une nouvelle carte de probabilité de dépassement de seuil est obtenue (Figure 18).

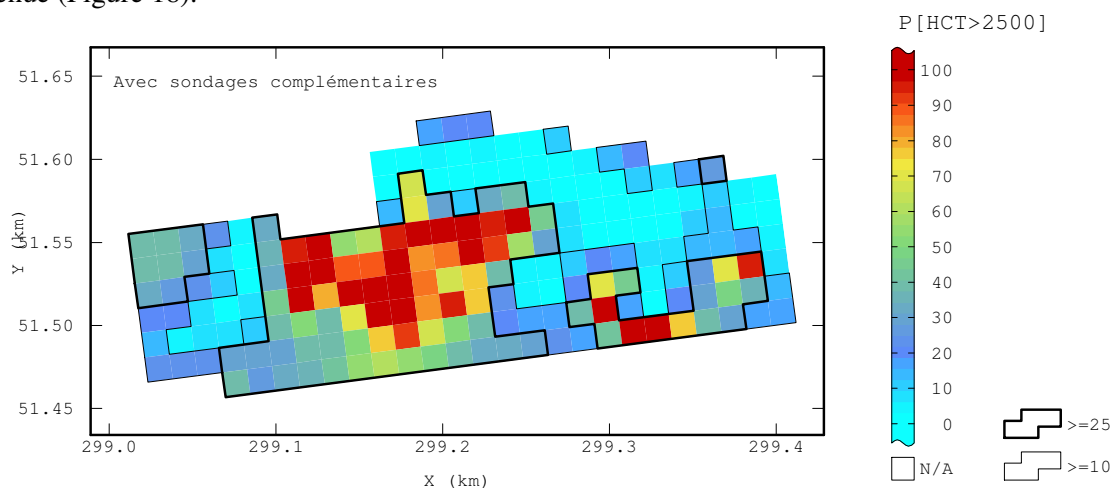


Figure 18 : Cartographie de la probabilité de dépassement de 2 500 ppm en HCT au sein de la couche potentiellement contaminée à maille 15 m et en incluant les nouveaux sondages. Contours d'isoprobabilité à 10% et à 25 %.

Les différences avec la carte précédente se situent principalement pour les mailles contenant les nouveaux sondages. La Figure 19 présente le nuage de corrélation entre les probabilités de dépassement de seuil maille à maille avec et sans les nouveaux sondages et fait ressortir les mailles pour lesquelles la probabilité est fortement modifiée. Ainsi les mailles bleues qui, auparavant, avaient des probabilités de dépassement comprises entre 20 et 70% ont maintenant des probabilités vraiment très faibles de dépasser le seuil. Au contraire, les mailles grises qui avaient des probabilités situées aux alentours de 50 % s'avèrent être polluées de manière quasi-certaine. La modification des probabilités intermédiaires grâce aux sondages complémentaires souligne bien l'apport d'information de ces derniers.

L'intérêt de la première étude géostatistique a donc aussi été d'orienter un second échantillonnage permettant de réduire les zones d'incertitude. Cela démontre aussi les bénéfices d'une démarche itérative pour l'échantillonnage avec une première reconnaissance aboutissant à des recommandations pour un second échantillonnage. Lorsque cette démarche est compatible avec la durée consacrée à l'échantillonnage d'un site elle permet d'améliorer les estimations géostatistiques ainsi que très souvent de diminuer les coûts de dépollution grâce à une meilleure connaissance finale du site.

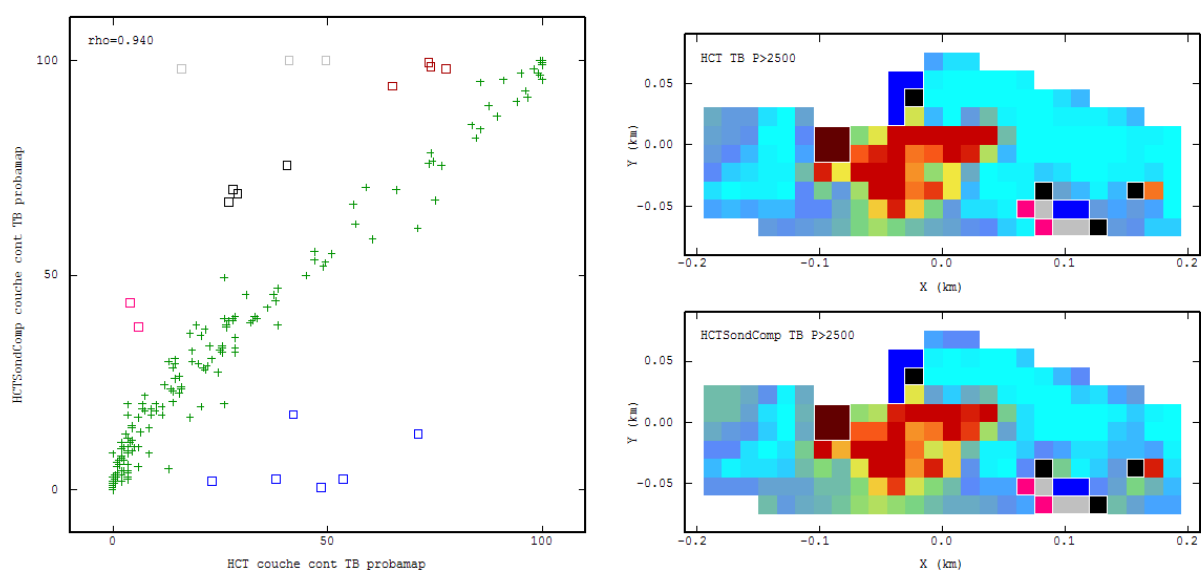


Figure 19 : Nuage de corrélation entre probabilités de dépassement de seuil avant et après intégration des nouveaux sondages et carte correspondantes. Les couleurs des mailles mises en évidence sont les mêmes sur le nuage de corrélation et sur les cartes.

5.2.2 Estimation des volumes pollués et comparaison aux résultats de dépollution

L'intégration des nouveaux sondages conduit aux volumes pollués suivants (Figure 20) :

- volume probable de 12 059 m³ ;
- intervalle de confiance à 90% compris entre 10 028 m³ et 15 421 m³.

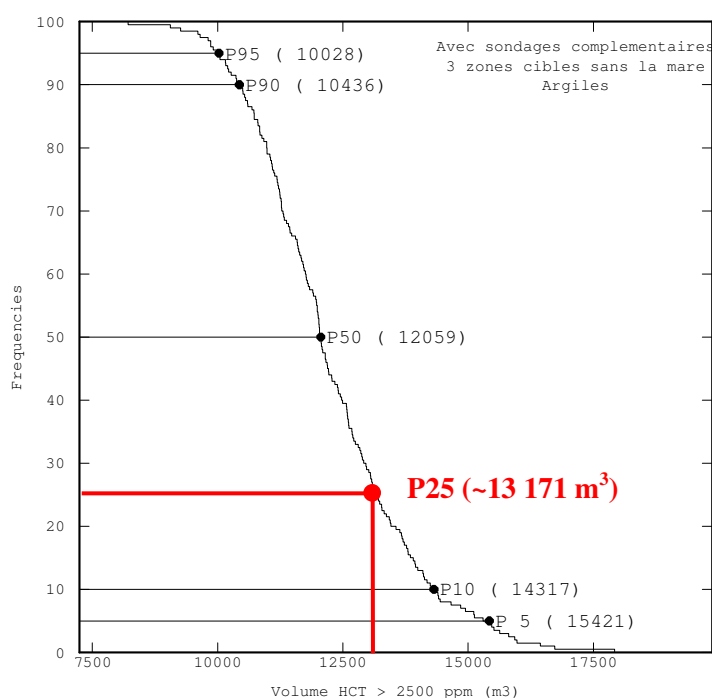


Figure 20 : Distribution (histogramme cumulé inverse) des volumes contaminés en HCT (estimés à maille 15 m et incluant les nouveaux sondages) au sein des trois zones cibles incluant la contamination de 25 cm du toit des argiles ; indication des volumes correspondants aux quantiles à 5, 10, 50, 90 et 95%.

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

Le volume probable est encore plus proche du volume pollué réel qu'il ne sous-estime plus que de 8,4 %. Si l'on considère un scénario plus sécuritaire et que l'on s'intéresse au quantile 25% on trouve alors presque exactement le volume pollué réel de 13 171 m³.

Ainsi lorsque l'on considère une maille identique à la maille utilisée lors de la dépollution et que l'on intègre toutes les informations disponibles sur le site, l'estimation des volumes par des méthodes géostatistiques montre sa pertinence.

5.2.3 Estimation des volumes à excaver et comparaison aux résultats de dépollution

Si l'on prend maintenant en compte l'incertitude due à la géométrie de la couche et que l'on calcule les volumes qu'il sera nécessaire d'excaver pour accéder aux terres polluées, un choix doit être effectué : quel risque accepte-t-on de prendre, à la fois en ce qui concerne la géométrie de la couche et en ce qui concerne les teneurs.

Différents scénarii sont présentés dans le Tableau 2 ; ils correspondent à des situations plus ou moins sécuritaires. Les volumes qui en découlent vont de 14 112 m³, soit à peine plus que le volume pollué, lorsque que l'on considère le scénario le plus probable à la fois pour la géométrie de la couche et pour l'estimation des teneurs en HCT, à presque 35 000 m³ lorsque l'on considère le scénario le plus sécuritaire.

Tableau 2 : Volumes à excaver pour différents scénarios de risque liés à la géométrie de la couche potentiellement contaminée ainsi qu'au dépassement de 2500 ppm en HCT (résultats bruts prenant en compte une contamination probable du toit des argiles sur 25 cm, la maille de dépollution et les nouveaux sondages).

Quantile considéré pour le toit de la couche	Quantile considéré pour la base de la couche	Quantile considéré pour l'estimation de la teneur en HCT	Volume à excaver
Q50 (probable)	Q50 (probable)	Q50 (probable)	14 112 m ³
Q25 (sécuritaire)	Q75 (sécuritaire)	Q50 (probable)	22 160 m ³
Q50 (probable)	Q50 (probable)	Q25 (sécuritaire)	17 750 m ³
Q25 (sécuritaire)	Q75 (sécuritaire)	Q25 (sécuritaire)	31 239 m ³
Q50 (probable)	Q50 (probable)	Q10 (très sécuritaire)	18 414 m ³
Q25 (sécuritaire)	Q75 (sécuritaire)	Q10 (très sécuritaire)	34 736 m ³

Si l'on considère le scénario avec le quantile 25% pour les teneurs en HCT (valeur la plus proche du volume pollué réel), les volumes à excaver sont de 17 750 m³ pour la géométrie la plus probable et de 31 239 m³ pour la géométrie la plus sécuritaire, ce qui englobe le volume excavé réel de 22 348 m³. Selon le risque choisi au départ, le volume à excaver est ici sous ou surestimé et le volume réel est compris entre deux scénarii.

5.2.4 Comparaison des mailles polluées estimées et excavées

L'excavation a principalement été réalisée à maille 15 m. Cependant il est difficile d'utiliser les résultats de dépollution maille par maille pour deux raisons :

- les limites de maille retenues ne sont pas connues de manière précise ;
- l'excavation a commencé par les parties les plus contaminées, là où la pollution était bien connue puis a été faite progressivement au fur et à mesure des contrôles en bordure de fouille. Cela conduit à des zones excavées selon d'autres géométries et d'autres dimensions que la maille de 15 x 15 m considérée dans l'étude (Figure 21).

Seule une comparaison maille à maille au centre de la zone la plus polluée serait donc possible. S'agissant de la zone pour laquelle les incertitudes sont les plus faibles (Figure 22), la comparaison serait favorable et biaisée puisque ne prenant pas en compte les zones incertaines pour lesquelles la probabilité de dépassement du seuil est entre 40 et 60%.

Le choix a donc été fait de superposer les figures présentant d'une part l'excavation et d'autre part la probabilité de dépassement du seuil.

La Figure 21 reprend ainsi en fond l'excavation telle qu'elle a été réalisée, sur lequel ont été superposés deux isocontours pour les probabilités de dépassement de seuil à 25 et 50%. Sur la Figure 22 sont représentés la carte de probabilité de dépassement de seuil (identique à la Figure 18) ainsi que le contour de la zone effectivement excavée et les passes successives.

Les deux figures montrent une très bonne concordance entre les estimations réalisées et les terres ayant dû être excavées. Le contour excavé est proche du contour de probabilité de dépassement à 50%. L'excavation n'ayant pas toujours été réalisée à maille 15 m on observe aisément, pour quelques mailles à la frontière, qu'elle a pu être interrompue avant et que ces situations correspondent à des probabilités de dépassement intermédiaires, souvent autour de 40%. Enfin, les plus grandes différences sont observées pour les mailles décrites comme étant bitumineuses et présentant une problématique différente. Celles-ci ont donc été traitées à part lors de l'excavation : présentant une pollution très localisée, l'excavation y a été réalisée très finement.

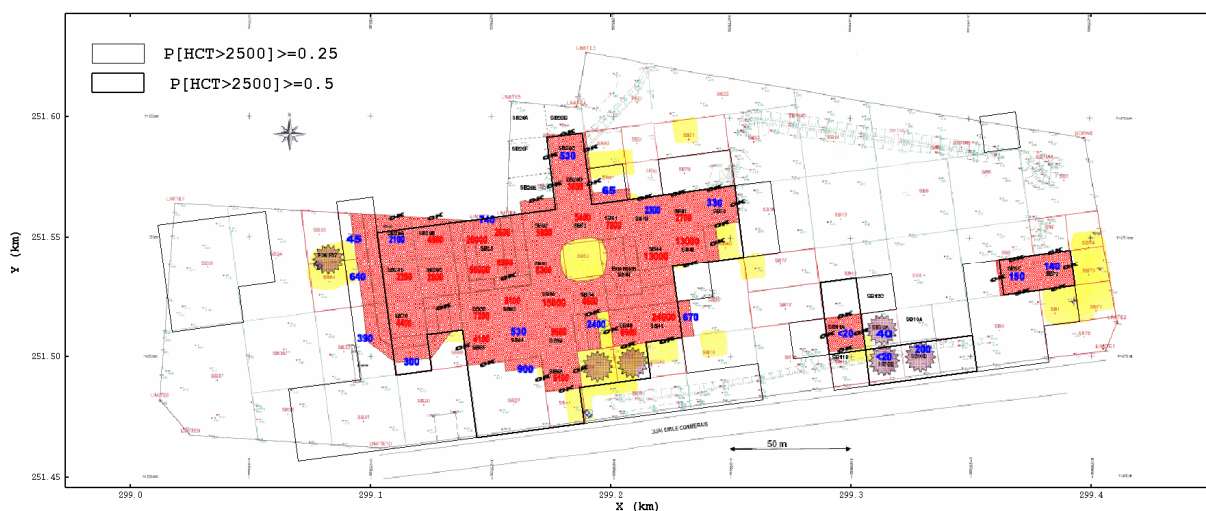


Figure 21 : Superposition de la zone réhabilitée (en rouge) aux contours de probabilités de dépassement de seuil à 25 et 50%.

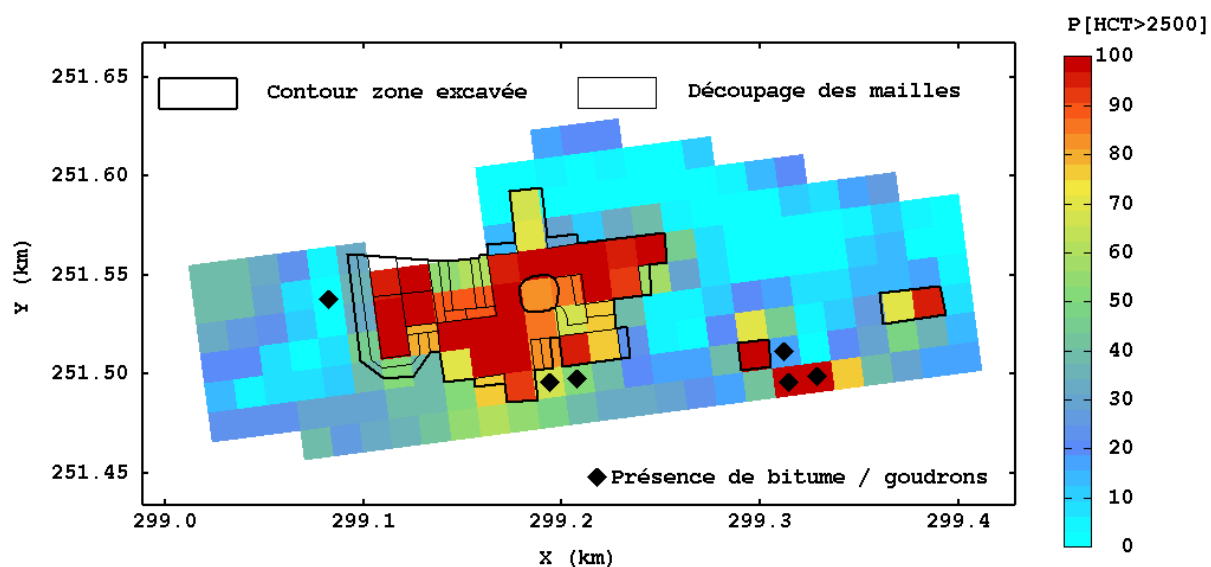


Figure 22 : Superposition de la carte de probabilité de dépassement du seuil de 2500 ppm en HCT et des contours de la zone excavée ainsi que des mailles irrégulières utilisées.

6. CONCLUSION

L'étude géostatistique avait pour objectifs la cartographie de la pollution en HCT sur le site ainsi que l'évaluation des volumes de terres contaminées et à excaver en vue d'un chiffrage du coût de la dépollution du site.

L'intérêt d'une telle approche est de baser l'évaluation des coûts liés aux volumes contaminés sur les données disponibles, en intégrant cependant l'incertitude liée à l'hétérogénéité spatiale de la contamination et à la reconnaissance uniquement parcellaire de celle-ci.

Les conclusions basées sur l'analyse des données de la campagne de décembre 2005 ont montré que cette première étude géostatistique réalisée à maille fine était adaptée pour prévenir d'éventuels contrôles ponctuels, pour délimiter les zones polluées et pour conseiller la réalisation de nouveaux sondages dans les zones les plus incertaines. En revanche la maille d'estimation étant très différente de la maille retenue pour l'excavation, le calcul des volumes pollués n'est pas pertinent. En effet, en l'absence de connaissance de la maille retenue lors de l'excavation, aucune quantification des volumes ne peut être réalisée sans biais, l'estimation des volumes pollués au-delà d'un certain seuil dépendant du support d'estimation choisi.

Pour cela une deuxième étude a été réalisée. Intégrant de plus les nouveaux sondages réalisés à la suite de la première étude géostatistique, elle a montré la pertinence de la géostatistique pour l'estimation des volumes pollués. Le volume pollué réel de 13 171 m³ est en effet proche de la valeur estimée comme étant la plus probable (12 059 m³), correspond au quantile 25% des valeurs estimées et est donc largement inclus dans l'intervalle de confiance à 90% construit à partir des estimations.

L'étude n'a pas comporté d'analyse de la maille d'excavation la plus pertinente dans la mesure où celle-ci était fixée au moment de l'étude géostatistique. Dans le cadre d'une étude intervenant plus en amont, la géostatistique peut également permettre d'aider à définir la maille de traitement la plus adaptée, en prenant en compte les différentes contraintes de terrain et de manière à optimiser le ratio entre les coûts d'utilisation d'une maille très fine et le manque de sélectivité lié à l'usage d'une maille trop grande.

Après l'estimation du volume pollué, il est également nécessaire de calculer un volume qu'il conviendra d'excaver pour atteindre toute la pollution. En effet, en raison de la variabilité spatiale des teneurs et de l'incertitude liée à la profondeur réelle de la couche contaminée en HCT, il est recommandé de ne pas baser l'excavation sur l'estimation du volume pollué uniquement, en raison du risque important qu'il présente de laisser en place des sols contaminés. Ce risque découle à la fois :

- de la méconnaissance précise de l'extension verticale des remblais contaminés,
- de la variabilité spatiale latérale des teneurs en HCT au sein de cette couche contaminée.

Un scénario plus sécuritaire est recommandé. Il consiste à réhabiliter la couche potentiellement contaminée dès qu'il y a un certain risque, à définir, de s'y trouver. L'extension latérale de cette couche est quant à elle définie par un niveau de risque sécuritaire de dépassement du seuil de 2 500 ppm en HCT, à définir également.

En considérant le quantile 25% des estimations pour le volume pollué ainsi que le scénario le plus probable et le scénario sécuritaire (qui considère les quantiles 25% pour la profondeur du toit de la couche et 75% pour la base) pour la géométrie, un encadrement du volume à excaver entre 17 750 m³ et 31 239 m³ est réalisé. Cet intervalle comprend le volume réel excavé de 22 348 m³.

Il est évident que procéder de la sorte conduira à l'excavation de volumes non négligeables de terres saines, qui pourraient éventuellement être triées en cours de dépollution. Prévoir la réhabilitation de tels volumes permet cependant de s'affranchir du risque de laisser en place des terres contaminées et de beaucoup mieux prévoir les coûts de dépollution.

Enfin, la comparaison maille à maille entre la zone réellement excavée et les prédictions a été réalisée en superposant les cartes. Elle montre une très bonne concordance entre l'étude géostatistique et les

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

zones réellement excavées, confirmant la qualité des résultats globaux précédents et donc la pertinence de l'approche géostatistique mise en œuvre sur ce site.

Le tableau ci-dessous reprend les principales étapes réalisées pour cette étude de démonstration en mettant en avant les points spécifiques à cette étude et les étapes courantes d'une étude de caractérisation d'un site pollué.

Tableau 3 : Principales étapes réalisées pour l'étude de démonstration et pertinence de ces étapes dans le cadre d'une étude géostatistique classique.

Étape	Commentaire
Analyse de l'échantillonnage et des informations fournies	Indispensable, elle a conduit ici à de nombreuses remarques sur les spécificités de l'échantillonnage et a eu des conséquences sur tout le processus de traitement des données.
Préparation des données	Déoulant de l'étape précédente, des moyennes pondérées ont été calculées.
Analyse statistique	Elle permet un contrôle qualité des données et les premiers commentaires sur les niveaux de contamination rencontrés.
Modélisation géométrique de la couche contaminée	Spécifique à cette étude, en raison de l'échantillonnage réalisé. Elle a consisté à simuler la couche potentielle au sein de laquelle se trouve la pollution.
Modélisation des teneurs en HCT au sein de cette couche	2ème étape de modélisation (due à la méthodologie particulière retenue ici) visant à simuler les teneurs en HCT à 2D, au sein de la couche potentiellement contaminée préalablement modélisée.
Cartographies de probabilité de dépassement de seuil	Approche classique permettant de déterminer les zones qui devront être traitées en fonction d'un risque accepté choisi de laisser des terres contaminées en place.
Estimation des volumes contaminés et volumes à excaver	La distinction entre ces deux types de volumes est primordiale et chacun répond à une question différente. Ils sont classiquement calculés dans une étude de caractérisation.
Mise à jour de l'étude une fois la maille d'excavation connue	La maille d'excavation est souvent connue au moment de l'étude géostatistique et sa prise en compte doit être faite le plus tôt possible.
Mise à jour de l'étude pour prendre en compte les sondages complémentaires	Très intéressante, la démarche itérative présente un grand intérêt en termes de connaissance de la pollution. Elle gagne à être utilisée lors des étapes de caractérisation d'un site.
Comparaison aux mailles excavées	Dans le cadre de cette étude de démonstration, un retour d'expérience – qui peut rarement être réalisé car les données ne sont en général pas connues – a pu être fait. Cette comparaison a montré l'adéquation entre l'étude géostatistique et l'excavation réalisée.

7. ANNEXE A : DONNEES DES CAMPAGNES DE 2005 ET 2006

Les tableaux ci-dessous reprennent l'ensemble des données utilisées dans l'étude. Les coordonnées et profondeurs sont fournies en m et les teneurs en mg/kg sec.

Tableau 4 : Données de 2005 utilisées pour l'étude géostatistique : numéro de sondage, coordonnées X, Y en Lambert II, cote du sol Zsol en IGN69, profondeurs de la base des remblais et du toit des argiles, niveaux échantillonnés (min et max) et teneurs en HCT correspondantes (en ppm).

N	X	Y	Zsol	Base remblais	Toit argiles	Zmin A	Zmax A	Zmin B	Zmax B	ZmaxB avec argiles	HCT A	HCT B
1	299385	251518	4,83	2,00	2,00	0	0,6	0,6	2,3	2,30	160	220
2	299383	251556	4,99	2,60	2,60	0	1,0	1,0	2,8	2,80	140	200
3	299389	251574	4,90	1,90	1,90	0	1,0	1,0	2,1	2,10	33	10
4	299362	251511	4,65	2,20	2,20	0	1,0	1,0	2,3	2,30	190	580
5	299365	251542	4,15	3,50	3,50	0	1,0	1,0	2,5	2,50	10	270
6	299357	251576	3,80	1,10	1,10	0	0,5	0,5	1,3	1,30	200	220
7	299336	251586	3,80	1,10	1,10	0	1,0	1,0	1,4	1,40	41	580
8	299331	251568	4,15	1,50	1,50	0	1,0	1,5	1,7	1,70	10	10
9	299331	251534	4,45	2,40	2,40	0	0,5	0,5	2,4	2,40	250	96
10	299338	251505	4,90	2,50	2,50	0	1,4	1,4	2,5	2,75	47	1900
11	299296	251496	4,60	2,00	2,00	0	1,0	1,0	2,0	2,00	110	90
12	299300	251533	4,10	2,20	2,20	0	1,0	1,0	2,2	2,45	180	2000
13	299296	251558	4,15	1,50	1,50	0	1,5	1,5	1,7	1,70	10	10
14	299293	251590	3,60	0,80	0,80	0	0,8	0,8	1,1	1,10	140	16
15	299260	251590	3,80	1,00	1,00	0	1,0	1,0	1,2	1,20	10	10
16	299266	251560	3,75	0,80	0,80	0	0,6	0,6	1,5	1,50	340	160
17	299273	251520	4,35	2,00	3,00	0	2,0	2,0	2,8	2,80	120	10
18	299275	251498	4,50	2,50	2,50	0	1,5	1,5	2,3	2,30	240	450
19	299241	251500	4,00	0,90	0,90	0	0,4	0,4	0,9	0,90	37	400
20	299245	251525	4,40	1,30	2,50	0	1,3	1,3	2,5	2,50	160	10
21	299233	251591	4,30	1,00	1,00	0	0,3	0,3	1,4	1,40	140	30
22	299247	251608	4,40	1,00	1,00	0	1,0	1,0	2,0	2,00	130	10
23	299210	251606	4,40	1,40	1,40	0	1,4	1,4	1,6	1,60	140	10
24	299210	251589	4,37	1,20	1,20	0	1,2	1,2	1,7	1,70	120	41
26	299174	251594	4,41	0,80	0,80	0,2	0,8	0,8	1,1	1,10	1100	4200
27	299159	251480	5,23	2,00	2,00	0,4	1,1	1,1	2,0	2,25	61	2100
28	299150	251544	4,53	2,10	2,10	0	1,3	1,3	2,0	2,00	1300	330
29	299119	251541	4,30	1,90	1,90	0	0,8	0,8	1,9	2,15	300	7000
30	299122	251480	5,42	2,00	2,00	0,4	1,6	1,6	2,0	2,00	1200	480
31	299097	251473	5,40	2,10	2,10	0	1,0	1,0	2,0	2,00	730	61
32	299089	251502	5,00	1,40	2,30	0	1,4	1,4	2,3	2,30	170	130
33	299079	251551	4,42	2,20	2,20	0	1,2	1,2	2,2	2,20	10	110
34	299061	251541	4,68	2,10	2,30	0,2	1,3	1,3	2,3	2,30	460	420
35	299065	251501	5,00	2,35	2,35	0,1	1,0	1,0	2,4	2,40	66	120
36	299066	251476	5,10	2,20	2,20	0,3	0,8	0,8	2,2	2,20	1300	690
37	299037	251490	5,00	2,00	2,00	0,2	1,0	1,0	2,0	2,00	43	170
38	299033	251537	4,90	2,50	2,50	0,1	1,2	1,2	2,4	2,75	40	1400

N	X	Y	Zsol	Base remblais	Toit argiles	Zmin A	Zmax A	Zmin B	Zmax B	ZmaxB avec argiles	HCT A	HCT B
39	299246	251561	4,20	1,40	1,40	0	0,5	0,5	1,4	1,65	150	25000
40	299248	251546	3,70	1,60	1,60	0	0,8	0,8	1,8	1,80	120	410
41	299232	251563	4,10	0,80	0,80	0	0,8	0,8	1,5	1,50	190	3200
42	299232	251546	4,00	2,10	2,10	0	1,1	1,1	2,2	2,35	3000	3400
43	299217	251560	4,85	1,30	1,30	0	1,0	1,0	1,5	1,50	980	12000
44	299217	251544	4,20	1,40	1,60	0	1,6	1,6	1,7	1,85	22000	9800
45	299223	251513	4,30	0,40	1,70	0	0,8	0,8	1,5	1,50	2700	290
46	299220	251496	4,29	1,20	2,40	0	1,2	1,2	2,4	2,40	940	34
47	299207	251499	4,37	1,25	2,50	0	1,1	1,1	2,5	2,50	1600	320
48	299207	251514	4,32	1,50	1,50	0	0,3	0,3	1,5	1,75	490	8100
49	299202	251528	4,68	1,70	1,70	0	0,7	0,7	1,7	1,95	420	1300
50	299203	251537	4,55	1,70	1,70	0	0,8	0,8	1,7	1,70	470	320
51	299202	251557	4,20	1,10	1,50	0	1,1	1,1	1,5	1,50	21000	80
52	299187	251558	3,80	0,80	0,80	0	0,8	0,8	1,2	1,20	13000	19000
54	299186	251523	4,90	1,50	1,50	0	0,5	0,5	1,5	1,75	2000	4600
55	299190	251508	4,08	1,50	1,50	0	0,5	0,5	1,5	1,75	470	1700
56	299190	251494	4,85	1,40	3,00	0	1,4	1,4	3,0	3,00	730	2000
57	299180	251479	5,23	1,40	2,70	0	1,4	1,4	2,7	2,70	1700	1400
58	299177	251494	5,30	1,90	3,00	0	1,9	1,9	3,0	3,00	3700	3400
59	299183	251504	5,10	3,00	3,00	0	0,8	0,8	3,0	3,00	3200	9900
60	299175	251520	4,46	1,90	1,90	0	0,6	0,6	1,5	2,15	1200	12000
61	299173	251540	4,38	1,80	1,80	0	1,3	1,3	1,8	2,05	2600	11000
62	299168	251553	4,45	2,10	2,10	0	1,3	1,3	2,1	2,35	10	8200
63	299159	251522	4,45	2,00	2,00	0	0,6	0,6	1,9	2,25	630	28000
64	299165	251507	4,80	2,20	2,20	0	0,9	0,9	2,2	2,45	110	11000
65	299146	251507	4,65	1,80	1,80	0	0,9	0,9	1,7	2,05	47	1200
66	299145	251521	4,35	2,10	2,10	0	1,1	1,1	2,1	2,35	160	17000
67	299129	251516	4,50	1,80	1,80	0	0,5	0,5	1,8	2,05	620	1300
68	299136	251500	4,90	2,00	2,00	0,3	0,8	0,8	2,0	2,25	1400	1100
69	299117	251498	5,00	2,20	2,20	0	0,9	0,9	2,2	2,20	90	990
70	299114	251513	4,70	2,20	2,20	0	1,0	1,0	2,2	2,45	200	9900
71	299402	251519	4,83	2,50	2,50	0	1,1	1,1	2,5	2,50	54	470
72	299385	251533	4,80	2,00	2,00	0	1,5	1,5	2,0	2,25	200	11000
73	299399	251534	4,80	3,40	3,40	0	1,0	1,0	3,4	3,40	82	230
74	299399	251546	5,35	2,00	2,00	0,1	1,0	1,0	2,3	2,30	74	230
75	299397	251508	4,97	2,10	2,10	0	0,9	0,9	2,2	2,20	370	140
76	299314	251502	4,75	2,10	2,10	0	0,5	0,5	2,3	2,30	2500	660
77	299271	251539	3,80	1,80	1,80	0	1,0	1,0	1,8	1,80	160	170
79	299231	251577	4,15	1,10	1,10	0	1,0	1,0	1,3	1,30	35	1900
80	299217	251576	4,29	1,35	1,35	0	1,2	1,2	1,4	1,40	58	230
81	299198	251573	4,40	1,40	1,40	0	1,3	1,3	1,5	1,50	97	670
82	299196	251587	4,60	1,40	1,40	0	1,4	1,4	1,5	1,50	51	10
83	299095	251553	4,51	2,00	2,20	0	1,2	1,2	2,2	2,45	46	1800
84	299083	251531	4,40	2,15	2,15	0	1,3	1,3	2,2	2,20	110	100
85	299095	251532	4,32	1,55	1,55	0	1,2	1,2	1,5	1,50	120	580

Tableau 5 : Sondages complémentaires de juin 2006 : numéro de sondage, coordonnées X, Y en Lambert II, profondeurs de la base des remblais et du toit des argiles, niveaux échantillonnés (min et max) et teneurs en HCT correspondantes (en ppm).

N	X	Y	Base remblais	Toit argiles	Zmin A	Zmax A	Zmin B	Zmax B	ZmaxB avec argiles	HCT A	HCT B
86	299370	251535	2,2	2,2	1,5	2,2	2,2	2,5	2,5	5500	0
87	299329	251514	2,6	2,6	1,9	2,5	2,5	2,6	2,6	0	470
88	299331	251500	2	2	0	1,6	1,6	2	2,25	65	35000
89	299297	251510	2,5	2,5	1	2,3	2,3	2,5	2,5	20000	110
90	299299	251497	0,2	1,9	0,2	1,9	1,9	2,1	2,1	310	0
91	299310	251526	1,8	1,8	1,3	1,8	1,8	2	2	1800	350
92	299166	251597	1	1	0	1	1	1,6	1,6	85	0
93	299181	251597	1,1	1,1	1,1	1,7	1,7	2,1	2,1	60	490
94	299184	251583	1,5	1,5	1,2	1,5	1,5	1,8	1,8	14000	1600
95	299168	251568	1,8	1,8	1	1,8	1,8	2	2	330	0
96	299167	251583	1,2	1,2	0	1,2	1,2	2	2	190	0
97	299113	251548	1,7	1,7	0,9	1,7	1,7	2	2	5800	6000
98	299126	251549	1,8	1,8	1,3	1,8	1,8	2,1	2,1	11000	1500
99	299130	251533	1,1	1,9	1,1	1,9	1,9	2,4	2,4	13000	2100
100	299114	251531	1,9	1,9	0,4	1,9	1,9	2,1	2,1	11000	2000
101	299311	251512	1	2	1	2	2	2,3	2,3	370	0
102	299314	251497	1	2,1	1	2,1	2,1	2,2	2,35	46000	25000

8. ANNEXE B : DONNEES D'EXCAVATION

Tableau 6 : Données de dépollution : volumes de terres extraits et pollués.

TABLEAU DE DETAIL DES VOLUMES DE TERRES EXTRAITS ET POLLUES																									
SITA REMEDIATION																									
N° maille	Volume pollué								Terrassement global										Goudron		commentaire				
	Volume pollué théorique	Cumul volume pollué théorique	Volume pollué réel initial	Volume pollué reprise	Volume pollué reprise fin août	Total volume pollué réel	Cumul volume pollué réel	Volume restant à extraire (m3)	Volume potentiel (m3)	vol terrassement ent théo	cumul vol terrassement ent théo	Vol. ter en juillet	reprise ter. En juillet	total juillet	Vol. ter en août	reprise ter. En août	total aout	Vol terrassement ent réel initial	Volume terrassement ent reprise	Total maille		cumul vol terrassement réel	Vol goudron	cumul vol goudron	
SB10B												690		690				690		690	690	195	195		
SB11A	315	315	382			382	382			540	540	495		495				495		495	1185		195		
SB26 C	90	405	149,6			149,6	532			360	900	265		265				265		265	1450		195		
SB26 D	90	495	243,6			243,6	775			360	1260	360		360	383		383	743		743	2193		195		
SB28		495			280	280	1055				1260					500	500		500	500	2693	45	240	maille plus étendue	
SB29A	240	735	300	132	100	532	1587			540	1800	600	242	842		200	200	600	442	1042	3735		240	maille plus étendue	
SB29B	149	884	198	90	121,5	410	1997			446	2246	494		494		527	527	494	527	1021	4755		240	maille plus étendue	
SB29C	225	1109	240	195	26	461	2458			450	2696	495		495		448	448	495	448	943	5698	105	345	extension vers argile 20 cm	
SB29D	315	1424	285	83	60	428	2885			405	3101	405	158	563				405	158	563	6260	30	375	extension vers argile 20 cm	
F36 théo		1424					2885			450	3551	450		450				450		450	6710	45	420		
SB39	225	1649	215			215	3100			315	3866	300		300				300		300	7010		420		
SB41	112,5	1761	122			122	3222			293	4158	262		262				262		262	7272		420		
SB42	500	2261	375	126		501	3723			495	4653	495		495		156	156	495	156	651	7923	30	450		
SB43	270	2531	210			210	3933			360	5013	280		280				280		280	8203		450		
SB44	400	2931	292	232,5	210	734,5	4668			400	5413	383		383		405	405	383	405	788	8991		450		
SB45	337,5	3269	405	105	127,5	637,5	5305	165		338	5751	428		428		245	245	428	245	673	9663	20	470		
SB47 théo		3269					5305			270	6021	270		270				270		270	9933	45	515		
SB48	270	3539	276,25	120	80	476,25	5782			405	6426	416		416		150	150	416	150	566	10499	30	545		
SB51	315	3854	382,5		143,75	526,25	6308			315	6741	383		383		144	144	383	144	526	11026		545		
SB52	270	4124	187			187	6495			270	7011	187		187				187		187	11213		545		
SB53		4124	90			90	6585				7011	90		90				90		90	11303		545		
SB54	315	4439	270	80	40	390	6975			405	7416	450		450		100	100	450	100	550	11853	45	590		
SB56 théo		4439					6975			360	7776	315		315				315		315	12168		590		
SB58	675	5114	442			442	7417			675	8451	442		442				442		442	12610		590		
SB59	433	5547	315	109,5	150	574,5	7991	113		315	8766	450		450		150	150	450	150	600	13210	20	610		
SB5C	180	5727	251			251	8242			495	9261	499		499				499		499	13708		610		
SB60	315	6042	360			360	8602			450	9711	495		495				495		495	14203		610	extension vers argile 20 cm	
SB61	315	6357	382,5	85	97,5	565	9167			405	10116	473		473		203	203	473	203	675	14878		610		
SB62	225	6582	427,5	266	146,25	839,75	10007			495	10611	518		518		472	472	518	472	990	15868		610		
SB63	315	6897	337	135		472	10479			450	11061	495		495		180	180	495	180	675	16543		610	extension vers argile 20 cm	
SB64	315	7212	285		120	405	10884	105		495	11556	495		495		180	180	495	180	675	17218	30	640		
SB65	405	7617	318,75			318,75	11203			405	11961	405		405	484		484	889		889	18107	30	670		
SB66	270	7887	263		135	398	11601			495	12456	270		270	495	235	730	765	235	1000	19107	7	677		
SB70	270	8157	285	340	464,38	1089,375	12690	360		495	12951	529		529		1239	1239	529	1239	1768	20875	90	767		
SB72	135	8292	270			270	12960			450	13401	450		450				450		450	21325		767		
SB76A		8292					12960			495	13896	495		495				495		495	21820	225	992		
SB76B		8292					12960			495	14391	462		462				462		462	22282	120	1112		
TAS n°2		8292	225			225	13185				14391	225		225				225		225	22507		1112		
OK																									
TOTAUX			8785	2099	2302	13185		743																	

Note : les « volumes pollués théoriques » de la colonne 1 se basent exclusivement sur une comparaison des résultats analytiques par rapport au seuil et sur la hauteur prélevée correspondante. Ils ne se basent par conséquent pas du tout sur l'étude géostatistique de 2006.

Étude géostatistique d'un ancien site de stockage de produits chimiques : le site 1bis ATTENA.

Fournisseur du cas d'étude :	BURGEAP dans le cadre du projet R&D ADEME/ATTENA
Réalisation de l'étude :	KIDOVA
Contrôle qualité :	Mines ParisTech
Rédacteurs :	Jean-Baptiste Mathieu, Michel Garcia
Outil logiciel utilisé :	SoilRemediation® (KIDOVA)
Relecteurs :	Jean-Paul Chilès, Gaëlle Leloch

TABLE DES MATIERES

1. INTRODUCTION.....	53
2. DESCRIPTION DES TRAVAUX REALISES	55
3. PRESENTATION DU CAS D'ETUDE ATTENA	56
3.1 DESCRIPTION DU SITE.....	56
3.2 DONNEES DISPONIBLES	57
3.2.1 <i>Investigations réalisées</i>	57
3.3 DEFINITION DU CADRE DE L'ETUDE DE DEMONSTRATION.....	59
3.3.1 <i>Sélection des composants d'intérêt</i>	59
3.3.2 <i>Constitution du jeu de donnée restreint</i>	59
3.3.3 <i>Définition de la « zone source »</i>	60
4. ETUDE GEOSTATISTIQUE – MODELE GEOSTATISTIQUE DE BASE	60
4.1 ANALYSE EXPLORATOIRE ET STATISTIQUE DES DONNEES	60
4.1.1 <i>Jeu de donnée restreint</i>	60
4.1.2 <i>Etapes de l'analyse exploratoire et statistique des données</i>	61
4.1.3 <i>Analyse unidimensionnelle des teneurs et mesures PID</i>	61
4.1.4 <i>Corrélations entre composants</i>	63
4.1.5 <i>Corrélations entre teneurs et mesures PID</i>	64
4.1.6 <i>Extrapolation des teneurs en TCE et DCE au-dessous des limites de quantification</i>	65
4.1.7 <i>Transformation gaussienne des teneurs</i>	65
4.1.8 <i>Corrélation spatiale des teneurs</i>	66
4.2 MODELISATION GEOSTATISTIQUE A PARTIR DU JEU DE DONNEE RESTREINT	68
4.2.1 <i>Définition du domaine</i>	68
4.2.2 <i>Simulation conjointe des teneurs en PCE, TCE et DCE</i>	69
4.3 CONSIDERATIONS METHODOLOGIQUES	69
5. RESULTATS DU MODELE GEOSTATISTIQUE	71
5.1 SIMULATIONS DES TENEURS EN POLLUANTS	71
5.1.1 <i>Résultats</i>	71
5.2 DELIMITATION DE LA « ZONE SOURCE » DE LA POLLUTION	72
5.2.1 <i>Critères de définition de la zone source</i>	72
5.2.2 <i>Incertitude sur le volume de la zone source</i>	72
5.2.3 <i>Incertitude spatiale sur la position de la zone source</i>	73
6. CONFRONTATION DU MODELE GEOSTATISTIQUE AUX DONNEES DE VALIDATION.....	76
6.1 DONNEES DE VALIDATION	76
6.2 COMPARAISONS STATISTIQUES ENTRE DONNEES DU MODELE ET DONNEES DE VALIDATION.....	78
6.3 PREDICTIONS DU MODELE GEOSTATISTIQUE PAR RAPPORT AUX DONNEES DE VALIDATION.....	81
6.3.1 <i>Modes de comparaison</i>	81
6.3.2 <i>Comparaison entre teneurs simulées et teneurs mesurées</i>	81
6.3.3 <i>Evaluation de la précision et justesse du modèle</i>	83
6.3.4 <i>Prédictions du modèle par rapport à la zone source</i>	84
6.4 CONCLUSIONS	87
7. AMELIORATION DU MODELE GEOSTATISTIQUE AVEC DES STATISTIQUES COMPLETES.....	88
7.1 AMELIORATIONS APPORTEES AU MODELE	88
7.2 RESULTATS DU MODELE GEOSTATISTIQUE AMELIORE	94
7.2.1 <i>Simulation des teneurs</i>	94
7.2.2 <i>Incertitude sur le volume de la zone source</i>	94
7.2.3 <i>Incertitude spatiale sur la position de la zone source</i>	94
7.3 CONFRONTATION DU MODELE AVEC LES DONNEES DE VALIDATION.....	97
7.3.1 <i>Comparaison entre teneurs simulées et teneurs mesurées</i>	97
7.3.2 <i>Evaluation de la précision et justesse du modèle amélioré</i>	99

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

7.3.3	<i>Prédictions du modèle amélioré par rapport à la zone source</i>	99
7.4	CONCLUSIONS SUR LE MODELE AMELIORE	102
8.	MODELISATION GEOSTATISTIQUE COMPLETE AVEC L'ENSEMBLE DES DONNEES	103
8.1	CARACTERISTIQUES DU MODELE.....	103
8.2	RESULTATS DU MODELE GEOSTATISTIQUE COMPLET.....	103
8.2.1	<i>Simulation des teneurs</i>	103
8.2.2	<i>Incertitude sur le volume de la zone source</i>	104
8.2.3	<i>Incertitude spatiale sur la position de la zone source</i>	104
9.	SYNTHESE DES VOLUMES DE LA ZONE SOURCE ESTIMES ET LOCALISES A PARTIR DES TROIS MODELES GEOSTATISTIQUES	108
9.1	ESTIMATION DES VOLUMES DE LA ZONE SOURCE SUR L'ENSEMBLE DU DOMAINE	108
9.1.1	<i>Modèle amélioré comparé au modèle initial</i>	108
9.1.2	<i>Modèle complet comparé au modèle amélioré</i>	109
9.2	ESTIMATION DES VOLUMES DE LA ZONE SOURCE SUR UN DOMAINE RESTREINT	110
9.2.1	<i>Modèle amélioré comparé au modèle de référence</i>	113
9.2.2	<i>Modèle complet comparé au modèle amélioré</i>	113
10.	CONCLUSIONS	115
	ANNEXES SITE « ATTENA »	117
11.	ANNEXE A : RESUMES STATISTIQUE DE L'ENSEMBLE DES TENEURS MESUREES DANS LES SOLS	117
12.	ANNEXE B : LOGS DES SONDAGES OU SE SITUENT LES DONNEES SINGULIERES	120
13.	ANNEXE C : PRESENTATION DE METHODES ET MESURES STATISTIQUES UTILISEES DANS L'ETUDE ATTENA.....	123
13.1	STATISTIQUES BIDIMENSIONNELLES	123
13.2	DIAGRAMME QUANTILE-QUANTILE	123
14.	ANNEXE D : PRINCIPE DE L'EVALUATION DE LA JUSTESSE ET DE LA PRECISION D'UN MODELE DE SIMULATION GEOSTATISTIQUE.....	125

TABLE DES FIGURES

FIGURE 23 : CARTE ILLUSTRANT LE CONTEXTE DU SITE ET LA POSITION DES DIVERS OUVRAGES EN FONCTION DE LEUR CAMPAGNE : BSS ET HISTORIQUE (2004–08) – PRE-DIAGNOSTIQUE (2008) – DIAGNOSTIQUE (FIN 2008 ET 2009).....	56
FIGURE 24 : COUPE DE SONDAGE QUI MONTRE LA LITHOLOGIE DU SITE SUR LES PREMIERS METRES DE PROFONDEUR, LA ZONE DE BATTEMENT DE LA NAPPE (EN BLEU) ET LA POSITION D’ECHANTILLONS DE SOLS SUR LESQUELS DES TENEURS EN PCE ONT ETE MESUREES (EN ROUGE).....	57
FIGURE 25 : POSITION DES OUVRAGES RECENTS EN FONCTION DE LEUR CAMPAGNE : PRE-DIAGNOSTIQUE – DIAGNOSTIQUE.	58
FIGURE 26 : VUE EN PERSPECTIVE DE LA REPARTITION SPATIALE DES 319 ECHANTILLONS DE SOL (SPHERES ROUGES) ANALYSES LORS DE LA CAMPAGNE DE DIAGNOSTIC. LES SPHERES VIOLETTES ET JAUNES, QUI DELIMITENT LES LIGNES DE SONDAGE, RENSEIGNENT SUR LES CAMPAGNES DE RECONNAISSANCE (PRE-DIAGNOSTIQUE EN VIOLET, DIAGNOSTIQUE EN JAUNE). LA SURFACE EN VERT EST UNE ESTIMATION DE LA TOPOGRAPHIE DU SITE ETABLI A PARTIR DES DONNEES TOPOGRAPHIQUES DISPONIBLES AU DROIT DES SONDAGES.	58
FIGURE 27 : DELIMITATION DU DOMAINE D’ETUDE (LIGNE ROUGE) ET POSITIONS DES OUVRAGES SELECTIONNES (SPHERES ROUGES). 60	
FIGURE 28 : DIAGRAMMES DE DISPERSION ENTRE POIDS DE DEGROUPEMENT (VERTICAL) ET DONNEES DE TENEUR (HORIZONTAL EN ECHELLE LOG) POUR LES TROIS COMPOSANTS. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).....	62
FIGURE 29 : HISTOGRAMMES DES TENEURS EN PCE (EN HAUT A GAUCHE), TCE (EN HAUT A DROITE) ET DCE (EN BAS A GAUCHE) ET DU PID (EN BAS A DROITE) DES DONNEES DU MODELE GEOSTATISTIQUE (UNITE : PPM).	62
FIGURE 30 : DIAGRAMMES DE DISPERSION EN ECHELLE LOG ENTRE A) TCE ET PCE, B) DCE ET PCE ET C) DCE ET TCE POUR LES DONNEES DE VALIDATION. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES).	63
FIGURE 31 : DIAGRAMMES DE DISPERSION EN ECHELLE LOG ENTRE TENEURS (AXE VERTICAL) ET MESURES PID (AXE HORIZONTAL) POUR LES TROIS COMPOSANTS D’INTERET (UNITE : PPM) LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES).	64
FIGURE 32 : DIAGRAMMES DE DISPERSION ENTRE A) TCE ET PCE, B) DCE ET PCE ET C) DCE ET TCE APRES TRANSFORMATION GAUSSIENNE DES DONNEES. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES).	66
FIGURE 33 : VARIOGRAMMES EXPERIMENTAUX (POINTS) ET MODELES CALES (LIGNES) POUR LE PCE, TCE ET DCE (DONNEES TRANSFORMEES, PROFONDEUR EN COORDONNEE VERTICALE) ETABLI A PARTIR DU JEU DE DONNEES RESTREINT.	67
FIGURE 34 : REPRESENTATION 3D DU DOMAINE DE SIMULATION AVEC VISUALISATION DU JEU DE DONNEES RESTREINT DU MODELE GEOSTATISTIQUE (SPHERES JAUNES) ET LE JEU DE DONNEES DE VALIDATION DU MODELE (SPHERES ROUGES). LA DISTORSION VERTICALE DE L’IMAGE EST DE 7 POUR 1.	68
FIGURE 35 : COUPES HORIZONTALES DE DEUX SIMULATIONS CONJOINTES DES TENEURS POUR LES TROIS COMPOSANTS (UNITE : PPM). LES SONDAGES AFFICHES SONT CEUX D’OU PROVIENNENT LES DONNEES DE TENEUR UTILISEES POUR CONDITIONNER LES SIMULATIONS.....	71
FIGURE 36 : DIAGRAMMES QUANTILE-QUANTILE ENTRE LES TENEURS SIMULEES DE CHAQUE REALISATION ET LES DONNEES DE TENEUR UTILISEES POUR CONDITIONNER LES SIMULATIONS. LES DISTRIBUTIONS DE REFERENCE TIREES DES DONNEES SONT CORRIGEEES DES BIAIS D’ECHANTILLONNAGE (DONNEES PONDEREES).....	72
FIGURE 37 : HISTOGRAMMES DES VOLUMES DE LA ZONE SOURCE EN M ³ CALCULES A PARTIR DES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS).	73
FIGURE 38 : COUPES HORIZONTALES DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE.	74
FIGURE 39 : COUPES VERTICALES VUES EN PERSPECTIVE DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE.	74
FIGURE 40 : HISTOGRAMMES ET RESUMES STATISTIQUES DES PROBABILITES DE PRESENCE DE LA ZONE SOURCE CALCULEES A PARTIR DES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS).	75
FIGURE 41 : VISUALISATIONS 3D DE LA POSITION LA PLUS PROBABLE DE LA ZONE SOURCE, CORRESPONDANT A UNE PROBABILITE DE PRESENCE DE LA SOURCE > 0.8 (PLUS DE 80% DE CHANCES), POUR DES SEUILS DE 20 ET 40 PPM. LES COUPES VERTICALES ET HORIZONTALES SUR LA GRILLE MONTRENT LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE POUR LE SEUIL CONSIDERE. LA DISTORSION VERTICALE DE L’IMAGE EST DE 7 POUR 1.....	76

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

FIGURE 42 : VISUALISATIONS 3D DES DONNEES DE VALIDATION AINSI QUE DE LA POSITION LA PLUS PROBABLE DE LA ZONE SOURCE CALCULEE POUR UNE PROBABILITE DE PRESENCE SUPERIEURE A 0.5 (PLUS DE 50% DE CHANCES) ET DES SEUILS DE 20 (EN HAUT) ET 40 PPM (EN BAS). LES SPHERES CORRESPONDENT AUX DONNEES DE VALIDATION QUI SONT EFFECTIVEMENT DANS LA ZONE SOURCE ET LES CROIX A CELLES QUI SONT EN DEHORS. LA COULEUR DES CROIX ET DES SPHERES CORRESPOND A LA SOMME DES TENEURS EN DCE, PCE ET TCE AU POINT CONSIDERE (ECHELLE DE COULEUR LOGARITHMIQUE). LA DISTORSION VERTICALE DE L'IMAGE EST DE 7 POUR 1.	77
FIGURE 43 : DIAGRAMMES DE DISPERSION ENTRE POIDS DE DEGROUPEAGE (VERTICAL) ET DONNEES DE TENEUR (HORIZONTAL EN ECHELLE LOG) POUR LES DONNEES DE VALIDATION DES TROIS COMPOSANTS. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).	78
FIGURE 44 : HISTOGRAMMES DES TENEURS EN PCE (EN HAUT A GAUCHE), TCE (EN HAUT A DROITE) ET DCE (EN BAS A GAUCHE) DU JEU DE DONNEES DE VALIDATION (UNITE : PPM).	78
FIGURE 45 : DIAGRAMMES QUANTILE-QUANTILE ENTRE LES DISTRIBUTIONS DES JEUX DE DONNEES DU MODELE ET DE VALIDATION POUR LES TROIS COMPOSANTS PCE, TCE ET DCE. POUR CHAQUE COMPOSANT, LES DISTRIBUTIONS DEDUITES DE DONNEES DEGROUPEES OU NON SONT PRESENTEES. LISTE DES QUANTILES VISUALISES : Q0.1, Q0.2, ..., Q0.6, Q0.65, Q0.7, Q0.75, Q0.8, Q0.85, Q0.9, Q0.95, Q0.975.	80
FIGURE 46 : DIAGRAMMES DE DISPERSION EN ECHELLE LOG ENTRE A) TCE ET PCE, b) DCE ET PCE ET c) DCE ET TCE POUR L'ENSEMBLE DES DONNEES DE VALIDATION. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES).	80
FIGURE 47 : DIAGRAMMES DE DISPERSION ENTRE LES VALEURS SIMULEES ET LES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).	82
FIGURE 48 : DIAGRAMMES DE DISPERSION ENTRE LES MOYENNES DES VALEURS SIMULEES ET LES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).	82
FIGURE 49 : DIAGRAMMES DE DISPERSION ENTRE LES MEDIANES DES VALEURS SIMULEES ET LES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).	82
FIGURE 50 : DIAGRAMMES DE PRECISION ET JUSTESSE CALCULES SUR LES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS EN REPOSITIONNANT A LA LIMITE DE QUANTIFICATION LES VALEURS SIMULEES AU-DESSOUS.	83
FIGURE 51 : HISTOGRAMMES DES PROBABILITES DE DEPASSEMENT DES SEUILS, A L'ENDROIT DES DONNEES DE VALIDATION POUR LES SEUILS DE 400, 250, 1200, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). DIFFERENTIATION EN ROUGE ET EN BLEU DES DONNEES SITUEES RESPECTIVEMENT DANS ET HORS DE LA ZONE SOURCE POUR LE SEUIL DE TENEUR CONSIDERE.	84
FIGURE 52 : DIAGRAMMES DES ERREURS RELATIVES ENTRE LA PROPORTION DE DONNEES DE VALIDATION MAL CLASSEES ET LA PROBABILITE D'ERREUR DE CLASSIFICATION MOYENNE DES DONNEES CLASSEES SELON LA VALEUR DU SEUIL CRITIQUE (EN PPM).	86
FIGURE 53 : HISTOGRAMMES DES TENEURS EN PCE (EN HAUT A GAUCHE), TCE (EN HAUT A DROITE) ET DCE (EN BAS A GAUCHE) POUR L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION (UNITE : PPM).	90
FIGURE 54 : DIAGRAMMES DE DISPERSION ENTRE POIDS DE DEGROUPEAGE (VERTICAL) ET DONNEES DE TENEUR (HORIZONTAL EN ECHELLE LOG) POUR L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES).	90
FIGURE 55 : HISTOGRAMMES EN ECHELLE LOG ET RESUMES STATISTIQUES DES TENEURS EN DCE CORRESPONDANT A DES ECHANTILLONS OU LA TENEUR EN TCE EST SOUS LA LIMITE DE QUANTIFICATION.	91
FIGURE 56 : HISTOGRAMMES ET STATISTIQUES DES TENEURS EN PCE (EN HAUT), TCE (AU CENTRE) ET DCE (EN BAS) APRES TRANSFORMATION GAUSSIENNE DE L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION (COLONNE DE GAUCHE) PUIS APRES CORRECTION DE LA VALEUR TRANSFORMEE CORRESPONDANT A LA LIMITE DE QUANTIFICATION POUR CENTRER LES DISTRIBUTIONS SUR 0 (COLONNE DE DROITE). LA COURBE BLEUE INDIQUE LE MODELE GAUSSIEN CALE SUR CHAQUE DISTRIBUTION.	91
FIGURE 57 : DIAGRAMMES DE DISPERSION EN ECHELLE LOG ENTRE A) TCE ET PCE, b) DCE ET PCE ET c) DCE ET TCE POUR L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION. LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES AVEC ECHELLE DE COULEUR).	92
FIGURE 58 : VARIOGRAMMES EXPERIMENTAUX (POINTS) ET MODELES CALES (LIGNES) POUR LE PCE, TCE ET DCE (DONNEES TRANSFORMEES, PROFONDEUR EN COORDONNEE VERTICALE) ETABLIS A PARTIR DE L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION.	92

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

FIGURE 59 : COUPES HORIZONTALES DE DEUX SIMULATIONS CONJOINTES DES TENEURS POUR LES TROIS POLLUANTS (UNITE : PPM). LES SONDAGES AFFICHES SONT CEUX D'OU PROVIENNENT LES DONNEES DE TENEUR UTILISEES POUR CONDITIONNER LES SIMULATIONS. MODELE GEOSTATISTIQUE AMELIORE.	93
FIGURE 60 : HISTOGRAMME DES VOLUMES EN M ³ DE LA ZONE SOURCE CALCULES SUR LES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE AMELIORE.....	94
FIGURE 61 : COUPES HORIZONTALES DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE. MODELE GEOSTATISTIQUE AMELIORE.	95
FIGURE 62 : COUPES VERTICALES VUES EN PERSPECTIVE DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE. MODELE GEOSTATISTIQUE AMELIORE.	95
FIGURE 63 : HISTOGRAMMES ET RESUMES STATISTIQUES DES PROBABILITES DE PRESENCE DE LA ZONE SOURCE CALCULEES A PARTIR DES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE AMELIORE.....	96
FIGURE 64 : VISUALISATIONS 3D DE LA POSITION LA PLUS PROBABLE DE LA ZONE SOURCE A 20 PPM SI ON NE CHOISIT QUE LES MAILLES POUR LESQUELLES LA PROBABILITE DE DEPASSEMENT DU SEUIL CRITIQUE EST SUPERIEURE OU EGALE A 0.8 (VOLUME : 993 M ³). LA DISTORSION VERTICALE DE L'IMAGE EST DE 7 POUR 1. MODELE GEOSTATISTIQUE AMELIORE.....	97
FIGURE 65 : DIAGRAMMES DE DISPERSION ENTRE LES VALEURS SIMULEES ET L'ENSEMBLE DES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES). MODELE GEOSTATISTIQUE AMELIORE.....	97
FIGURE 66 : DIAGRAMMES DE DISPERSION ENTRE LES MOYENNES DES VALEURS SIMULEES ET L'ENSEMBLE DES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES). MODELE GEOSTATISTIQUE AMELIORE.....	98
FIGURE 67 : DIAGRAMMES DE DISPERSION ENTRE LES MEDIANES DES VALEURS SIMULEES ET L'ENSEMBLE DES DONNEES DE VALIDATION POUR LES TROIS COMPOSANTS (ECHELLES = VARIABLES TRANSFORMEES). LES CARTES DE FOND RENSEIGNENT SUR LA DENSITE DE PROBABILITE (HISTOGRAMMES BIDIMENSIONNELS LISSES ASSOCIES A UNE ECHELLE DE COULEUR ARC-EN-CIEL POUR FAIRE RESSORTIR LES STRUCTURES). MODELE GEOSTATISTIQUE AMELIORE.....	98
FIGURE 68 : DIAGRAMMES DE PRECISION ET JUSTESSE CALCULE SUR LES DONNEES DE VALIDATION NON EXTRAPOLEES POUR LES TROIS COMPOSANTS (LES VALEURS SIMULEES SOUS LES SEUILS DE QUANTIFICATION ONT ETE PLACEES AU SEUIL). MODELE GEOSTATISTIQUE AMELIORE.	99
FIGURE 69 : HISTOGRAMMES DES PROBABILITES DE DEPASSEMENT DES SEUILS, A L'ENDROIT DES DONNEES DE VALIDATION POUR LES SEUILS DE 400, 250, 1200, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). DIFFERENTIATION EN ROUGE ET EN BLEU DES DONNEES SITUEES RESPECTIVEMENT DANS ET HORS DE LA ZONE SOURCE POUR LE SEUIL DE TENEUR CONSIDERE. MODELE GEOSTATISTIQUE AMELIORE.	100
FIGURE 70 : DIAGRAMMES DES ERREURS RELATIVES ENTRE LA PROPORTION DE DONNEES DE VALIDATION MAL CLASSEES ET LA PROBABILITE D'ERREUR DE CLASSIFICATION MOYENNE DES DONNEES CLASSEES SELON LA VALEUR DU SEUIL CRITIQUE (EN PPM). MODELE GEOSTATISTIQUE AMELIORE.	101
FIGURE 71 : COUPES HORIZONTALES DE DEUX SIMULATIONS CONJOINTES DES TENEURS POUR LES TROIS POLLUANTS (UNITE : PPM). LES SONDAGES AFFICHES SONT CEUX D'OU PROVIENNENT LES DONNEES DE TENEUR UTILISEES POUR CONDITIONNER LES SIMULATIONS. MODELE GEOSTATISTIQUE COMPLET.....	103
FIGURE 72 : HISTOGRAMME DES VOLUMES EN M ³ DE LA ZONE SOURCE CALCULES SUR LES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE COMPLET.....	104
FIGURE 73 : COUPES HORIZONTALES DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE. MODELE GEOSTATISTIQUE COMPLET.	105
FIGURE 74 : COUPES VERTICALES VUES EN PERSPECTIVE DE LA PROBABILITE DE PRESENCE DE LA ZONE SOURCE CALCULEE A PARTIR DES 100 SIMULATIONS CONJOINTES POUR UN SEUIL CRITIQUE DE 20 PPM APPLIQUE A LA SOMME DES TENEURS EN PCE, TCE ET DCE. MODELE GEOSTATISTIQUE COMPLET.	105
FIGURE 75 : HISTOGRAMMES ET RESUMES STATISTIQUES DES PROBABILITES DE PRESENCE DE LA ZONE SOURCE CALCULEES A PARTIR DES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE COMPLET....	106

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

FIGURE 76 : VISUALISATIONS 3D DE LA POSITION LA PLUS PROBABLE DE LA ZONE SOURCE A 20 PPM SI ON NE CHOISIT QUE LES MAILLES POUR LESQUELLES LA PROBABILITE DE DEPASSEMENT DU SEUIL CRITIQUE EST SUPERIEURE OU EGALE A 0.8 (VOLUME : 1 346 M3) (DISTORSION VERTICALE = 7 POUR 1). MODELE GEOSTATISTIQUE COMPLET.	107
FIGURE 77 : DELIMITATION DU DOMAINE RESTREINT ET DE LA GRILLE REDUITE ASSOCIEE UTILISEE POUR ESTIMER LES VOLUMES DE LA ZONE SOURCE.	111
FIGURE 78 : HISTOGRAMME DES VOLUMES EN M3 DE LA ZONE SOURCE CALCULES SUR LES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE INITIAL LIMITE AU DOMAINE RESTREINT.	111
FIGURE 79 : HISTOGRAMME DES VOLUMES EN M3 DE LA ZONE SOURCE CALCULES SUR LES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE AMELIORE LIMITE AU DOMAINE RESTREINT.	112
FIGURE 80 : HISTOGRAMME DES VOLUMES EN M3 DE LA ZONE SOURCE CALCULES SUR LES 100 SIMULATIONS CONJOINTES DES TROIS COMPOSANTS EN APPLIQUANT A LA SOMME DES TENEURS LES SEUILS DE 400, 250, 120, 60, 40 ET 20 PPM (DE GAUCHE A DROITE ET DE HAUT EN BAS). MODELE GEOSTATISTIQUE COMPLET LIMITE AU DOMAINE RESTREINT.	112

TABLE DES TABLEAUX

TABEAU 7 : RESUME STATISTIQUE DES TENEURS CORRIGES DU BIAS D'ECHANTILLONNAGE POUR LES COMPOSANTS SELECTIONNES ET DE LA MESURE PID (UNITE : PPM).	62
TABEAU 8 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE LA TRANSFORMEE LOG10 DES TENEURS DE CHACUN DES COMPOSANTS POUR TOUTES LES DONNEES, SANS LES DONNEES SOUS LES LIMITES DE QUANTIFICATION (LQ), SANS LES DONNEES SINGULIERES ET SANS LES DONNEES SOUS LQ ET SINGULIERES.	64
TABEAU 9 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE LA TRANSFORMEE LOG10 DES TENEURS ET DES MESURES PID POUR TOUTES LES DONNEES ET SANS LES DONNEES SOUS LES LIMITES DE QUANTIFICATION (LQ).	65
TABEAU 10 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG DES TENEURS EN PCE, TCE ET DCE APRES TRANSFORMATION GAUSSIENNE DES TENEURS. LES CORRELATIONS AVEC DCE SONT CALCULEES SANS LES TROIS DONNEES SINGULIERES.	66
TABEAU 11 : MODELES DE VARIOGRAMMES DIRECTES ET CROISES CALES SUR LES VARIOGRAMMES EXPERIMENTAUX. LES VARIOGRAMMES CROISES ONT ETE DEDUITS DES VARIOGRAMMES DIRECTS EN PRENANT EN COMPTE UNE CORRELATION LINEAIRE DE 0.9 ENTRE TCE ET PCE ET DE 0.75 ENTRE DCE ET TCE.	68
TABEAU 12 : NOMBRE DE MODELES DE VARIOGRAMME A ETABLIR ET NOMBRE MAXIMAL DE DONNEES DE CONDITIONNEMENT A PRENDRE EN COMPTE DANS LE SYSTEME DE COKRIGEAGE EN FONCTION DU NOMBRE DE POLLUANTS A SIMULER CONJOINTEMENT PAR UNE METHODE DE COKRIGEAGE MULTICOLOCALISE.	70
TABEAU 13 : RESUMES STATISTIQUES DES TENEURS EN PCE, TCE ET DCE DU JEU DE DONNEES DE VALIDATION (UNITE : PPM).	79
TABEAU 14 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE LA TRANSFORMEE LOG10 DES TENEURS DE CHACUN DES COMPOSANTS POUR TOUTES LES DONNEES DE VALIDATION ET SANS LES DONNEES SOUS LES LIMITES DE QUANTIFICATION (LQ).	81
TABEAU 15 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE L'ENSEMBLE DES DONNEES DE VALIDATION (VARIABLES TRANSFORMEES) ET LES VALEURS SIMULEES OU DES STATISTIQUES CENTRALES TIREES DE CES VALEURS.	83
TABEAU 16 : CORRESPONDANCE ENTRE LA PROPORTION DE DONNEES DE VALIDATION MAL CLASSEES ET LA PROBABILITE D'ERREUR DE CLASSIFICATION MOYENNE DES DONNEES CLASSEES. CALCULS REPETES POUR DIFFERENTS INTERVALLES DE CONFIANCE ET LES SIX SEUILS DE DEFINITION DE LA ZONE SOURCE.	86
TABEAU 17 : RESUMES STATISTIQUES DES TENEURS EN PCE, TCE ET DCE POUR L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION (UNITE : PPM).	90
TABEAU 18 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE LA TRANSFORMEE LOG10 DES TENEURS DE CHACUN DES POLLUANTS POUR L'ENSEMBLE DES DONNEES (JEUX DE DONNEES COMPLET) ET SANS LES DONNEES SOUS LES LIMITES DE QUANTIFICATION (LQ).	92
TABEAU 19 : MODELES DE VARIOGRAMMES DIRECTS ET CROISES CALES SUR LES VARIOGRAMMES EXPERIMENTAUX, A PARTIR DE L'ENSEMBLE DES DONNEES DU MODELE GEOSTATISTIQUE ET DE VALIDATION, ET CORRIGES POUR QUE LES SIMULATIONS REPRODUISSENT CONVENABLEMENT LES DISTRIBUTIONS DE VALEURS ET LES CORRELATIONS ENTRE COMPOSANTS SOUHAITEES (VOIR ITEM 6 P 101).	93
TABEAU 20 : COEFFICIENTS DE CORRELATION LINEAIRE ET DE RANG ENTRE L'ENSEMBLE DES DONNEES DE VALIDATION (VARIABLES TRANSFORMEES) ET LES VALEURS SIMULEES OU DES STATISTIQUES CENTRALES TIREES DE CES VALEURS. MODELE GEOSTATISTIQUE AMELIORE.	98
TABEAU 21 : CORRESPONDANCE ENTRE LA PROPORTION DE DONNEES DE VALIDATION MAL CLASSEES ET LA PROBABILITE D'ERREUR DE CLASSIFICATION MOYENNE DES DONNEES CLASSEES. CALCULS REPETES POUR DES DIFFERENTS INTERVALLES DE CONFIANCE ET DIFFERENTS SEUILS DE DEFINITION DE LA ZONE SOURCE. MODELE GEOSTATISTIQUE AMELIORE.	101

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

TABLEAU 22 : RECAPITULATIF DE L'INCERTITUDE SUR LES VOLUMES DE LA ZONE SOURCE ET SUR LEUR LOCALISATION POUR LES TROIS MODELES GEOSTATISTIQUES ET POUR UN SEUIL DE 20 PPM EN CONSIDERANT L'ENSEMBLE DU DOMAINE.....	110
TABLEAU 23 : RECAPITULATIF DE L'INCERTITUDE SUR LES VOLUMES DE LA ZONE SOURCE ET SUR LEUR LOCALISATION POUR LES TROIS MODELES GEOSTATISTIQUES ET POUR UN SEUIL DE 400 PPM EN CONSIDERANT L'ENSEMBLE DU DOMAINE.....	110
TABLEAU 24 : RECAPITULATIF DE L'INCERTITUDE SUR LES VOLUMES DE LA ZONE SOURCE ET SUR LEUR LOCALISATION POUR LES TROIS MODELES GEOSTATISTIQUES ET POUR UN SEUIL DE 20 PPM EN CONSIDERANT LE DOMAINE RESTREINT DE 36 673 m ³	114
TABLEAU 25 : RECAPITULATIF DE L'INCERTITUDE SUR LES VOLUMES DE LA ZONE SOURCE ET SUR LEUR LOCALISATION POUR LES TROIS MODELES GEOSTATISTIQUES ET POUR UN SEUIL DE 400 PPM EN CONSIDERANT LE DOMAINE RESTREINT DE 36 673 m ³	114
TABLEAU 26 : RESUMES STATISTIQUES DES COMPOSANTS MESURES MOINS DE 10 FOIS.....	117
TABLEAU 27 : RESUMES STATISTIQUES DES COMPOSANTS MESURES ENTRE 10 ET 100 FOIS.	118
TABLEAU 28 : RESUMES STATISTIQUES DES COMPOSANTS MESURES PLUS DE 100 FOIS, CLASSES SELON LE NOMBRE DE LEURS MESURES AU-DESSUS DE LA LIMITE DE QUANTIFICATION DES APPAREILS DE MESURE.	119

1. INTRODUCTION

Le site Ibis ATTENA est un ancien site de stockage, conditionnement et expédition de produits chimiques. Les sols y sont pollués par des COHV, des BTEX, des HCT et des composés organo-solubles (alcools, cétones). Une ESR a déjà été réalisée et un suivi annuel de la qualité de la nappe existe depuis 2001. Les données de pollution disponibles (dites historiques et pré-diagnostic) proviennent de sondages, piezairs, piézomètres et mesures géophysiques réalisés dans le cadre du projet R&D ATTENA, cofinancé par l'ADEME, coordonné par le BRGM et dont les participants sont INERIS, BURGEAP, TOTAL PETROCHEMICALS, ARCELOR MITTAL et RHODIA. De nombreuses investigations complémentaires y ont été réalisées jusqu'à mi-2009. Elles incluent des sondages à 5 et 30 m avec des prélèvements de sols entre 4 et 30m de profondeur, des piézomètres, des piezair et des analyses en laboratoire sur les matrices sol, air et eau. L'accord a été donné à GeoSiPol par les participants à ATTENA pour utiliser ces données dans les études de démonstration.

Ce site de recherche n'a pas fait l'objet d'une dépollution avec excavation des sols. Par conséquent, les résultats de l'étude géostatistique n'ont pas pu être validés de façon classique en les confrontant à des volumes de sols pollués excavés. En rapport avec les travaux de recherche menés dans le cadre du projet ATTENA, d'autres objectifs d'étude tout aussi intéressants ont cependant pu être définis en s'intéressant à la délimitation de la « zone source » définie comme étant constituée des sols dont la somme des teneurs en PCE, TCE et DCE est suffisamment grande pour qu'une phase organique soit présente. Les trois modèles géostatistiques suivants ont ainsi été produits en exploitant progressivement les nombreuses données disponibles (287 échantillons de sols analysés).

1. **Modèle géostatistique de base.** Le jeu de données complet a été décomposé en deux sous-ensembles de 49 et 238 échantillons, le premier devant servir à construire le modèle géostatistique de la pollution des sols, le second à vérifier les bonnes prédictions du modèle. Le choix des données du modèle a été effectué par un ingénieur BURGEAP n'ayant pas participé à l'étude ATTENA en se déterminant par rapport à des conditions d'étude habituelles sur le nombre de données et leur répartition (en particulier verticale).
2. **Modèle géostatistique amélioré.** Une première amélioration a porté sur les modèles de distribution et variographiques (distributions, corrélations entre composants et variogrammes). Le jeu de données complet (287 échantillons) a ainsi été utilisé pour estimer plus justement ces modèles et étudier l'impact de cette amélioration sur la simulation géostatistique des teneurs en utilisant toujours le jeu de données restreint (49 échantillons) pour conditionner les simulations. Comme précédemment, les 238 données de validation, non prises en compte dans les simulations, ont servi à analyser les prédictions du modèle géostatistique amélioré.
3. **Modèle géostatistique complet.** En s'appuyant sur les modèles statistiques améliorés établis précédemment, les teneurs ont été simulées à nouveau mais en les conditionnant sur le jeu de données complet (287 échantillons). Le modèle géostatistique ainsi obtenu montre comment les 238 données nouvelles contribuent à réduire considérablement l'incertitude spatiale. Ce modèle peut être considéré comme un modèle de référence de la zone source par rapport aux modèles précédents.

Ce cas d'étude présente deux particularités intéressantes. D'une part, il nécessite de simuler conjointement les trois composants, qui sont corrélés entre eux, pour modéliser la zone source. D'autre part, il permet d'aborder le problème des données sous les limites de détection analytiques, qui représentent de 15 à 50% des données selon les composants et dont la prise en compte dans les modèles géostatistiques nécessite des précautions particulières.

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

Toujours en liaison avec le projet ATTENA, deux autres pistes avaient été initialement évoquées pour valider le modèle géostatistique en s'intéressant à la zone source :

- comparaison avec les modèles physiques réalisés par les partenaires du projet ATTENA (simulations numériques),
- comparaison avec les concentrations en polluant mesurées dans les autres milieux eau et air présents dans les sols.

Sans pouvoir disposer des données nécessaires dans les délais prévus pour les études de démonstration, ces pistes ont été abandonnées pour se concentrer plus largement sur les autres aspects de la démonstration décrits précédemment. Si l'intérêt devait être confirmé, ces autres modes de validation pourraient cependant faire l'objet de travaux complémentaires dans le cadre des études de démonstration ou du projet ATTENA.

2. DESCRIPTION DES TRAVAUX REALISES

Pour cette étude de démonstration, les travaux suivants ont été réalisés.

1. Etapes préliminaires
 - 1.1. Compilation des données disponibles
 - Synthèse des données sous un format unique et facilement exploitable
 - Analyse exploratoire préliminaire
 - 1.2. Sélection des polluants « d'intérêt » pour l'étude de démonstration
 - 1.3. Constitution du jeu de données à utiliser pour construire le modèle géostatistique
2. Réalisation de l'étude géostatistique à partir du jeu de données restreint
 - 2.1. Analyse exploratoire des données
 - Analyse statistique élémentaire (unidimensionnelle) des composants sélectionnés
 - Analyse des corrélations entre composants
 - Calcul des corrélations linéaires et d'ordre entre les variables d'intérêt
 - Dégroupement des données (correction du biais d'échantillonnage, les données n'étant pas régulièrement réparties latéralement et en profondeur), correction des mesures au-dessous des seuils de détection et transformation gaussienne des variables d'intérêt
 - Etude des diagrammes de dispersion entre les différentes variables transformées
 - Analyse critique des résultats : identification de données « singulières »
 - Validation avec Burgeap
 - Analyse variographique des données de teneur transformées avec et sans les données jugées « singulières »
 - 2.2. Modélisation géostatistique des teneurs en polluants dans les sols
 - Définition du domaine et génération de la grille de simulation
 - Simulation conjointe des trois polluants sélectionnés (simulation séquentielle gaussienne)
 - Application de différents seuils de teneur de présence de la « zone source » définis pour l'ensemble du site (non prise en compte des faciès)
 - 2.3. Analyse des résultats
 - Volume et localisation de la « zone source »
 - Analyse des incertitudes
3. Validation de l'étude géostatistique
 - 3.1. Confrontation des résultats du modèle aux données de validation non prises en compte dans le modèle géostatistique
 - Comparaison des valeurs simulées et mesurées
 - Diagramme de justesse et de précision du modèle par rapport aux données de validation
 - Prédications du modèle par rapport à la zone source sur les données de validation
4. Amélioration du modèle géostatistique
 - 4.1. Etablissement de modèles de distribution et de variogramme plus justes à partir de l'ensemble des données (répétition des travaux 2.1)
 - 4.2. Simulation géostatistique des teneurs en polluants dans les sols à partir des modèles statistiques améliorés en n'utilisant que le jeu de données restreint pour conditionner les simulations (répétition des travaux 2.2 à l'exception de la génération de la grille)
 - 4.3. Analyse des résultats (répétition des travaux 2.3)
 - 4.4. Confrontation du modèle géostatistique amélioré aux données de validation (répétition des travaux 3.1)
 - 4.5. Comparaison avec le modèle géostatistique de base
5. Modélisation géostatistique complète
 - 5.1. Simulation géostatistique des teneurs en polluants dans les sols à partir des modèles statistiques améliorés établis précédemment (étape 4) et en utilisant l'ensemble des données disponibles pour conditionner les simulations
 - 5.2. Comparaison avec les modèles géostatistiques de base et amélioré simulés avec le jeu de données restreint

3. PRESENTATION DU CAS D'ETUDE ATTENA

3.1 Description du site

Le site 1bis ATTENA est un ancien site de stockage, conditionnement et expédition de produits chimiques. Les sols y sont principalement pollués par des COHV, des BTEX, des HCT et dans une moindre mesure des composés organo-solubles (alcools, cétones). Ce site est étudié dans le cadre du projet de recherche ATTENA co-financé par l'ADEME, coordonné par le BRGM et avec pour partenaires l'INERIS, BURGEAP, TOTAL PETROCHEMICALS, ARCELOR MITTAL et RHODIA. Ce projet porte sur l'atténuation naturelle des composés organochlorés, des hydrocarbures et des HAP. En accord avec les autres partenaires du projet ATTENA, Burgeap a mis à la disposition de GeoSiPol les nombreuses données disponibles sur ce site pour que puisse être menée une étude de démonstration.



Figure 23 : Carte illustrant le contexte du site et la position des divers ouvrages en fonction de leur campagne : BSS et Historique (2004-08) – Pré-diagnostique (2008) – Diagnostique (fin 2008 et 2009).

L'étude de démonstration ne s'intéresse qu'aux cinq premiers mètres de sols qui sont les mieux renseignés par les données disponibles. A ces profondeurs, les sondages révèlent la présence de trois lithologies auxquelles se rajoute le revêtement du site comme le montre la coupe de la Figure 24 :

- « Remblais » : couche de remblais hétérogènes d'environ 70 cm.
- « Limons » : couche de 2 à 3m composée de limons tourbeux et de limons argileux.
- « Alluvions » : couche d'alluvions récentes de cours d'eau (sables fins et grossiers).

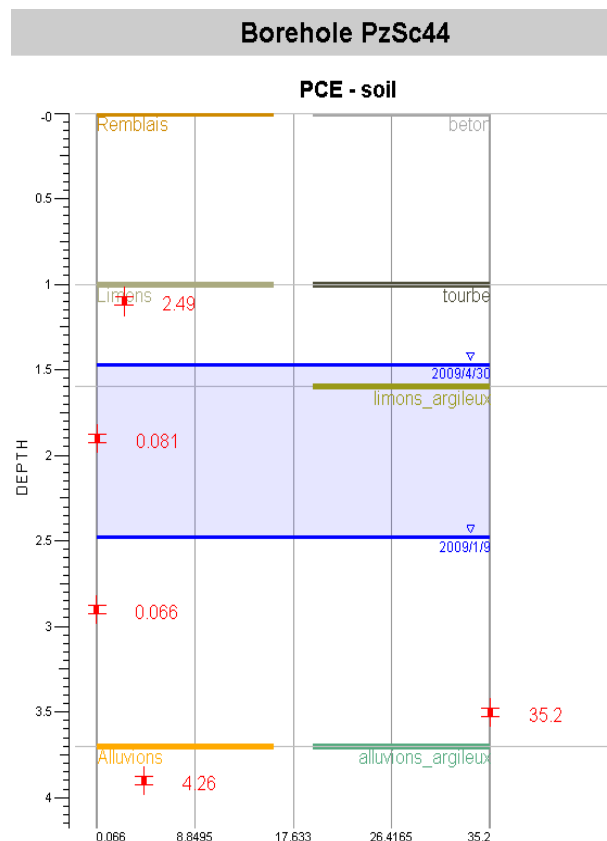


Figure 24 : Coupe de sondage qui montre la lithologie du site sur les premiers mètres de profondeur, la zone de battement de la nappe (en bleu) et la position d'échantillons de sols sur lesquels des teneurs en PCE ont été mesurées (en rouge).

3.2 Données disponibles

3.2.1 Investigations réalisées

Une étude simplifiée des risques (ESR) a déjà été réalisée et un suivi annuel de la qualité de la nappe existe depuis 2001. Les données de pollution disponibles proviennent de sondages, piézais, piézomètres et mesures géophysiques. Leurs positions géographiques sont illustrées sur les Figure 23, Figure 25 et Figure 26.

Les données d'investigation se répartissent comme suit :

- 19 coupes lithologiques issues de la base de données sols (BSS).
- 6 piézomètres historiques :
 - 22 échantillons d'eaux souterraines analysés entre octobre 2004 et octobre 2008 (BTEX et/ou organochlorés).
 - 128 relevés de côte de nappe entre octobre 2004 et février 2009.
- 6 piézomètres entre 12 et 20m de profondeur mis en place lors de l'étape de pré-diagnostic :
 - 10 échantillons d'eaux souterraines analysés entre février et octobre 2008 (organochlorés).
 - 92 relevés de côte de nappe entre février 2008 et février 2009.
- 13 piézomètres entre 7 et 15m de profondeurs réalisés lors du diagnostic du site.
 - 12 échantillons d'eaux souterraines analysés en novembre 2008 (organochlorés).
 - 101 relevés de côte de nappe entre novembre 2008 et février 2009.
- 52 sondages de l'ordre de 5m de profondeur réalisés lors du diagnostic du site (début 2009) dont 15 ont été équipés en piézomètres.
 - 15 échantillons d'eaux souterraines analysés en laboratoire (organochlorés) dont 2 ont aussi analysés les métaux lourds et les phénols.

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

- 29 relevés de cote de nappe entre janvier et février 2009.
- 319 échantillons de sols prélevés le long de 50 d'entre eux (hydrocarbures et/ou organochlorés et/ou solvants polaires) (Figure 26).
- 750 mesures de PID réparties tous les 20cm sur certains tronçons entre 0 et 5m.

Les échantillons de sols et les mesures PID ont été réalisés lors de la mise en œuvre des ouvrages.

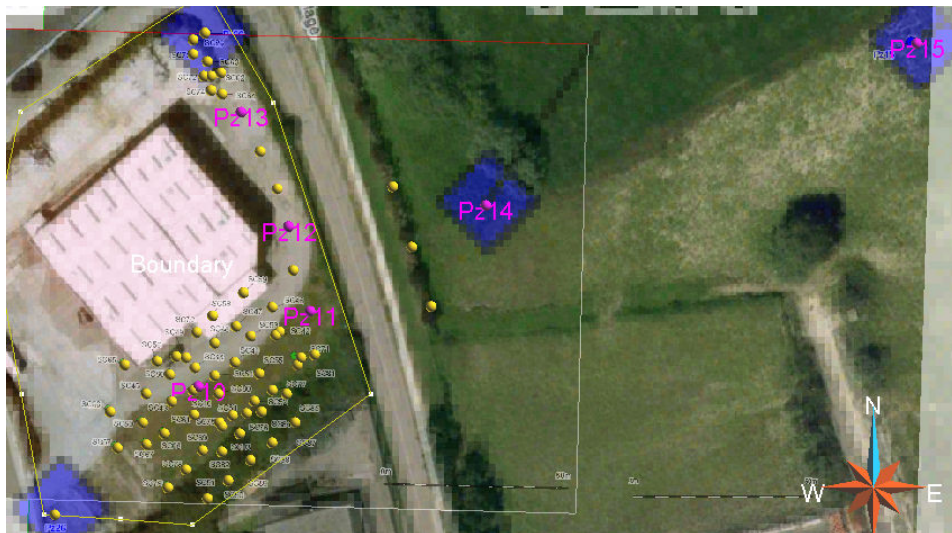


Figure 25 : Position des ouvrages récents en fonction de leur campagne : Pré-diagnostique – Diagnostique.

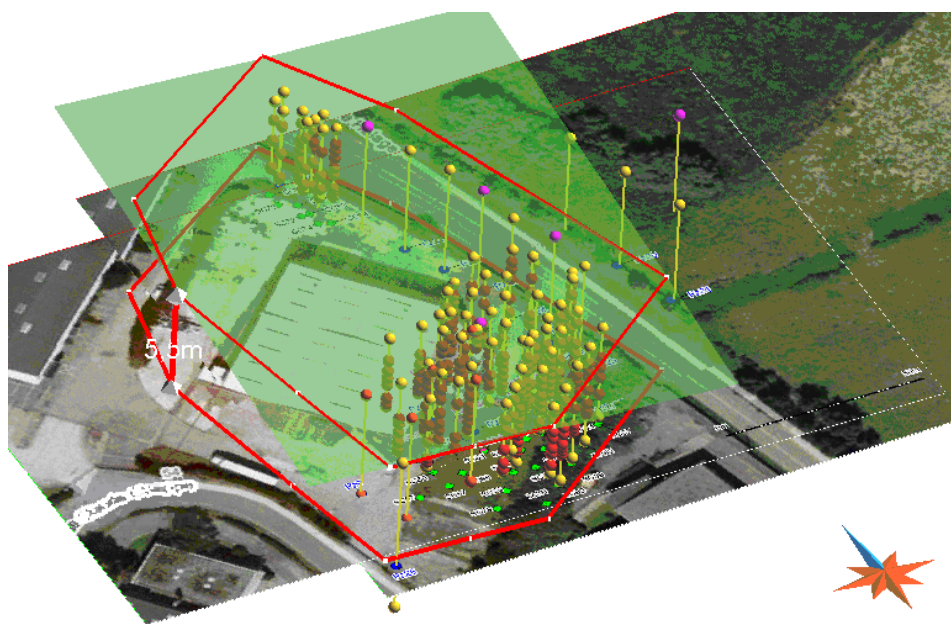


Figure 26 : Vue en perspective de la répartition spatiale des 319 échantillons de sol (sphères rouges) analysés lors de la campagne de diagnostic. Les sphères violettes et jaunes, qui délimitent les lignes de sondage, renseignent sur les campagnes de reconnaissance (pré-diagnostique en violet, diagnostique en jaune). La surface en vert est une estimation de la topographie du site établie à partir des données topographiques disponibles au droit des sondages.

3.3 Définition du cadre de l'étude de démonstration

Le cadre de l'étude de démonstration a été défini comme suit en concertation avec Burgeap.

- Trois composants corrélés entre eux ont été sélectionnés pour étudier la pollution des sols.
- Compte tenu de leur nombre inhabituellement important, les données ont été divisées par Burgeap en deux jeux de données dont un a servi à générer le modèle géostatistique (données du modèle initial) et l'autre à en vérifier les prédictions (données de validation).
- Le modèle géostatistique a été exploité pour estimer le volume et la position de la « zone source » dans laquelle une phase organique est susceptible d'exister.

Ces aspects sont détaillés ci-après.

3.3.1 Sélection des composants d'intérêt

Parmi les 50 composants chimiques analysés dans les sols, tous n'ont pas été mesurés un même nombre de fois. Certains l'ont été très peu (< 10 fois), d'autres moyennement (entre 40 et 80 fois) et les derniers beaucoup (environ 280 fois). Les résumés statistiques de ces données sont présentés en annexe (**Tableau 27** à **Tableau 29**) en distinguant les composants en fonction du nombre de mesures.

L'étude géostatistique devant servir de démonstration à l'application de la géostatistique aux sites et sols pollués, le choix des composants à modéliser s'est porté sur ceux qui ont été les plus mesurés et présentent des teneurs suffisamment élevées pour être assimilées à une pollution des sols. En favorisant les polluants étudiés dans le projet ATTENA (les COHV), le nombre de composants d'intérêt se réduit à 3 :

- Dichloréthylène-cis (DCE-cis abrégé DCE dans la suite)
- Trichloréthylène (TCE)
- Perchloréthylène (PCE).

3.3.2 Constitution du jeu de donnée restreint

Ce site de recherche a été très largement échantillonné et les nombreux échantillons de sols collectés ont tous été analysés. Afin de mener à bien la démonstration de l'intérêt de la géostatistique, il a été décidé avec Burgeap de ne prendre en compte qu'une partie des données disponibles dans l'étude géostatistique proprement dite et d'utiliser les autres données comme données de validation pour juger des prédictions du modèle sur l'état de pollution des sols.

Le jeu de données restreint a été établi par un ingénieur d'étude de Burgeap, étranger au projet, à partir uniquement de photos aériennes du site en activité et des campagnes de suivi de l'impact du site sur la nappe sous-jacente. Sur les 77 sondages disponibles, vingt ont été sélectionnés dont la répartition correspond aux pratiques usuelles dans le domaine. Sur ces 20 sondages, les règles d'élection des données ont été les suivantes.

- Les mesures PID sont divisées de moitié (soit une mesure tous les 40 cm).
 - Le choix des mesures PID a été dicté en partie par les échantillons de sols sélectionnés de façon à conserver les mesures correspondant à ces échantillons.
- Un échantillon au moins est conservé tous les mètres.
 - La préférence a été donnée, chaque fois que possible (cas où plusieurs échantillons sont disponibles), à l'échantillon associé à la mesure de PID la plus élevée sur un même mètre.

3.3.3 Définition de la « zone source »

Dans le cadre du projet ATTENA, les travaux menés s'intéressent à déterminer la « zone source » de la pollution des sols et qui est définie comme étant constituée des sols dont la somme des teneurs en DCE, TCE et PCE est suffisamment grande pour qu'une phase organique soit présente dans le milieu poreux. Les seuils de teneur totale au-dessus desquels la phase organique est présente se calculent à partir de lois d'équilibre thermodynamique en conditions triphasiques qui dépendent de la lithologie de la matrice (nature et propriétés des sols). La position et l'extension de cette zone sont déterminées expérimentalement, à partir des nombreuses données disponibles.

En appliquant ces lois aux teneurs simulées par le modèle géostatistique, il peut être quantifié l'incertitude sur le volume de la zone source (incertitude dite globale) et celle relative à la position de cette zone (incertitude dite locale). L'extension probable de la zone source prédite par le modèle géostatistique peut alors être comparée à celle observée sur les données de validation.

4. ETUDE GEOSTATISTIQUE – MODELE GEOSTATISTIQUE DE BASE

4.1 Analyse exploratoire et statistique des données

4.1.1 Jeu de donnée restreint

Les données de reconnaissance sélectionnées se répartissent comme suit :

- 7 ouvrages historiques d'environ 6m de profondeur réalisés en Octobre 2004.
 - 9 échantillons de sols analysés (hydrocarbures, organochlorés et/ou solvants polaires).
 - 163 relevés de côte de nappe entre octobre 2004 et octobre 2009.
- 13 ouvrages « récents » d'environ 5m de profondeur réalisés lors du diagnostic du site début 2009.
 - 13 relevés de côte de nappe entre janvier et juin 2009.
 - 49 échantillons de sols analysés (hydrocarbures, organochlorés et/ou solvants polaires).
 - 89 mesures de PID réparties tous les 40cm sur certains tronçons entre 0 et 5m.

Les échantillons de sols et les mesures PID ont été réalisés lors de la mise en œuvre des ouvrages.

Toutefois, sur les ouvrages anciens, les méthodes de prélèvement et de conservation ne sont pas aussi bonnes que ceux réalisées lors de la campagne « récente ». Ils n'ont donc pas été pris en compte lors de l'étude statistique qui suit.



Figure 27 : Délimitation du domaine d'étude (ligne rouge) et positions des ouvrages sélectionnés (sphères rouges).

4.1.2 Etapes de l'analyse exploratoire et statistique des données

L'analyse exploratoire et statistique des données a comporté les étapes suivantes.

1. Analyse unidimensionnelle des teneurs et mesures PID : étude et si nécessaire correction des biais d'échantillonnage, calcul de statistiques relatives à chaque variable, détermination de la complexité des distributions à prendre en compte dans les modèles géostatistiques.
2. Analyse des corrélations entre composants : détermination du degré et de la complexité des corrélations, confirmation de la nécessité ou non à simuler les composants conjointement.
3. Analyse des corrélations entre mesures PID et teneurs : détermination de l'apport d'information des mesures PID sur les teneurs, confirmation de l'intérêt ou non à exploiter ces données comme information secondaire sur les teneurs.
4. Extrapolation des teneurs en TCE et DCE au-dessous des limites de détection : utilisation des données de PCE pour extrapoler les teneurs en TCE et DCE sous les limites de quantification analytiques afin d'en faciliter la transformation gaussienne (voir étape suivante).
5. Transformation gaussienne des teneurs : étape nécessaire à la simulation géostatistique des teneurs par une méthode de simulation séquentielle gaussienne.
6. Analyse de la corrélation spatiale des teneurs : détermination des modèles de variogramme à utiliser pour simuler conjointement les teneurs par une approche géostatistique.

4.1.3 Analyse unidimensionnelle des teneurs et mesures PID

Les données n'étant pas régulièrement réparties sur le site ni le résultat d'un échantillonnage systématique des sols car conditionné par les mesures PID (voir §3.3.2), le biais d'échantillonnage, dû à des regroupements de données ou à une sélection préférentielle des échantillons analysés, a été étudié pour en corriger les effets possibles sur les statistiques déduites des données. Sans entrer dans le détail qui sort du cadre de ce rapport, cette correction consiste à pondérer les données en fonction de leurs distances par rapport aux autres : plus une donnée se situe à proximité d'autres, moins elle a de poids. Une telle correction est nécessaire et se justifie du fait que les données présentent une corrélation spatiale qui se traduit par une redondance possible de l'information.⁴ La correction du biais d'échantillonnage a été effectuée à l'aide d'une méthode dite par blocs. Elle a conduit à des poids qui varient entre 0.4 et 2.4 et confirment des regroupements de données. Les diagrammes de dispersion de la Figure 28 renseignent sur la répartition des poids en fonction des teneurs. Les poids les plus grands portent sur des teneurs faibles à fortes, les poids les plus faibles sur des teneurs plutôt fortes et les poids intermédiaires touchent les teneurs faibles à très fortes. Cette répartition des poids pourrait s'interpréter comme un léger échantillonnage préférentiel des zones plutôt polluées.

Sauf indication contraire, les statistiques présentées dans la suite portent sur des données dégroupées. Le Tableau 7 contient les statistiques élémentaires de chacun des composants sélectionnés. On remarquera qu'une part non négligeable des teneurs est indéterminée car au-dessous des seuils de détection des instruments de mesure utilisés (de 5 à 20% des données selon le composant). Ces statistiques ont été calculées en affectant la limite de quantification aux données au-dessous de cette limite (choix le plus pénalisant).

Les écarts importants, jusqu'à deux ordres de grandeur, entre les moyennes et les médianes, associés aux coefficients de variation très supérieurs à 1, révèlent des distributions fortement asymétriques vers les valeurs fortes avec une grande proportion de teneurs faibles et quelques mesures très élevées (Figure 29).

⁴ L'apport d'information de deux données est d'autant plus faible que les données sont proches et que leur corrélation augmente.

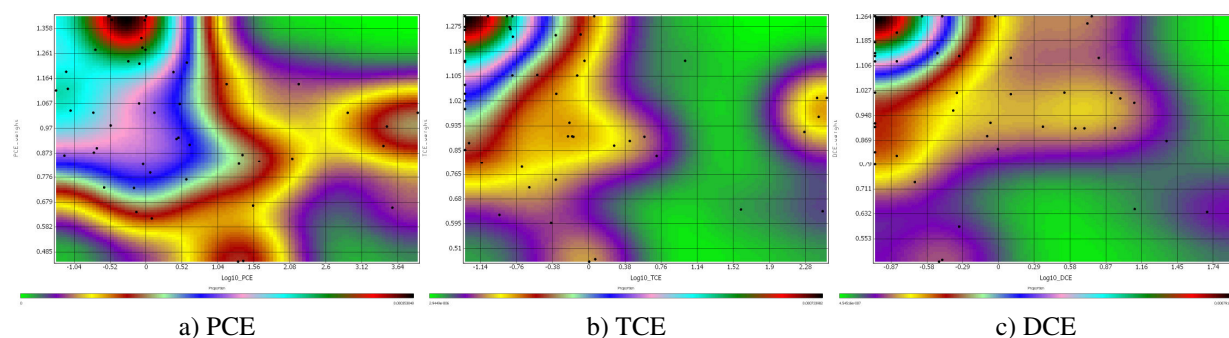


Figure 28 : Diagrammes de dispersion entre poids de dégroupage (vertical) et données de teneur (horizontal en échelle log) pour les trois composants. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

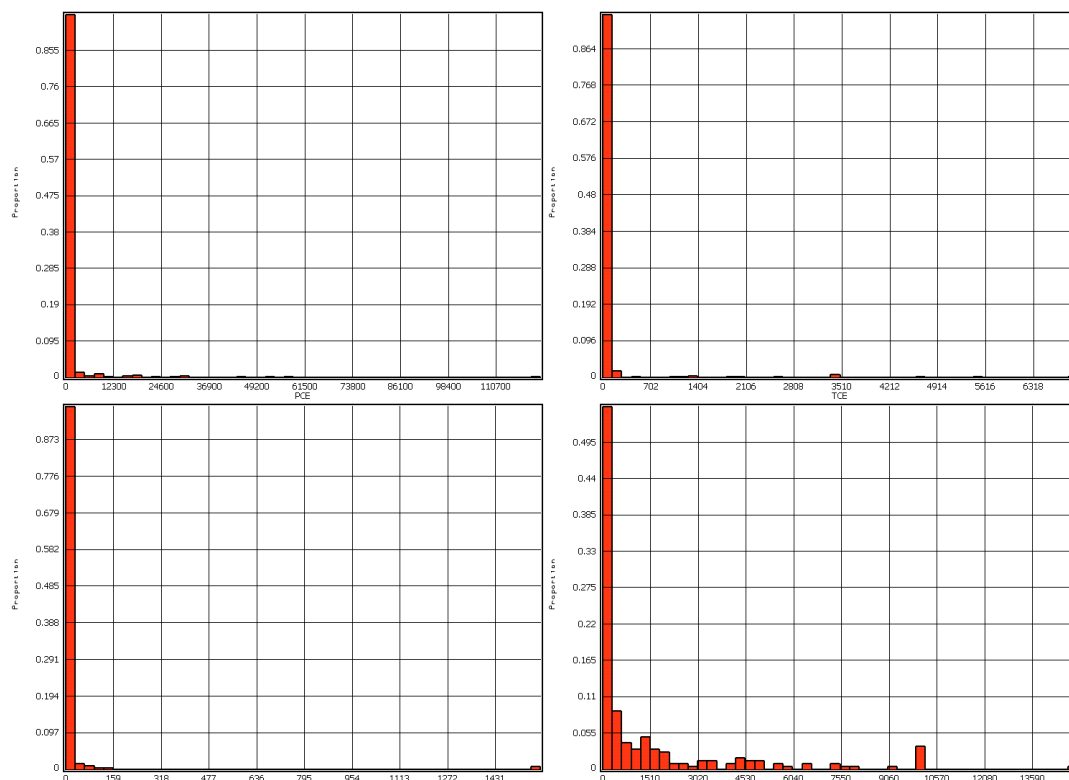


Figure 29 : Histogrammes des teneurs en PCE (en haut à gauche), TCE (en haut à droite) et DCE (en bas à gauche) et du PID (en bas à droite) des données du modèle géostatistique (unité : ppm).

Tableau 7 : Résumé statistique des teneurs corrigées du biais d'échantillonnage pour les composants sélectionnés et de la mesure PID (unité : ppm).

	Mesures	Mesures > LD	Moyenne	Ecart-type	Coef. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LQ)
PCE	49	48	363.7	1 372	3.77	0.05	8 460	0.29	0.95	7.43	<0.05
TCE	49	38	25.6	76.8	2.85	0.05	318	0.05	0.24	1.03	<0.05
DCE-cis	49	32	4.03	11.5	2.86	0.1	85.2	0.1	0.34	3.74	<0.15
PID	89	89	626.7	1261	2.01	0	6 896	7.07	35.29	524.97	-

4.1.4 Corrélations entre composants

La modélisation géostatistique d'une pollution de sols causée par plusieurs composants nécessite de simuler les composants conjointement en prenant en compte les corrélations entre eux. En procédant ainsi, l'état de pollution des sols peut être estimé plus précisément et l'incertitude sur la classification des sols comme pollués ou sains être réduite.

Compte tenu des distributions très asymétriques des données et de la proportion non négligeable de teneurs en TCE et DCE sous les limites de quantification, les corrélations ont été étudiées en échelle log et l'analyse a été répétée en écartant les données de TCE et DCE sous les limites de quantification.⁵ Les corrélations ont également été calculées en éliminant quelques données jugées singulières (voir ci-après). Les coefficients de corrélation linéaire (Pearson) et de rang (Spearman) correspondants sont indiqués dans le Tableau 8 (voir définition en Annexe, §13.1). Les diagrammes de dispersion de la Figure 30 illustrent les relations entre les composants.

Sur ces diagrammes, trois points, liés à des teneurs en DCE, apparaissent comme étant situés à l'écart du reste du nuage. En supprimant ces données, les corrélations entre le DCE et les autres composants sont augmentées de près de 50% (Tableau 8, colonnes 4 et 5).

Ces trois données peuvent être qualifiées de singulières. Elles ne reflètent cependant pas des situations identiques et peuvent s'interpréter comme suit (voir logs de sondages en Annexe, §12).

- **Point entouré en bleu :** cette donnée provient de Sc50 et correspond à une valeur de DCE anormalement faible par rapport à PCE et TCE. Aucune explication physique satisfaisante n'a pu être trouvée pour justifier cette donnée singulière.
- **Points entourés en rouge :** ces données proviennent de PzSc44 et Sc53 et montrent des teneurs en DCE anormalement supérieures à celles en PCE et TCE. Sur PzSc44, l'échantillon de sol se situerait dans la zone saturée, ce qui pourrait être une explication. Cette dernière ne peut cependant pas être confirmée par Sc53 dont les cotes de nappe sont inconnues (non communiquées).

Note : le Cis-DCE est un produit de la biodégradation du PCE et du TCE en zone anaérobie. Les conditions de biodégradation sont vraisemblablement plus favorables sous la nappe qu'au-dessus, ce qui pourrait justifier ces données singulières (Précision technique apportée par BURGEAP).

Afin de dégager les caractéristiques générales de la pollution, il a été vérifié l'impact de ces données sur l'analyse variographique de la teneur en DCE (§4.1.8). Cet impact s'est finalement révélé négligeable. Sans pouvoir lier ces données à une quelconque erreur de mesure, il a été choisi de les conserver pour simuler les teneurs en DCE dans les sols.

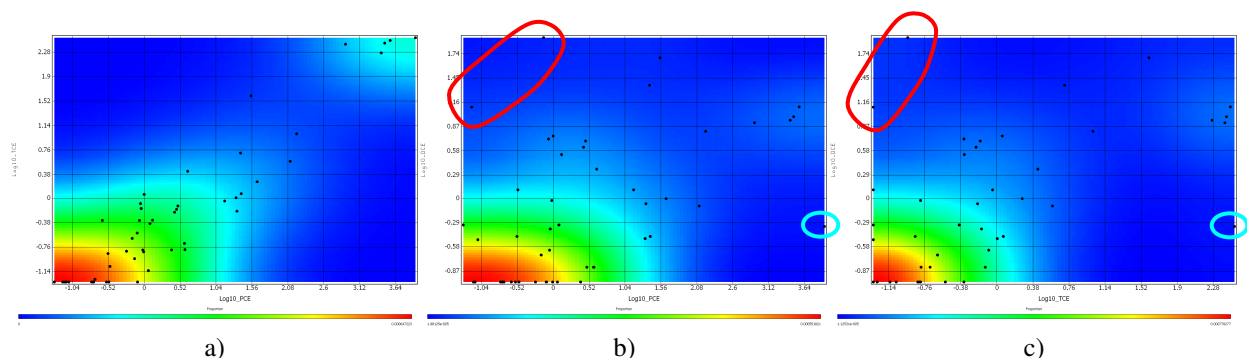


Figure 30 : Diagrammes de dispersion en échelle log entre a) TCE et PCE, b) DCE et PCE et c) DCE et TCE pour les données de validation. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés).

⁵ Les données sous les limites de quantification présentent l'inconvénient de ne pas pouvoir être ordonnées. Quand leur proportion est grande, ces données tendent à dégrader les corrélations entre variables et en particulier les corrélations d'ordre (voir Annexes site ATTENA, §13.1).

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

Tableau 8 : Coefficients de corrélation linéaire et de rang entre la transformée Log10 des teneurs de chacun des composants pour toutes les données, sans les données sous les limites de quantification (LQ), sans les données singulières et sans les données sous LQ et singulières.

	Ensemble des données	Sans données sous LQ	Sans données singulières	Sans données sous LQ et singulières
PCE - TCE	Données : 49 Linéaire : 0.94 Rang : 0.89	Données : 38 Linéaire : 0.92 Rang : 0.84	Données : 46 Linéaire : 0.93 Rang : 0.87	Données : 26 Linéaire : 0.91 Rang : 0.86
PCE – DCE	Données : 49 Linéaire : 0.48 Rang : 0.51	Données : 32 Linéaire : 0.26 Rang : 0.31	Données : 46 Linéaire : 0.67 Rang : 0.61	Données : 26 Linéaire : 0.51 Rang : 0.55
TCE – DCE	Données : 49 Linéaire : 0.60 Rang : 0.57	Données : 28 Linéaire : 0.40 Rang : 0.40	Données : 46 Linéaire : 0.78 Rang : 0.72	Données : 26 Linéaire : 0.69 Rang : 0.66

4.1.5 Corrélations entre teneurs et mesures PID

Les mesures PID sont supposées être proportionnelles à la somme de tout ou partie des teneurs en PCE, DCE et TCE ainsi que des autres composants organiques volatiles présents dans les sols. La prise en compte de ces données dans les modèles géostatistiques supposerait d'intégrer des corrélations entre ces mesures et la sommes des teneurs pour les composants simulés. Sans disposer immédiatement de méthode de simulation adaptée à ce type de corrélation complexe, seules les corrélations directes entre les mesures PID et chacun des composants ont été étudiées.

Comme précédemment, les corrélations ont été calculées à partir des seules données de teneur supérieures aux limites de quantification des différents composants. La Figure 31 montre les diagrammes de dispersion entre teneurs et mesures PID. Les coefficients de corrélation linéaire et de rang correspondants, ainsi que ceux entre le PID et la somme des teneurs, se trouvent dans le Tableau 9.

Ces résultats montrent des corrélations très faibles (< 0.5) qui ne justifient pas d'utiliser les mesures PID comme information secondaire (auxiliaire) dans la simulation géostatistique des teneurs. Ces données n'ont pas été exploitées davantage.

Note : Les précisions techniques suivantes ont été apportées par BURGEAP.

La relation ou corrélation entre le PID et les teneurs dans les sols ($[T_i]$) dépend de la teneur totale en composés organiques volatils ($[T_{cov}]$). En effet, pour chaque composé i présent,

- si $[T_{cov}] > \text{seuil source}$: $PID \propto \sum_i P_{vap_i} \times [T_i]$, avec P_{vap_i} = pression de vapeur saturante du composé i ;
- si $[T_{cov}] < \text{seuil source}$: $PID \propto \sum_i H_i \times \rho_i \times [T_i]$, avec H_i = constante d'Henry du composé i .

La mesure PID est ainsi proportionnelle à la somme (pondérée) de teneurs en composés organiques dans les sols et peut donc être corrélée aux teneurs mais de façon non linéaire, non homogène ou les deux à la fois : selon que $[T_{cov}]$ est inférieur ou supérieur au seuil d'apparition d'une phase organique, la relation est différente et le seuil d'apparition peut varier en fonction du type de sols.

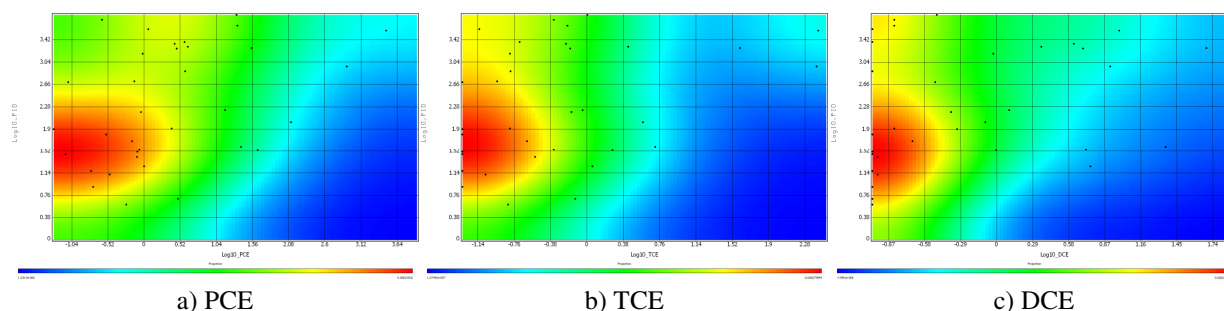


Figure 31 : Diagrammes de dispersion en échelle log entre teneurs (axe vertical) et mesures PID (axe horizontal) pour les trois composants d'intérêt (unité : ppm). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés).

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

Tableau 9 : Coefficients de corrélation linéaire et de rang entre la transformée Log10 des teneurs et des mesures PID pour toutes les données et sans les données sous les limites de quantification (LQ).

	Ensemble des données	Sans données sous LD
PID – PCE	Données : 34 Linéaire : 0.42 Rang : 0.47	Données : 33 Linéaire : 0.42 Rang : 0.48
PID – TCE	Données : 34 Linéaire : 0.36 Rang : 0.36	Données : 27 Linéaire : 0.25 Rang : 0.15
PID – DCE	Données : 34 Linéaire : 0.32 Rang : 0.38	Données : 24 Linéaire : 0.18 Rang : 0.16
PID – Σ (PCE, TCE et DCE)	Données : 34 Linéaire : 0.42 Rang : 0.47	Données : 22 Linéaire : 0.32 Rang : 0.29

4.1.6 Extrapolation des teneurs en TCE et DCE au-dessous des limites de quantification

La proportion non négligeable de teneurs en TCE et DCE situées au-dessous des limites de quantification complexifie et peut rendre inefficace la transformation gaussienne de ces données qui est nécessaire à la mise en œuvre de l'approche géostatistique suivie (voir §4.1.7). En effet, sans pouvoir ordonner convenablement ces données, qui selon toute vraisemblance devraient varier entre 0 et leur limite de quantification, cette transformation échoue à produire une variable transformée qui suit véritablement une distribution gaussienne : après transformation, les données au-dessous de la limite de quantification d'un composant continuent d'avoir une valeur identique dans une même proportion.

Tel n'est pas le cas du PCE dont la teneur « exacte » a pu être mesurée sur tous les échantillons sauf un. Compte tenu des fortes corrélations linéaires et de rang observées précédemment entre le PCE et le TCE (§4.1.4), il peut être justifié, à défaut d'autre moyen, de prendre les données de PCE comme référence pour réordonner les teneurs en TCE sous les limites de quantification. Une simple loi linéaire a été utilisée pour extrapoler ces teneurs. De la même façon, TCE et DCE présentant d'assez bonnes corrélations, les données de TCE, extrapolées à partir de PCE, ont été utilisées pour ordonner les teneurs en DCE sous les limites de quantification. Ces corrections ne portent pas préjudice au modèle car elle ne concerne que les teneurs inférieures aux limites de quantification et qui sont donc en principe très faibles. Elle tend simplement à augmenter un peu et de façon artificielle la corrélation entre les composants. En procédant ainsi, en revanche, les transformations gaussiennes peuvent être opérées convenablement pour l'ensemble des composants.

4.1.7 Transformation gaussienne des teneurs

La méthode de simulation géostatistique utilisée est une méthode de simulation séquentielle gaussienne qui suppose que les variables simulées suivent une distribution gaussienne et qu'elles soient corrélées linéairement pour pouvoir être simulées conjointement (voir par exemple Manuel méthodologique GéoSiPol, 2005). L'asymétrie des distributions de teneurs vers les valeurs fortes, telle qu'établie précédemment (§4.1.3) et qui est caractéristique des phénomènes de pollution, ne répond pas à cette condition. Pour s'en affranchir, la pratique consiste à opérer une transformation gaussienne des données pour les variables d'intérêt, de simuler les variables transformées et d'effectuer une transformation inverse de ces simulations pour en tirer des simulations des variables d'intérêt.

Dans le cas présent, les données de PCE, TCE et DCE ont été transformées indépendamment, après extrapolation des teneurs en TCE et DCE inférieures aux limites de quantification, à partir des distributions de probabilité établies en utilisant les données dégroupées (voir §4.1.3). Il peut être vérifié à partir des diagrammes de dispersion de la Figure 32 et des coefficients de corrélation du Tableau 10 que les variables transformées présentent des corrélations relativement linéaires et qui justifient de pouvoir les simuler conjointement.

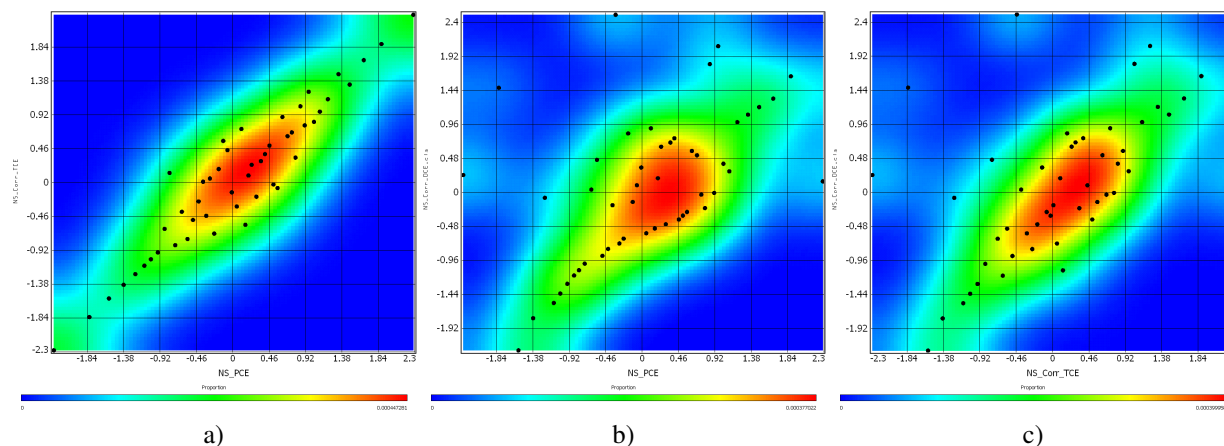


Figure 32 : Diagrammes de dispersion entre a) TCE et PCE, b) DCE et PCE et c) DCE et TCE après transformation gaussienne des données. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés).

Tableau 10 : Coefficients de corrélation linéaire et de rang des teneurs en PCE, TCE et DCE après transformation gaussienne des teneurs. Les corrélations avec DCE sont calculées sans les trois données singulières.

	Données	Coefficient de corrélation linéaire	Coefficient de corrélation de rang
PCE – TCE	49	0.95	0.91
PCE – DCE	46	0.71	0.71
TCE – DCE	46	0.75	0.77

4.1.8 Corrélation spatiale des teneurs

Compte tenu des corrélations fortes observées entre les composants et de leur même niveau d'information (nombre et positions des données identiques), les choix suivants ont été effectués pour simuler les teneurs.

1. Ordre de simulation des composants en chaque point de simulation : PCE suivi de TCE suivi de DCE. Le PCE est simulé en premier car mieux défini (une seule donnée sous la limite de quantification).
2. Corrélations entre composants prises en compte dans la simulation : TCE avec PCE, DCE avec TCE. La relation entre PCE et DCE a été ignorée car n'apportant pas d'information complémentaire (données redondantes).
3. Méthode de simulation : co-simulation séquentielle gaussienne avec cokrigage colocalisé. Le cokrigage colocalisé signifie ici que, en chaque point de simulation, seule la valeur de PCE au point de simulation est utilisée pour conditionner la simulation de TCE et la valeur simulée de TCE celle de DCE.

Dans ces conditions, l'analyse variographique a porté sur les données transformées (transformation gaussienne) et a conduit à modéliser les variogrammes directs des trois composants et les variogrammes croisés entre PCE et TCE et entre TCE et DCE. Les variogrammes croisés ont été déduit des variogrammes directs en les multipliant simplement par les coefficients de corrélation entre composants.

Il est à noter que l'analyse variographique du DCE a été répétée avec et sans les données singulières mises en évidence au cours de l'analyse des corrélations entre composants (§4.1.3). Leur impact étant négligeable, ces données n'ont finalement pas été écartées.

Il ressort de l'analyse des variogrammes expérimentaux directs les résultats suivants (Figure 33).

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

- La continuité spatiale latérale des teneurs apparaît plus clairement en utilisant les profondeurs, plutôt que les élévations, pour positionner verticalement les échantillons.
- Une anisotropie dite zonale existe pour tous les polluants. Elle se traduit par une variabilité contrastée (plateaux des variogrammes différents) latéralement et verticalement, la variabilité étant logiquement plus faible sur la verticale.
- Compte tenu du faible nombre de sondages, des variogrammes omnidirectionnels (isotropes) ont été calculés latéralement.
- Les variogrammes expérimentaux ont pu tous être calés à l'aide d'un modèle sphérique associé à un effet de pépité. Une deuxième structure sphérique a dû être introduite, avec une portée de corrélation très grande dans la direction verticale, pour reproduire l'anisotropie zonale.
- Les portées de corrélation sont de 25 m latéralement et de 1 m verticalement. Elles s'appliquent aux trois composants.

Les modèles de variogramme calés sur les variogrammes expérimentaux sont de la forme suivante où h désigne la distance entre points, C_0 l'effet de pépité, C_1 et C_2 la contribution des deux structures sphériques :

$$\begin{aligned} \text{Plan horizontal :} \quad & \gamma(h) = C_0 + C_1 \text{Sph}_{25}(h) + C_2 \text{Sph}_{25}(h) \\ \text{Direction verticale :} \quad & \gamma(h) = C_0 + C_1 \text{Sph}_1(h) + C_2 \text{Sph}_{10000}(h) \end{aligned}$$

Les valeurs des coefficients pour les différents modèles de variogrammes se trouvent dans le Tableau 11.

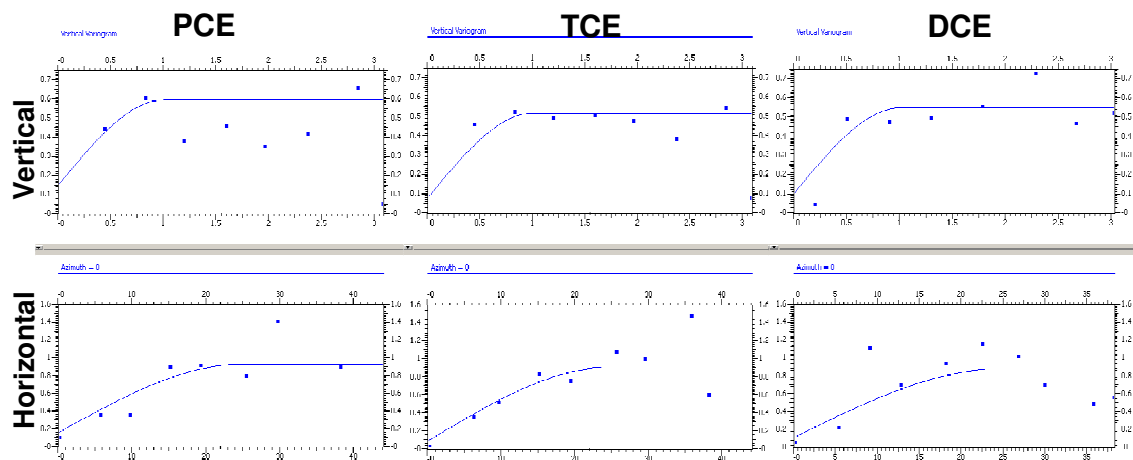


Figure 33 : Variogrammes expérimentaux (points) et modèles calés (lignes) pour le PCE, TCE et DCE (données transformées, profondeur en coordonnée verticale) établis à partir du jeu de données restreint.

Tableau 11 : Modèles de variogrammes directs et croisés calés sur les variogrammes expérimentaux. Les variogrammes croisés ont été déduits des variogrammes directs en prenant en compte une corrélation linéaire de 0.9 entre TCE et PCE et de 0.75 entre DCE et TCE.

Composant	Effet de pépité (C_0)	Structure 1	Structure 2
		Modèle : sphérique	Modèle : sphérique
		Portée horizontale : 25 m	Portée horizontale : 25 m
		Portée verticale : 1 m	Portée verticale : 10 000 m
		Seuil (C_1)	Seuil (C_2)
Variogrammes directs			
PCE	0.15	0.45	0.33
TCE	0.075	0.44	0.39
DCE	0.1	0.45	0.33
Variogrammes croisés			
TCE – PCE	0.072	0.40	0.35
DCE – TCE	0.075	0.34	0.25

4.2 Modélisation géostatistique à partir du jeu de donnée restreint

4.2.1 Définition du domaine

La grille utilisée pour la simulation géostatistique des teneurs en PCE, TCE et DCE comprend 255 102 mailles dont 193 492 sont des mailles actives situées dans la zone d'intérêt (Figure 34). Chaque maille représentant un volume de 0.4 m^3 ($2 \times 2 \times 0.1 \text{ m}^3$), le volume de la zone d'intérêt est de $77\,400 \text{ m}^3$.

La grille est attachée à la topographie du site qui a été estimée par krigeage (surface supérieure verte sur la Figure 34). Elle s'étend verticalement sur les 6 premiers mètres de profondeur.

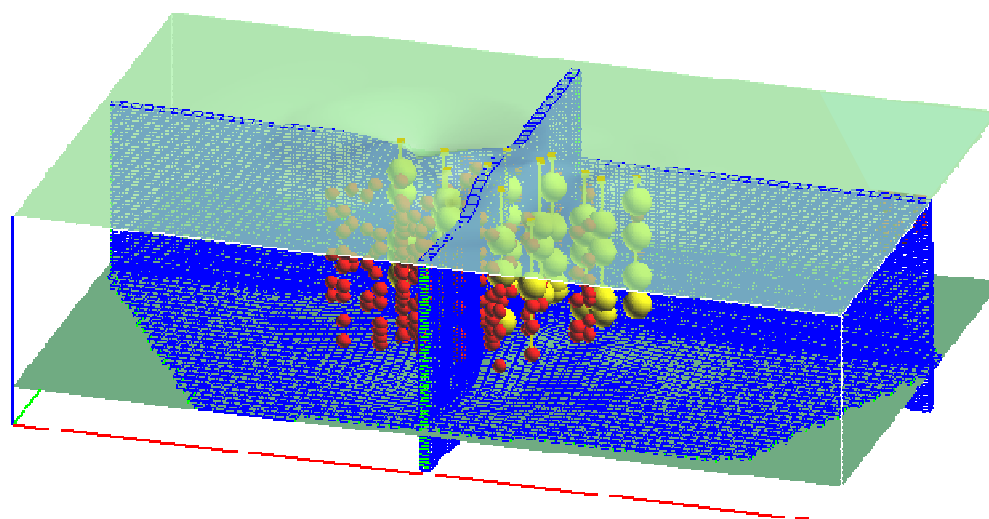


Figure 34 : Représentation 3D du domaine de simulation avec visualisation du jeu de données restreint du modèle géostatistique (sphères jaunes) et le jeu de données de validation du modèle (sphères rouges). La distorsion verticale de l'image est de 7 pour 1.

4.2.2 Simulation conjointe des teneurs en PCE, TCE et DCE

L'estimation du volume et de la localisation de la zone source est un problème d'estimation non linéaire qui requiert des méthodes géostatistiques spécifiques pour juger du dépassement du seuil de teneur au-delà duquel une phase organique constituée de PCE, TCE et DCE peut exister. Parmi les méthodes possibles (krigeage d'indicateur, krigeage disjonctif, simulations conditionnelles), les simulations conditionnelles ont été choisies en raison de leur souplesse d'utilisation, en particulier pour modéliser conjointement les trois composants corrélés, et de leur exploitation possible (post-traitement) pour quantifier l'incertitude globale sur le volume de la zone source et locale sur sa position.

Les simulations ont été générées par une méthode dite de co-simulation séquentielle gaussienne avec cokrigeage colocalisé en utilisant un voisinage local de 25 m de rayon dans le plan horizontal et de 1 m verticalement, en rapport avec les portées de corrélation observées.⁶ Le nombre de données maximal pour simuler la teneur d'un composant en chaque nœud de la grille a été fixé à 9 (8 données directes avoisinantes + 1 donnée colocalisée de la variable auxiliaire pour TCE et DCE). Les modèles de variogramme utilisés sont ceux décrits au §4.1.8. Les simulations ont été réalisées dans l'espace des variables transformées (transformations gaussiennes des données, voir §4.1.7) puis soumises à une transformation inverse pour en tirer des cartes de teneurs simulées.

Au total, 100 simulations conjointes des teneurs ponctuelles (à l'échelle des échantillons de sols) pour les trois polluants ont été générées sur la grille décrite précédemment (§4.2.1). Ce nombre de simulations a été choisi après avoir vérifié sur un polluant qu'il était suffisant pour évaluer convenablement l'incertitude locale sur les teneurs (comparaison de statistiques sur les valeurs simulées pour un nombre de simulations de 100 et 200). Un nombre de simulations plus important aurait signifié des temps de calcul plus long et davantage de résultats à sauvegarder pour un gain négligeable sur la qualité des estimations.

4.3 Considérations méthodologiques

Ce paragraphe aborde certains aspects théoriques et méthodologiques destinés à des lecteurs qui souhaitent mettre en pratique la géostatistique.

Comme indiqué précédemment, la méthode qui a été employée pour simuler les teneurs en PCE, TCE et DCE est une méthode dite de *co-simulation séquentielle gaussienne avec cokrigeage colocalisé*.⁷ Les trois composants sont ainsi simulés séquentiellement en chaque nœud de la grille de simulation parcourue aléatoirement, le PCE en premier à partir des seules données de PCE avoisinantes, le TCE en deuxième à partir des données de TCE avoisinantes et de la valeur de PCE précédemment simulée, le DCE en troisième à partir des données de DCE avoisinantes et de la valeur de TCE précédemment simulée. En procédant ainsi, l'objectif est de reproduire les corrélations relativement fortes entre TCE et PCE et entre DCE et TCE pour classer les sols plus justement comme étant situés dans ou hors de la zone source.

En toute rigueur, la simulation conjointe des teneurs en PCE, TCE et DCE suppose de prendre en compte les corrélations spatiales entre tous les composants. En gardant le principe de simuler le PCE suivi du TCE suivi du DCE en chaque nœud de la grille de simulation, il s'agit de tenir compte :

1. des données et valeurs déjà simulées⁸ de PCE, TCE et DCE dans le voisinage du nœud pour simuler la teneur en PCE,

⁶ Le voisinage de recherche aurait pu être doublé pour prendre en compte plus complètement la redondance des données en augmentant également le nombre de données maximal de conditionnement. Ce choix aurait conduit cependant à favoriser inutilement les effets dits "d'écran" compte tenu du fort regroupement des données.

⁷ Pour information ou rappel, le cokrigeage colocalisé doit être réalisé avec une méthode de cokrigeage simple, le cokrigeage ordinaire ne permettant pas de prendre en compte l'information secondaire (sans effet car le poids de la donnée secondaire devient nul).

⁸ Pour rappel (voir Manuel Méthodologique GeoSiPol), la méthode de simulation est une méthode dite de simulation séquentielle qui consiste à parcourir dans un ordre aléatoire les nœuds de la grille de simulation pour

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

2. de ces mêmes données et valeurs déjà simulées, complétées de la teneur en PCE simulée au nœud, pour simuler la teneur en TCE,
3. de ces mêmes données et valeurs déjà simulées, complétées des teneurs en PCE et TCE simulées au nœud, pour simuler la teneur en DCE.

Un tel modèle a pour conséquences :

1. de devoir modéliser l'ensemble des variogrammes directes et croisés pour les trois composants, soit 6 modèles de variogramme (3 directs et 3 croisés), qui doivent constituer un modèle dit de *corégionalisation* acceptable,
2. de devoir prendre en compte un plus grand nombre de données et donc de devoir résoudre des systèmes de cokrigage plus grands.

En pratique, la modélisation de pollutions de sols peut nécessiter de simuler jusqu'à 10 ou 15 composants qui se répartissent en groupes de composants corrélés, les composants d'un même groupe devant être simulés conjointement pour reproduire les corrélations. Une alternative plus complète au cokrigage colocalisé est le *cokrigage multicolocalisé*. Cette méthode consiste ici à simuler en chaque nœud de la grille chacun des polluants successivement en prenant en compte l'ensemble des polluants en chaque point de données avoisinant (teneurs mesurées⁹ ou teneurs simulées précédemment) et les teneurs des polluants déjà simulés au nœud considéré.

Le nombre de modèles de variogramme à produire et le nombre de données maximal à prendre en compte dans le système de cokrigage à résoudre (nombre d'équations du système de cokrigage $\geq$ nombre de données de conditionnement) sont renseignés dans le Tableau 12 pour un nombre de polluants croissant. Il apparaît ainsi que le nombre de variogrammes peut devenir rapidement dissuasif et le nombre de données à prendre en compte dans le système de cokrigage devenir excessif. Dans ces conditions, le cokrigage colocalisé se révèle être une solution certes plus approximative mais qui est plus facilement réalisable. En outre, la non prise en compte des autres composants hors du nœud simulé (données avoisinantes relatives uniquement au composant simulé) se justifie, en partie au moins, par l'effet d'écran des données du composant simulé par rapport aux données des autres composants : dès lors que la forme et la portée des modèles de corrélation spatiale (variogrammes) sont relativement similaires, l'apport d'information locale sur un composant provient essentiellement de la donnée de ce composant, les données sur les autres composants apportant une information largement redondante.

Tableau 12 : Nombre de modèles de variogramme à établir et nombre maximal de données de conditionnement à prendre en compte dans le système de cokrigage en fonction du nombre de polluants à simuler conjointement par une méthode de cokrigage multicolocalisé.

Nombre de composants	Nombre de points de données maximal	Nombre de variogrammes à modéliser	Nombre maximal de données de conditionnement
1	8	1	8
2	8	3	17
3	8	6	26
5	8	15	44
10	8	55	87

simuler en chaque nœud les teneurs en PCE, TCE et DCE en fonction des données sur ces teneurs et des valeurs simulées dans les nœuds déjà visités.

⁹ Les polluants sont généralement mesurés aux mêmes points d'échantillonnage.

5. RESULTATS DU MODELE GEOSTATISTIQUE

5.1 Simulations des teneurs en polluants

5.1.1 Résultats

Les trois composants ont été simulés conjointement en prenant en compte les corrélations observées sur les teneurs (voir §4.1.4 et 4.1.8). Cent simulations de teneurs ont été générées pour chaque composant.

La Figure 35 montre deux simulations conjointes particulières des teneurs en PCE, TCE et DCE correspondant à un même plan de coupe horizontal. Pour une même simulation, on peut y observer des tendances similaires pour les trois composants du fait de leurs corrélations. D'une simulation à l'autre, la variabilité des teneurs pour un même composant traduit l'incertitude locale sur les teneurs, l'incertitude étant d'autant plus grande que les teneurs simulées sont éloignées des sondages.

La variabilité des simulations apparaît également dans les distributions de teneurs simulées, la proportion de valeurs fortes ou faibles pouvant varier de façon importante d'une simulation à l'autre. Le diagramme quantile-quantile (QQ-plot en anglais, voir Annexe, §13.2) de la Figure 36 illustre ce type de variabilité. Comparées à la distribution de référence fournie par les données, les distributions de teneurs simulées révèlent de fortes fluctuations qui s'expliquent par les dimensions réduites du site par rapport aux portées de corrélation observées.¹⁰

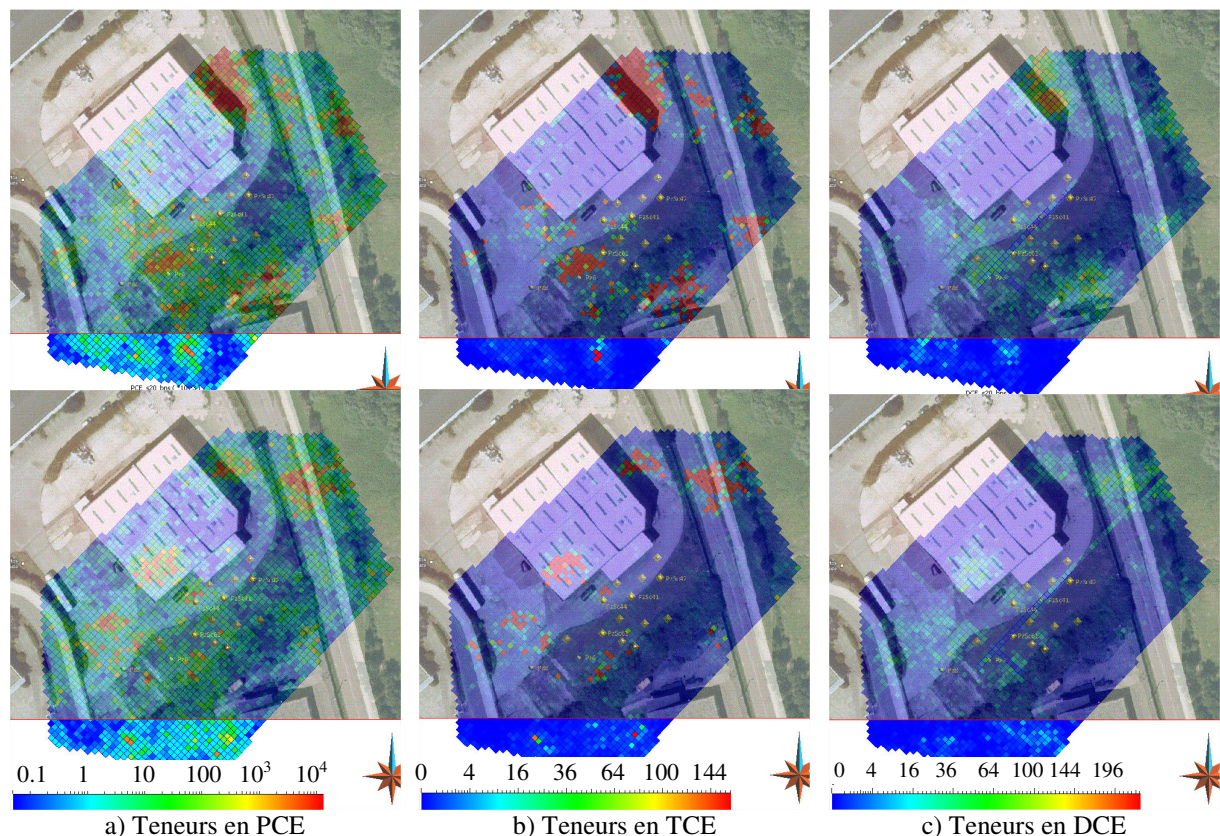


Figure 35 : Coupes horizontales de deux simulations conjointes des teneurs pour les trois composants (unité : ppm). Les sondages affichés sont ceux d'où proviennent les données de teneur utilisées pour conditionner les simulations.

¹⁰ Dans le jargon géostatistique, on parle de phénomène non ergodique.

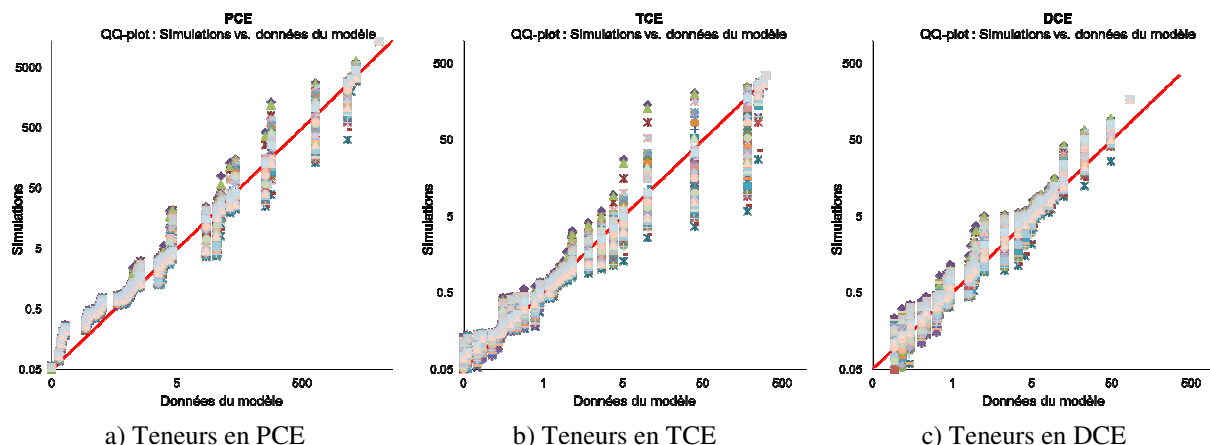


Figure 36 : Diagrammes quantile-quantile entre les teneurs simulées de chaque réalisation et les données de teneur utilisées pour conditionner les simulations. Les distributions de référence tirées des données sont corrigées des biais d'échantillonnage (données pondérées).

5.2 Délimitation de la « zone source » de la pollution

5.2.1 Critères de définition de la zone source

Dans le cadre du projet ATTENA, les sols sont jugés comme étant une source de pollution potentielle si les trois composants étudiés, PCE, TCE et DCE, forment ensemble une phase organique (phase liquide non aqueuse). Les conditions dans lesquelles une telle phase organique peut exister se calcule à partir des teneurs en ces trois composants dont la somme doit dépasser un certain seuil critique. Ce seuil dépend de la nature et des propriétés du sol ainsi que d'autres facteurs tels que la saturation en eau et la teneur en matières organiques du sol. Il est à noter que ce mode d'évaluation de l'état de pollution des sols nécessite que les composants soient simulés conjointement pour que, dans le calcul de la teneur totale, soient bien prises en compte leurs corrélations.

Pour le site étudié, des seuils critiques différents ont été calculés par Burgeap pour les limons, présents essentiellement sur les 4 premiers mètres de profondeur, et pour les sables rencontrés à plus grandes profondeurs. Ces calculs donnent les fourchettes de valeurs suivantes pour les deux milieux.

3. Limons : entre 470 et 670 ppm dans des conditions d'équilibre thermodynamique, seuils jusqu'à 2 à 4 fois plus petits dans d'autres conditions (non équilibre local).
4. Sables : entre 40 et 85 ppm.

Afin d'illustrer l'exploitation du modèle géostatistique pour estimer le volume et l'emplacement de la zone source, il a été choisi, en concertation avec Burgeap, d'appliquer un seuil unique partout et de le prendre suffisamment petit pour que la zone source ne soit pas trop restreinte. Plusieurs seuils ont été définis : 400, 250, 120, 60, 40 et 20 ppm. Les volumes de la zone source correspondants vont en moyenne de 6 000 m³ pour le seuil le plus grand (400 ppm) à 17 000 m³ pour le plus petit (20 ppm). Certains de ces seuils ont été utilisés également pour évaluer la position possible de la zone source et la confronter aux données de validation. Les résultats sur les volumes et la position de la zone source sont présentés aux paragraphes suivants.

5.2.2 Incertitude sur le volume de la zone source

Le volume de la zone source peut être calculé sur chaque simulation conjointe des trois composants en appliquant un des seuils définis précédemment. En répétant le calcul pour les 100 simulations, on obtient 100 valeurs de volume possibles à partir desquelles il peut être estimé l'incertitude sur le volume de la zone source.

La Figure 37 fournit les histogrammes et les résumés statistiques des volumes de la zone source ainsi calculés pour les 6 seuils de 400 à 20 ppm. Plus le seuil est petit, plus l'incertitude relative tend à diminuer. Pour le seuil de 400 ppm, le rapport volume maximal sur volume minimal est de 2.4, contre 1.8 pour le seuil de 20 ppm, et l'intervalle interquartile, correspondant à un intervalle de confiance de 50%, représente 25% de la médiane, contre un peu moins de 20% pour le seuil de 20 ppm.

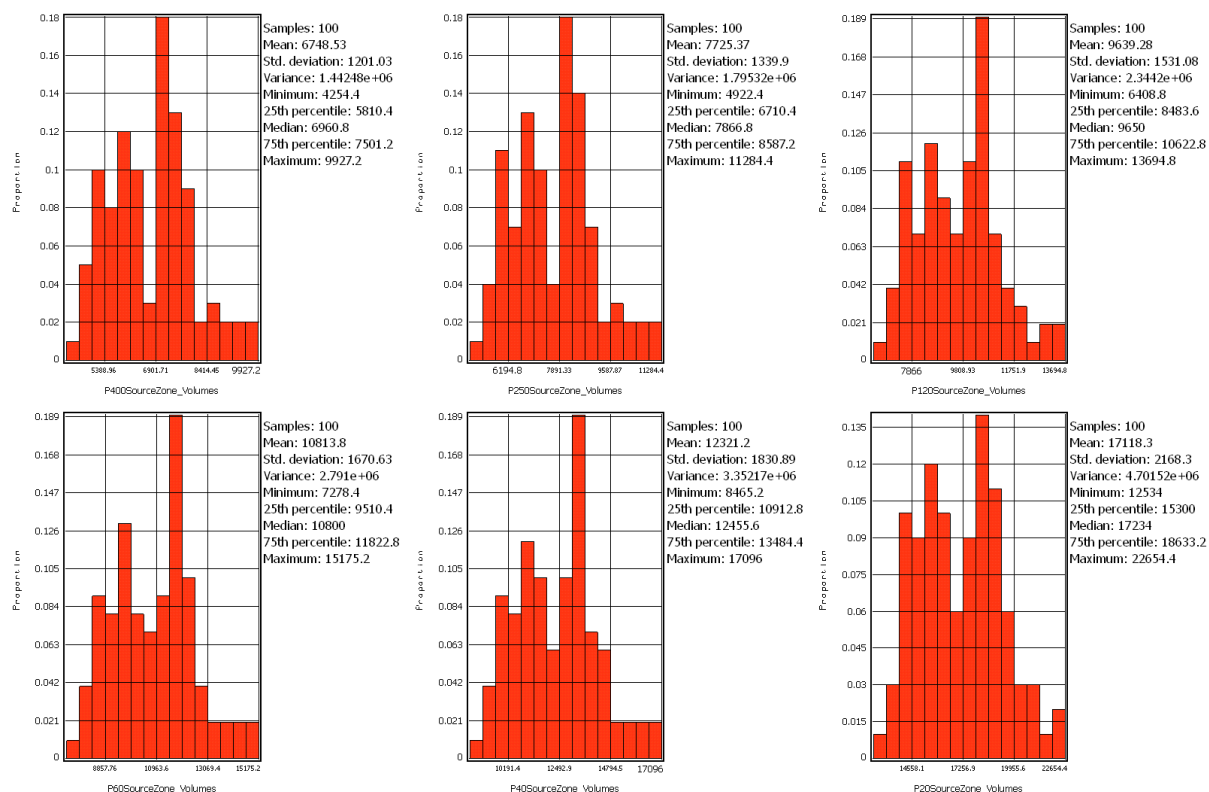


Figure 37 : Histogrammes des volumes de la zone source en m³ calculés à partir des 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas).

5.2.3 Incertitude spatiale sur la position de la zone source

En complément du volume, l'application du seuil de définition de la zone source aux simulations conjointes nous renseigne sur la position possible de cette zone. Pour chaque maille de la grille, il est en effet possible de calculer sa probabilité d'être située dans la zone source comme le rapport entre le nombre de simulations où la maille est dans la zone source sur le nombre de simulations total.

La Figure 38 montre trois coupes horizontales de la carte de probabilité de présence de la zone source calculée à partir des 100 simulations pour un seuil de teneur totale de 20 ppm. La Figure 39 fournit une vue en perspective de deux coupes verticales. Les probabilités les plus fortes (> 0.8) renseignent sur les endroits où il y a plus de 80% de chances de trouver une phase organique dans les sols pour ce seuil. Les probabilités les plus faibles (< 0.2) indiquent les lieux où cette phase organique a moins de 20% de chances d'exister. Entre les deux, l'incertitude sur la présence ou non d'une phase organique est d'autant plus grande que la probabilité est proche de 0.5 (50% de chances pour la phase organique d'être présente ou pas).

Il apparaît sur les coupes verticales deux zones de fortes probabilités qui pourraient correspondre à des spots de pollution plus ou moins dissociés. Ces zones se retrouvent sur les coupes horizontales.

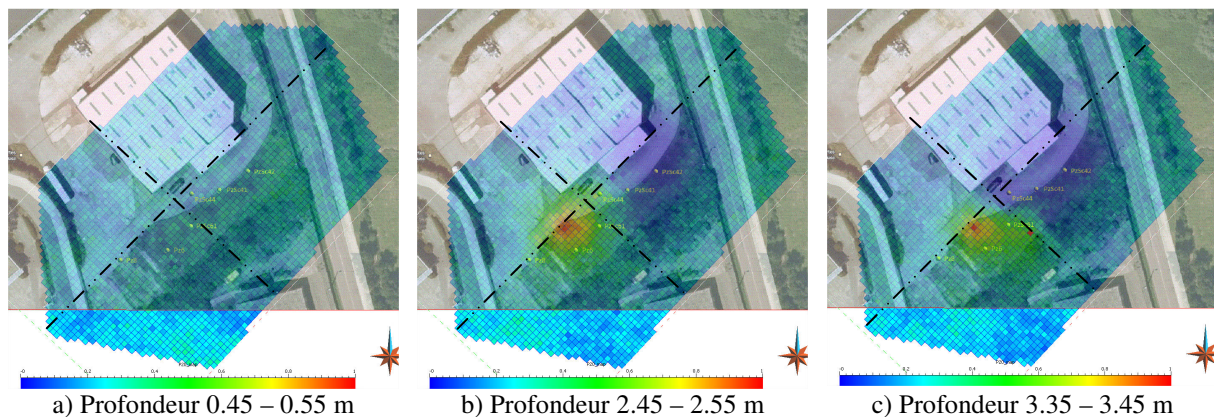


Figure 38 : Coupes horizontales de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE.

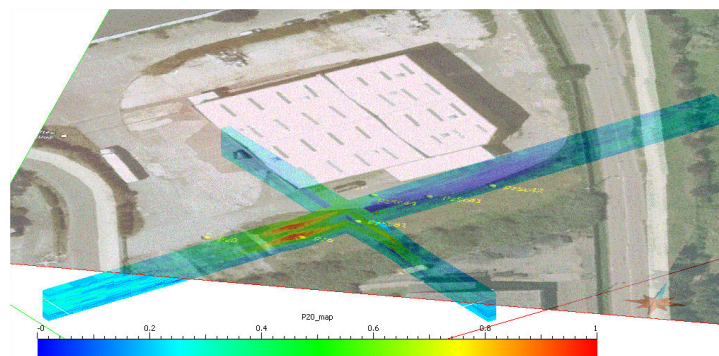


Figure 39 : Coupes verticales vues en perspective de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE.

Les histogrammes et les résumés statistiques des probabilités de présence de la zone source calculées pour les seuils de 400 à 20 ppm se trouvent à la Figure 40. Il peut en être tiré des informations sur le niveau de connaissance de la position de cette zone. Les mailles se répartissent ainsi en trois catégories.

- Les mailles que l'on peut classer avec confiance comme étant hors de la zone source. En se fixant par exemple un niveau de confiance de 80% (moins de 20% de chances d'erreur), il s'agit des mailles pour lesquelles la probabilité de dépassement du seuil critique est inférieure ou égale à 0.2. Pour le seuil de 400 ppm, ces mailles représentent 96.77% du volume, contre 44.4% pour le seuil de 20 ppm.
- Les mailles que l'on peut classer avec confiance comme situées dans la zone source. En conservant le même niveau de confiance de 80%, les mailles concernées sont celles pour lesquelles la probabilité de dépassement du seuil critique est supérieure ou égale à 0.8. Ces mailles représentent de 0.01% du volume pour le seuil de 400 ppm et seulement 0.2% pour le seuil de 20 ppm.
- Les mailles incertaines dont la probabilité de dépassement du seuil critique est supérieure à 0.2 et inférieure à 0.8. Avec les niveaux de confiance précédents, elles représentent 3.22% des mailles pour le seuil de 400 ppm et 55.4% pour celui de 20 ppm. Seules des données complémentaires permettraient de lever l'incertitude sur l'état de pollution de ces mailles.

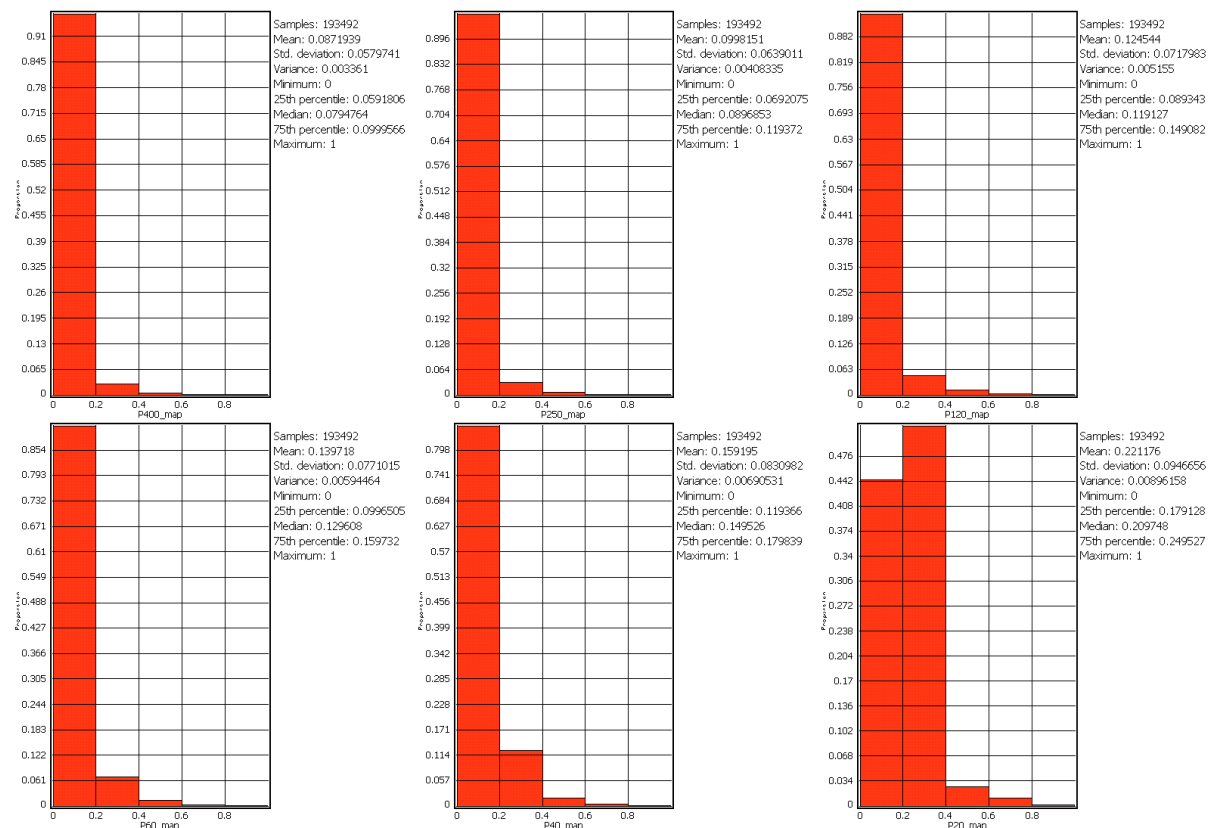


Figure 40 : Histogrammes et résumés statistiques des probabilités de présence de la zone source calculées à partir des 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas).

Comme attendu, il ressort de ces résultats que l'incertitude sur la position de la zone source dépend du seuil considéré. Pour un seuil élevé, le volume de cette zone est petit et difficilement localisable. En revanche, une très grande majorité des mailles peuvent être classées comme hors de la zone source, avec un niveau de confiance de 80%, et les mailles incertaines sont en nombre très réduit. Pour un petit seuil, il devient possible de localiser une (toute) petite partie de la zone source mais en contrepartie la proportion de mailles incertaines devient considérable. Plus généralement, comme le montre la Figure 40, la diminution du seuil se traduit par un déplacement des probabilités des valeurs faibles vers les valeurs fortes, les probabilités intermédiaires devenant progressivement de plus en plus importantes.

A partir de la carte de probabilité de présence de la source zone, il peut être visualisé la position la plus probable de cette zone. A titre d'exemple, la Figure 41 montre les positions que l'on obtient pour une probabilité de présence de la zone source supérieure à 0.8 (plus de 80% de chances) et des seuils de 20 et 40 ppm, pris suffisamment bas pour que les volumes ne soient pas trop petits (153 m³ pour le seuil de 20 ppm et 62 m³ pour celui de 40 ppm). Deux zones, situées à la perpendiculaires l'un de l'autre mais déconnectées, apparaissent sur ces figures.

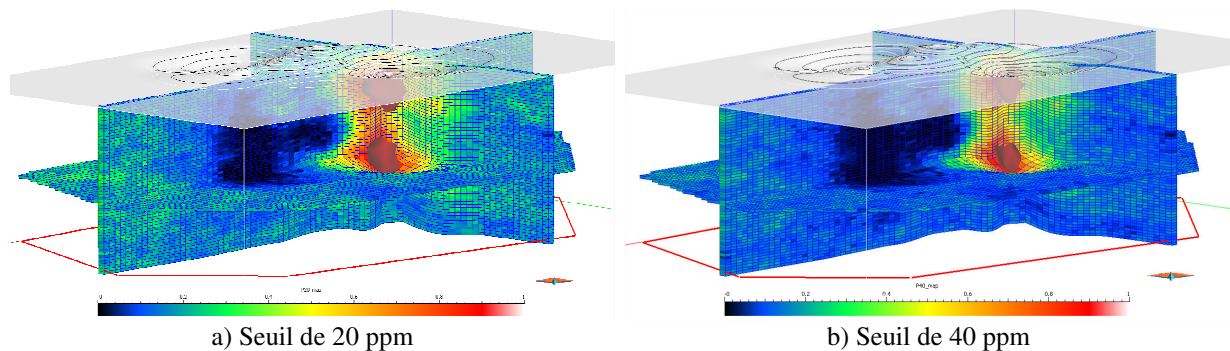


Figure 41 : Visualisations 3D de la position la plus probable de la zone source, correspondant à une probabilité de présence de la source > 0.8 (plus de 80% de chances), pour des seuils de 20 et 40 ppm. Les coupes verticales et horizontales sur la grille montrent la probabilité de présence de la zone source pour le seuil considéré. La distorsion verticale de l'image est de 7 pour 1.

6. CONFRONTATION DU MODELE GEOSTATISTIQUE AUX DONNEES DE VALIDATION

6.1 Données de validation

Sur les 319 échantillons prélevés lors du diagnostic « récent » du site, 49 ont été utilisés dans le jeu de donnée restreint et 32 sont sur des sondages situés hors du domaine simulé.

Les données de validation se composent donc de teneurs en PCE, TCE et DCE mesurées sur 238 échantillons dont 20 proviennent des mêmes sondages que les données du modèle géostatistique et les autres de 35 sondages différents. L'emplacement de ces données de validation est indiqué sur la Figure 42 ainsi que la position la plus probable de la zone source calculée pour une probabilité de présence supérieure à 0.5 (plus de 50% de chances) et des seuils de 20 et 40 ppm. On peut y voir que des données de validation, dont on sait qu'elles sont localisées dans la zone source et qui sont symbolisées par des sphères, se trouvent à l'extérieur de l'enveloppe où la zone source a plus de 50% de chances d'être rencontrée selon les prédictions du modèle géostatistique. Ces aspects sont abordés au paragraphe §6.3.3.

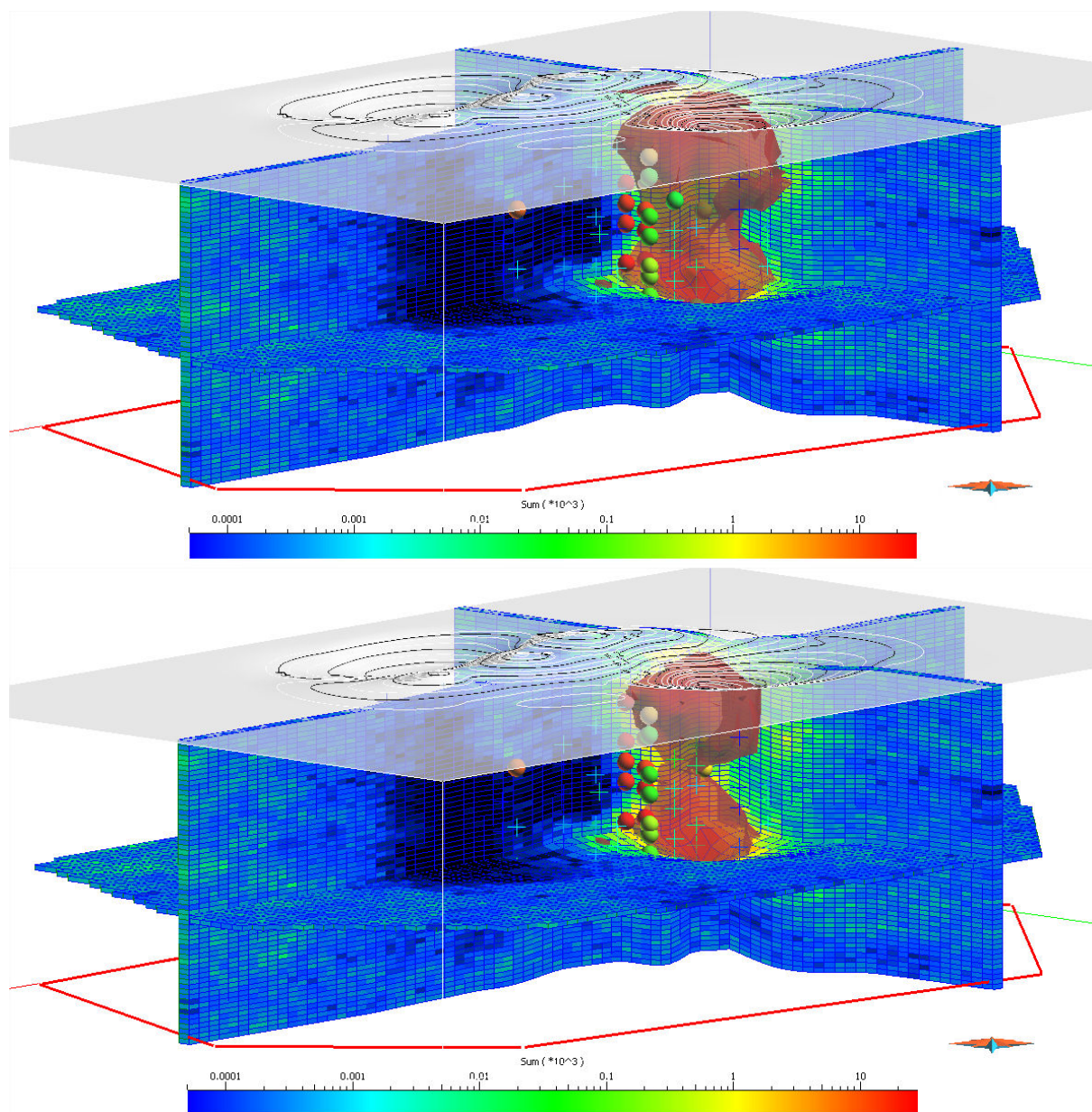


Figure 42 : Visualisations 3D des données de validation ainsi que de la position la plus probable de la zone source calculée pour une probabilité de présence supérieure à 0.5 (plus de 50% de chances) et des seuils de 20 (en haut) et 40 ppm (en bas). Les sphères correspondent aux données de validation qui sont effectivement dans la zone source et les croix à celles qui sont en dehors. La couleur des croix et des sphères correspond à la somme des teneurs en DCE, PCE et TCE au point considéré (échelle de couleur logarithmique). La distorsion verticale de l'image est de 7 pour 1.

6.2 Comparaisons statistiques entre données du modèle et données de validation

Comme pour le jeu de données du modèle géostatistique (voir §4.1.3), les données de validation ont été corrigées d'un biais d'échantillonnage possible en utilisant la même méthode de dégroupement par blocs. Elle a conduit à des poids qui varient entre 0.4 et 5 et confirment des regroupements de données. Les diagrammes de dispersion de la Figure 43 renseignent sur la répartition des poids en fonction des teneurs. Ils montrent clairement une tendance vers un échantillonnage préférentiel des sols pollués. Sauf indiqué explicitement, les statistiques présentées dans la suite portent sur les données dégroupées.

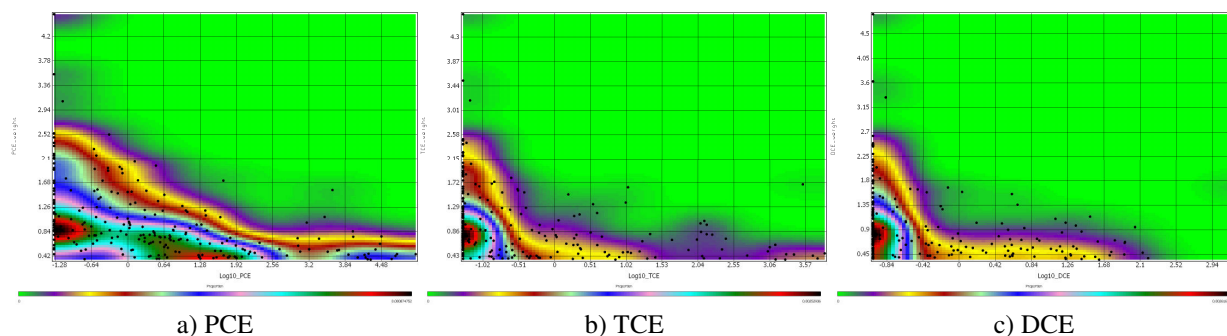


Figure 43 : Diagrammes de dispersion entre poids de dégroupage (vertical) et données de teneur (horizontal en échelle log) pour les données de validation des trois composants. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

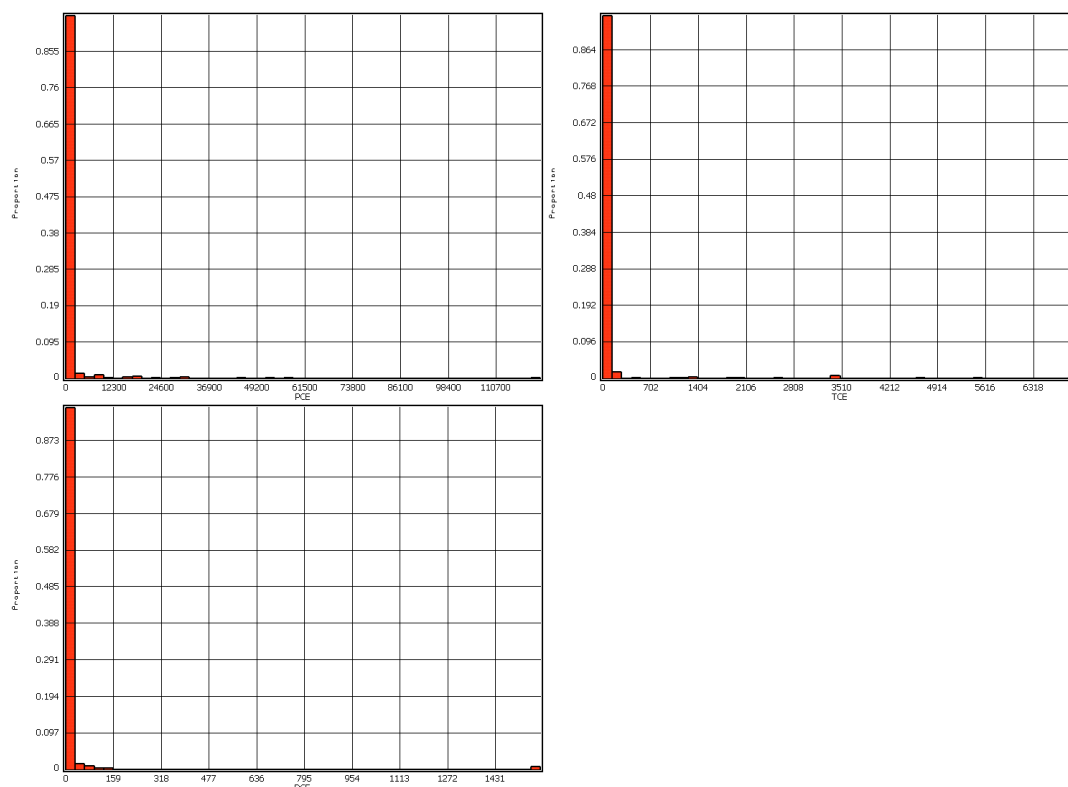


Figure 44 : Histogrammes des teneurs en PCE (en haut à gauche), TCE (en haut à droite) et DCE (en bas à gauche) du jeu de données de validation (unité : ppm).

Les histogrammes de la Figure 44 et les résumés statistiques du Tableau 13 renseignent sur les statistiques du jeu de données de validation. Ces statistiques sont à comparer à celles des données du modèle géostatistique, l'objectif étant de vérifier que les données de validation sont cohérentes avec

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

les données qui ont servi à générer le modèle géostatistique. Plus précisément, l'approche géostatistique ayant pour finalité de construire un modèle de la pollution des sols réaliste, dans le sens où il reproduit les caractéristiques statistiques et spatiales déduites des données de départ, si des écarts importants existent avec les données de validation, qui sont plus nombreuses et donc plus précises sur l'état de pollution des sols, on peut s'attendre déjà à ce que le modèle s'écarte de la réalité.

Il ressort de la comparaison des résumés statistiques que les données de validation présentent des teneurs plus élevées pour les trois composants (maximum et moyenne plus grands) ainsi que des coefficients de variation très supérieurs et qui indiquent une asymétrie des distributions beaucoup plus grande qu'initialement observée. Les médianes sont en revanche assez semblables. Il est à noter la grande proportion de données sous la limite de quantification pour les trois composants y compris le PCE qui était préservé dans le jeu de données restreint. Ces proportions varient de 15% pour le PCE à environ 50% pour le DCE. Après correction des biais d'échantillonnage, elles se trouvent encore augmentées.

Note : Ce biais d'échantillonnage est lié au fait que le choix des données initiales a été guidé par la présence des cuves et de fortes concentrations dans la nappe. L'ingénieur de BURGEAP a donc choisi des positions où il était susceptible de voir de fortes teneurs (Précision technique apportée par BURGEAP).

Tableau 13 : Résumés statistiques des teneurs en PCE, TCE et DCE du jeu de données de validation (unité : ppm).

	Mesures	Mesures > LD	Moyenne	Ecart-type	Coeff. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LQ)
PCE	238	196	1 086	7 223	6.7	0.05	122 000	0.05	0.46	4.43	<0.05
TCE	238	132	86.1	550.0	6.4	0.05	6 960	0.05	0.05	0.29	<0.05
DCE	238	112	14.9	133.0	8.9	0.1	1 580	0.1	0.1	0.39	<0.15

Pour compléter la comparaison, des diagrammes quantile-quantile (voir signification en Annexe, §13) ont été générés pour les trois composants. A titre d'illustration, l'exercice a été répété pour des données corrigées des biais d'échantillonnage (dégroupées) et non corrigées. Ces diagrammes se trouvent sur la Figure 45. Ils donnent un aperçu plus complet des différences entre les distributions des deux jeux de données. Plus les points sont proches de la bissectrice, plus les distributions sont semblables. Les points au-dessus de la bissectrice indiquent des valeurs (quantiles) supérieures pour les données de validation par rapport à celles du modèle, les points au-dessous des valeurs inférieures.

Ces diagrammes suscitent les commentaires suivants.

- Les queues de distribution (points les plus à droite et situés au-dessus de la bissectrice) montrent des écarts importants qui confirment les valeurs extrêmes plus grandes des données de validation révélées par les résumés statistiques. Il est à noter que la queue de distribution est la partie la plus difficile à estimer d'une distribution asymétrique à partir d'un nombre réduit de données. Elle peut aussi être celle qui a les conséquences les plus grandes quand il s'agit de données de pollution de sols et que la classification des sols comme pollués dépend d'un seuil correspondant à un quantile (valeur de teneur) élevé de la distribution.
- En s'intéressant aux données dégroupées, pour tous les composants, les données de validation présentent une proportion de valeurs faibles à intermédiaires supérieure à celle des données du modèle (points initialement sous la bissectrice) mais aussi une proportion de valeurs extrêmes plus grande (points au-dessus la bissectrice sur la droite).
- Les différences entre données dégroupées ou non illustrent l'impact que peut avoir la correction (pondération) sur les distributions estimées. Le biais d'échantillonnage a des caractéristiques similaires pour les deux jeux de données. La correction sur les données du modèle géostatistique tend essentiellement à diminuer le poids des teneurs moyennes à fortes et, pour le DCE, également

les valeurs extrêmes (déplacement vers la gauche des étoiles par rapport aux carrés pour un même quantile). Pour les données de validation, la correction tend à donner moins de poids aux valeurs moyennes à fortes pour tous les composants et également aux valeurs extrêmes pour le PCE et le TCE (déplacement vers le bas des étoiles par rapport aux carrés pour un même quantile). Dans les deux cas, le biais correspondrait à un échantillonnage préférentiel de zones polluées à très fortement polluées pour les données de validation.

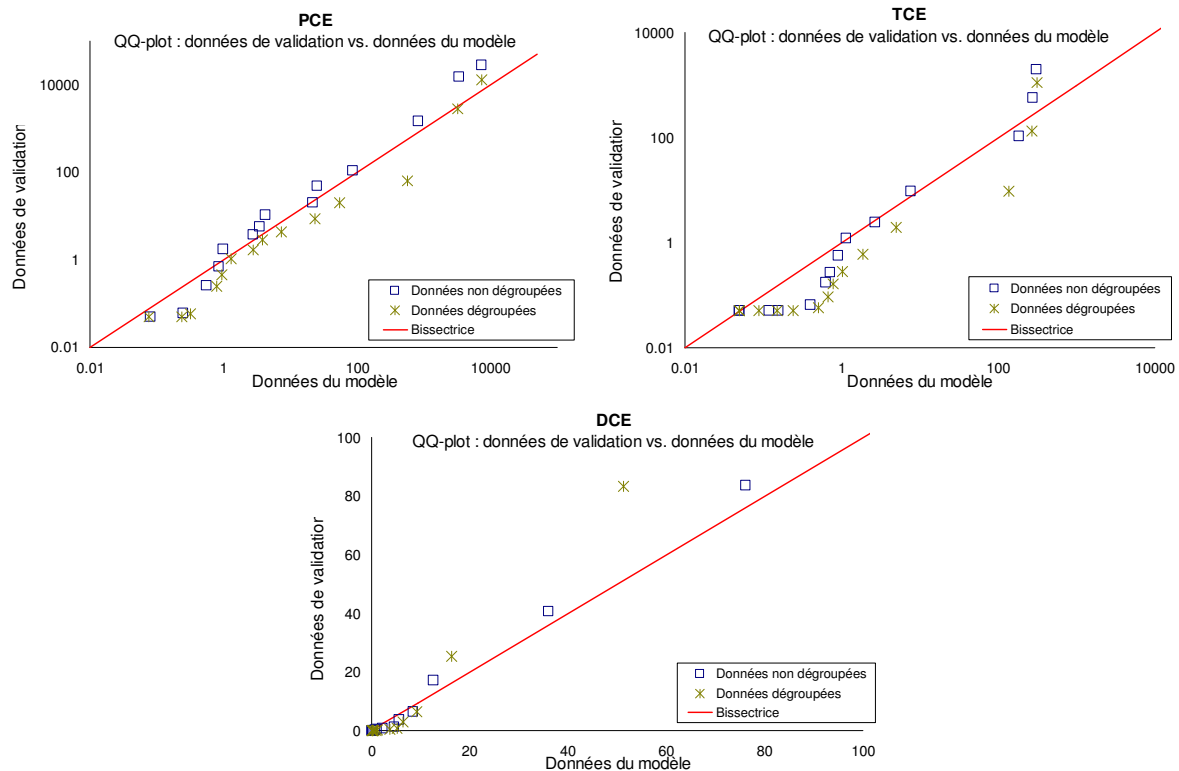


Figure 45 : Diagrammes quantile-quantile entre les distributions des jeux de données du modèle et de validation pour les trois composants PCE, TCE et DCE. Pour chaque composant, les distributions déduites de données dégroupées ou non sont présentées. Liste des quantiles visualisés : $Q_{0.1}$, $Q_{0.2}$, ..., $Q_{0.6}$, $Q_{0.65}$, $Q_{0.7}$, $Q_{0.75}$, $Q_{0.8}$, $Q_{0.85}$, $Q_{0.9}$, $Q_{0.95}$, $Q_{0.975}$.

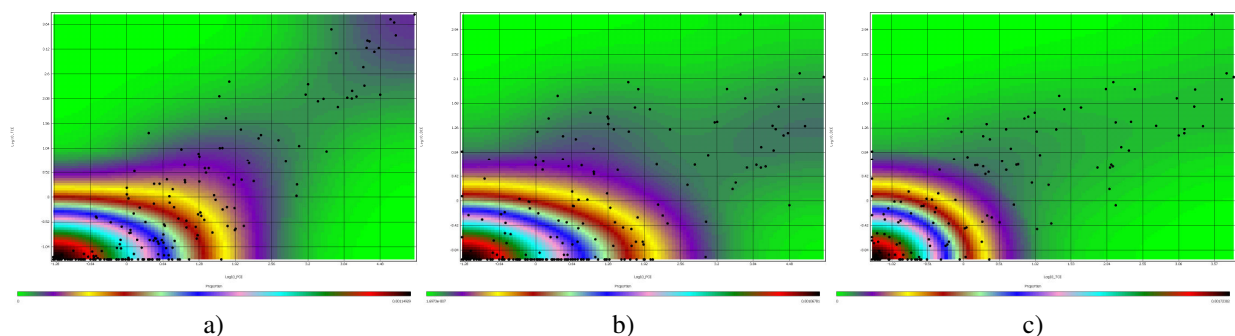


Figure 46 : Diagrammes de dispersion en échelle log entre a) TCE et PCE, b) DCE et PCE et c) DCE et TCE pour l'ensemble des données de validation. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés).

Les relations et corrélations entre composants déduites des données de validation sont illustrées à la Figure 46 et résumées dans le Tableau 14. Les coefficients de corrélation sont relativement similaires à ceux calculées sur les données du modèle géostatistique (§4.1.4) bien que légèrement plus grands entre TCE et DCE.

Tableau 14 : Coefficients de corrélation linéaire et de rang entre la transformée Log10 des teneurs de chacun des composants pour toutes les données de validation et sans les données sous les limites de quantification (LQ).

	Ensemble des données	Sans données sous LQ
PCE – TCE	Données : 238 Linéaire : 0.88 Rang : 0.67	Données : 132 Linéaire : 0.87 Rang : 0.81
PCE – DCE	Données : 238 Linéaire : 0.65 Rang : 0.35	Données : 114 Linéaire : 0.52 Rang : 0.46
TCE – DCE	Données : 238 Linéaire : 0.86 Rang : 0.71	Données : 106 Linéaire : 0.79 Rang : 0.78

6.3 Prédictions du modèle géostatistique par rapport aux données de validation

6.3.1 Modes de comparaison

La validation du modèle géostatistique, par rapport aux données de validation, a été opérée en plusieurs étapes de la façon suivante.

1. Comparaison des teneurs simulées par rapport aux teneurs mesurées.
2. Evaluation de la précision et de la justesse du modèle.
3. Prédictions du modèle par rapport à la zone source.

6.3.2 Comparaison entre teneurs simulées et teneurs mesurées

La comparaison la plus directe entre le modèle géostatistique et les données de validation consiste à vérifier comment se situent les données par rapport aux valeurs simulées. Pour mener à bien cette comparaison, les résultats suivants ont été produits.

- Diagrammes de dispersion entre valeurs simulées et données de validation (Figure 47).
- Diagrammes de dispersion entre moyennes des valeurs simulées et données de validation (Figure 48).
- Diagrammes de dispersion entre médianes des valeurs simulées et données de validation (Figure 49).

Les coefficients de corrélation correspondant aux différents diagrammes de dispersion sont regroupés dans le Tableau 15. Il peut être tiré de ces diagrammes les résultats suivants.

1. Tous les diagrammes montrent que les teneurs simulées tendent à augmenter avec les données de validation.
2. Cette tendance existe pour les trois composants et apparaît sur les moyennes et les médianes des valeurs simulées. Elle est cependant complexe et correspond à des corrélations linéaires ou de rangs très modestes. **Le niveau d'incertitude locale à l'endroit des données de validation variant d'une donnée à l'autre et pouvant être très grand, il n'est cependant pas surprenant que les corrélations soient faibles.**
3. Comme attendu de la comparaison des distributions des données du modèle et de celles de validation (§6.2), le modèle géostatistique tend à produire des teneurs faibles à moyennes supérieures aux données de validation et des teneurs fortes inférieures. Ces tendances contrastées se traduisent par une inflexion du nuage de points particulièrement visible sur les moyennes et les médianes (Figure 48 et Figure 49).

4. Les données de validation les plus extrêmes sont les moins bien reproduites. La raison est double : 1) la sous-évaluation des teneurs fortes mentionnée précédemment, 2) la méthode de simulation elle-même qui, en étant fondée sur un formalisme gaussien, a pour inconvénient de déstructurer les valeurs extrêmes.¹¹

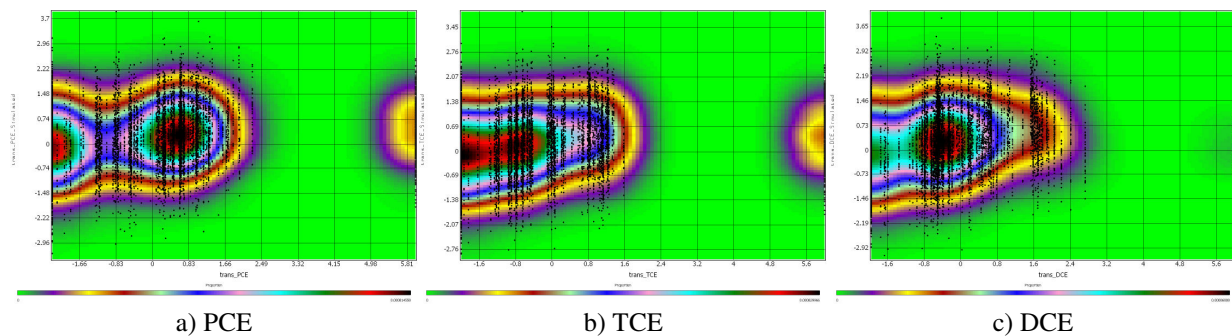


Figure 47 : Diagrammes de dispersion entre les valeurs simulées et les données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

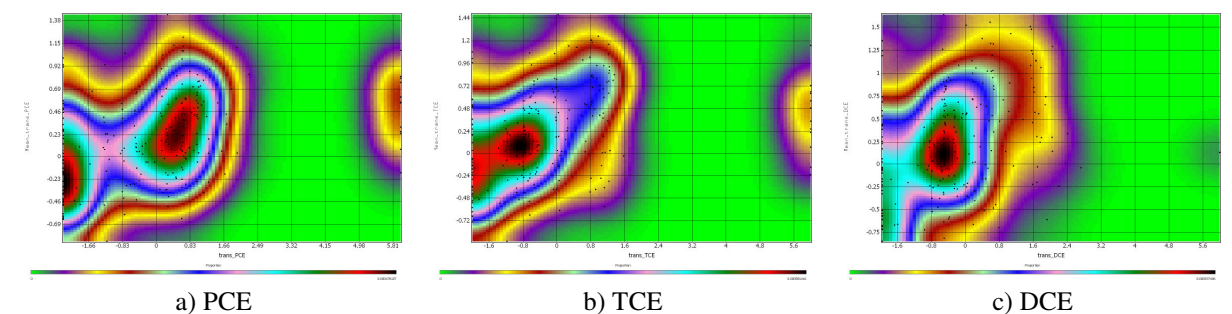


Figure 48 : Diagrammes de dispersion entre les moyennes des valeurs simulées et les données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

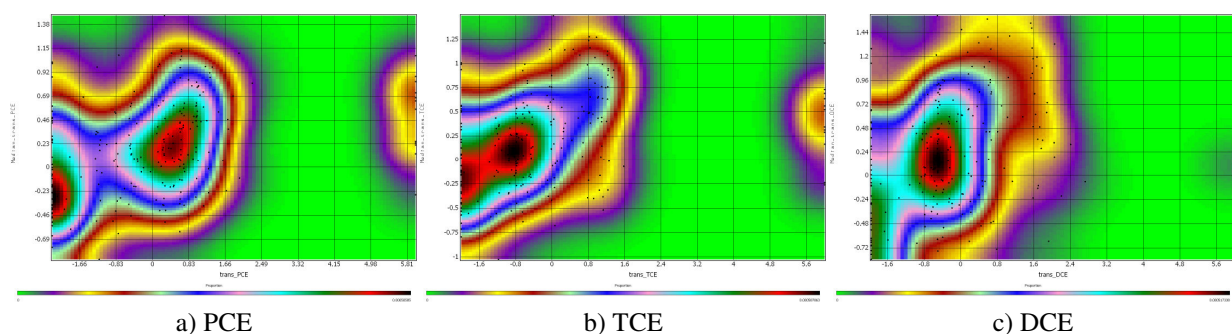


Figure 49 : Diagrammes de dispersion entre les médianes des valeurs simulées et les données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

¹¹ On parle « d'entropie maximale » des fonctions aléatoires gaussiennes avec pour conséquence une détérioration de la continuité spatiale des valeurs situées en queue de distribution.

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

Tableau 15 : Coefficients de corrélation linéaire et de rang entre l'ensemble des données de validation (variables transformées) et les valeurs simulées ou des statistiques centrales tirées de ces valeurs.

	Valeurs simulées	Moyenne des valeurs simulées	Médiane des valeurs simulées
PCE	Données : 23 800 Linéaire : 0.18 Rang : 0.19	Données : 238 Linéaire : 0.36 Rang : 0.42	Données : 238 Linéaire : 0.37 Rang : 0.42
TCE	Données : 23 800 Linéaire : 0.17 Rang : 0.22	Données : 238 Linéaire : 0.32 Rang : 0.44	Données : 238 Linéaire : 0.33 Rang : 0.44
DCE	Données : 23 800 Linéaire : 0.19 Rang : 0.20	Données : 238 Linéaire : 0.34 Rang : 0.37	Données : 238 Linéaire : 0.34 Rang : 0.37

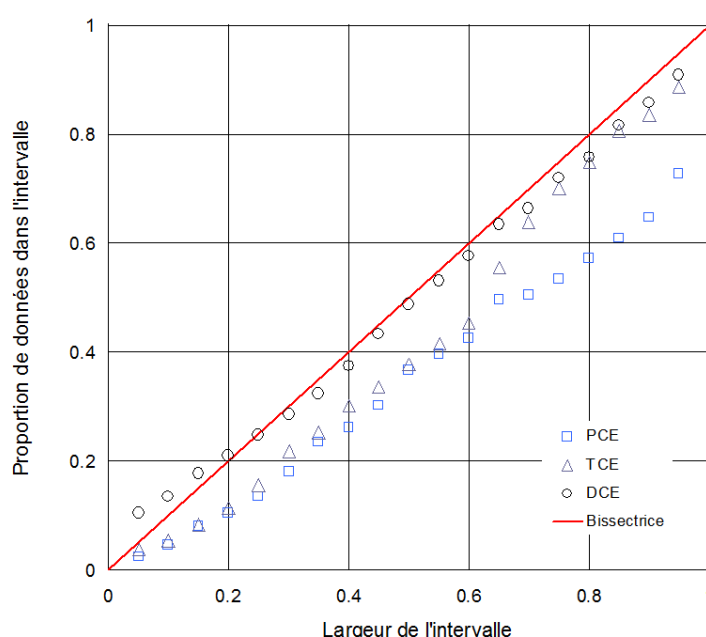


Figure 50 : Diagrammes de précision et justesse calculés sur les données de validation pour les trois composants en repositionnant à la limite de quantification les valeurs simulées au-dessous.

6.3.3 Evaluation de la précision et justesse du modèle

L'évaluation de la justesse et de la précision du modèle géostatistique consiste à tracer le diagramme de dispersion entre la largeur (probabilité p) des intervalles de confiance, calculés à partir des distributions de valeurs simulées à l'endroit des données de validation, et la proportion $\Pi(p)$ de teneurs mesurées tombant dans ces intervalles (voir Annexe, §14, pour une présentation détaillée de l'approche). Le modèle géostatistique peut être considéré comme juste lorsque la proportion $\Pi(p)$ est supérieure ou égale à p . Le modèle est d'autant plus précis que $\Pi(p)$ est proche de (et supérieur à) p . Le diagramme de dispersion entre p (abscisse) et $\Pi(p)$ (ordonnée) permet de contrôler visuellement la qualité du modèle : il est juste lorsque les points du graphique sont situés au-dessus de la première bissectrice et il est d'autant plus précis que ces points sont proches de cette bissectrice.

La Figure 50 fournit les diagrammes de dispersion et de justesse obtenus pour les trois composants. Ils montrent que les simulations des teneurs en PCE et TCE ne sont pas justes (points situés sous la bissectrice) et que les écarts avec les données de validation sont davantage marqués pour le PCE. Ces écarts sont à rapprocher aux différences observées précédemment entre valeurs simulées et données de validation (voir §6.3.2 commentaires 3 et 4). Les teneurs en DCE, dont la distribution est beaucoup moins étalée, semblent en revanche être simulées plus convenablement. Les valeurs de teneurs plus

faibles (distribution plus simple avec une asymétrie réduite) et la proportion plus grande de données sous la limite de quantification expliquent, en partie au moins, les meilleures prédictions pour DCE.

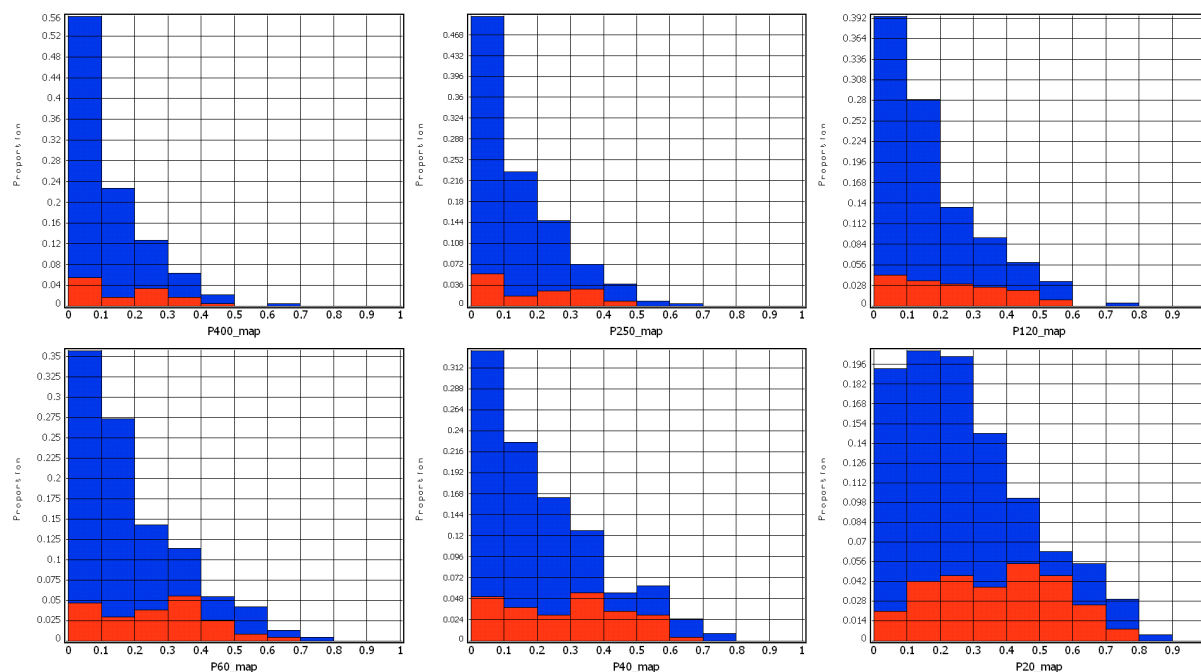


Figure 51 : Histogrammes des probabilités de dépassement des seuils, à l'endroit des données de validation pour les seuils de 400, 250, 1200, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Différentiation en rouge et en bleu des données situées respectivement dans et hors de la zone source pour le seuil de teneur considéré.

6.3.4 Prédictions du modèle par rapport à la zone source

Les données de validation peuvent être classifiées comme étant situées dans ou hors de la zone source en leur appliquant les seuils de teneur totale définis plus haut (voir §5.2.1). Cette classification des données de validation est à comparer aux prédictions du modèle géostatistique obtenues en appliquant les mêmes seuils aux simulations de teneurs (voir §5.2.2). Ces prédictions étant associées à une incertitude (erreur de classification) à prendre compte à nouveau pour juger de la justesse du modèle, les résultats suivants ont été produits pour mener les comparaisons.

1. Histogrammes des probabilités de dépassement des seuils à l'endroit des données de validation (un histogramme par seuil, voir §5.2.3), en différenciant les données dans et hors de la zone source. La Figure 51 montre les histogrammes des probabilités de dépassement pour des seuils de 400 à 20 ppm. Le modèle géostatistique peut être considéré comme convenablement prédictif si la proportion de données de validation hors de la zone source diminue, de gauche à droite, avec une probabilité croissante, tandis que la proportion de données dans la zone source augmente (barres de l'histogrammes majoritairement bleues à gauche et majoritairement rouge à droite).
2. Confrontation entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne des données classées. Pour un niveau de confiance donné, des données de validation prédites comme appartenant à la zone source peuvent se retrouver en réalité situées hors de cette zone et réciproquement. Plus le niveau de confiance est grand, moins grande doit être la proportion de ces données mal classées. Dans tous les cas, cette proportion doit se rapprocher de la probabilité d'erreur du modèle qui se déduit de la probabilité de dépassement du seuil de teneur considéré de la façon suivante :

Si $\text{Prob}(\text{somme des teneurs} > \text{seuil}) \geq 0.5$: probabilité d'erreur = $1 - \text{Prob}(\text{somme des teneurs} > \text{seuil})$
 Si $\text{Prob}(\text{somme des teneurs} > \text{seuil}) < 0.5$: probabilité d'erreur = $\text{Prob}(\text{somme des teneurs} > \text{seuil})$

Le Tableau 16 fournit la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne correspondante pour différents intervalles de confiance et les six seuils de définition de la zone source.

3. Diagrammes des erreurs relatives entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne. Ces diagrammes sont un résumé graphique du Tableau 16 précédent à partir d'erreurs relatives calculées sous la forme :

$$\text{Erreur relative} = \text{Proportion de données de validation mal classées} / \text{Probabilité d'erreur moyenne} - 1$$

L'erreur relative doit tendre vers zéro si les erreurs de classification des données de validation sont conformes aux incertitudes du modèle géostatistiques et ce quel que soit l'intervalle de confiance choisi. Elle est positive si les erreurs de classification sont supérieures à celles que le modèle géostatistique laisse prévoir et négative dans le cas contraire (prédictions du modèle supérieures aux attentes). Les diagrammes des erreurs relatives pour des niveaux de confiance de 50 à 95% sont présentés à la Figure 52.

Ces résultats peuvent s'interpréter comme suit.

- Comme attendu, les histogrammes des probabilités de dépassement des seuils de la Figure 51 montrent une proportion croissante de données situées dans la zone source (population en rouge dans les barres des histogrammes), accompagnée d'une augmentation de la probabilité maximale de dépassement du seuil, pour un seuil de teneur décroissant. Un seuil plus petit signifie en effet une zone source plus grande dans laquelle le modèle géostatistique est susceptible d'identifier avec davantage de confiance les échantillons de validation qui s'y trouvent. Comme évoqué précédemment à propos de l'incertitude spatiale (voir §5.2.3), la diminution du seuil de teneur conduit aussi à une incertitude spatiale plus grande sur la position de la zone source et qui se traduit, sur les échantillons de validation comme pour l'ensemble de la grille (Figure 40), par un déplacement des probabilités des valeurs faibles vers les valeurs fortes.
- Les barres bleues et rouges de chaque histogramme variant d'un intervalle à l'autre, il n'est pas aisé de vérifier à partir des histogrammes que les données de validation situées dans la zone source sont proportionnellement plus nombreuses lorsque la probabilité de dépassement du seuil augmente. Une telle tendance semble cependant se dessiner sur presque tous les histogrammes, soit que la hauteur des barres de l'histogramme diminue, soit que celle des barres rouges augmente.
- Dans le Tableau 16, les proportions de données de validation mal classées, comparées aux probabilités d'erreur de classification moyennes, permettent de vérifier plus en détail si les erreurs de classification des données de validation sont conformes aux incertitudes du modèle géostatistique. Pour un niveau de confiance de 50%, qui conduit à classer systématiquement toutes les données de validation comme étant soit dans la zone source (probabilité ≥ 0.5), soit à l'extérieur (probabilité < 0.5), les pourcentages d'erreurs (deuxième colonne) sont proches de ceux attendus (troisième colonne) et ceux d'autant mieux que le seuil de teneur est élevé (≥ 120 ppm). Tel n'est pas le cas des niveaux de confiance plus grands qui révèlent des écarts croissants entre la proportion d'erreurs commises et celle attendue au fur et à mesure que le niveau de confiance augmente.
- Les tendances tirées du Tableau 16 se retrouvent graphiquement sur les courbes de la Figure 52. Quel que soit le seuil de teneur considéré, l'erreur relative tend à croître avec le niveau de confiance. Ce phénomène se traduit par un déplacement des courbes vers le haut en s'éloignant de zéro.

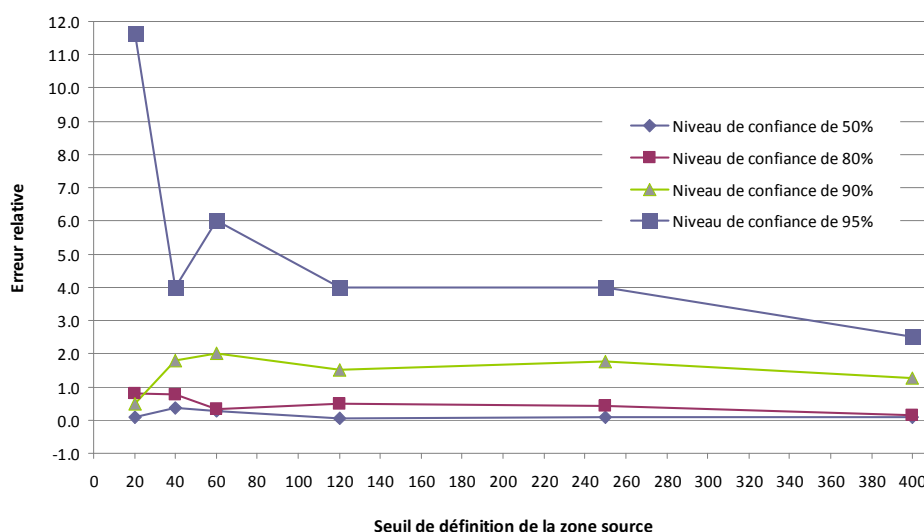


Figure 52 : Diagrammes des erreurs relatives entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne des données classées selon la valeur du seuil critique (en ppm).

Tableau 16 : Correspondance entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne des données classées. Calculs répétés pour différents intervalles de confiance et les six seuils de définition de la zone source.

Seuil	Nombre de données classées	Proportion de données mal classées (%)	Probabilité d'erreur moyenne (%)
Niveau de confiance de 50%			
400 ppm	238	13%	12%
250 ppm	238	15%	14%
120 ppm	238	18%	17%
60 ppm	238	23%	18%
40 ppm	238	27%	20%
20 ppm	238	27%	25%
Niveau de confiance de 80%			
400 ppm	183	8%	7%
250 ppm	171	10%	7%
120 ppm	156	12%	8%
60 ppm	147	12%	9%
40 ppm	132	16%	9%
20 ppm	88	18%	10%
Niveau de confiance de 90%			
400 ppm	117	9%	4%
250 ppm	107	11%	4%
120 ppm	89	10%	4%
60 ppm	81	12%	4%
40 ppm	72	14%	5%
20 ppm	43	9%	6%
Niveau de confiance de 95%			
400 ppm	75	7%	2%
250 ppm	68	10%	2%
120 ppm	52	10%	2%
60 ppm	42	14%	2%
40 ppm	34	15%	3%
20 ppm	8	38%	3%

6.4 Conclusions

Ce premier modèle géostatistique a été établi à partir du jeu de données restreint constitué de 49 échantillons de sols. Les prédictions de ce modèle ont été confrontées aux 238 données de validation en s'intéressant à la fois aux teneurs et à la classification des sols comme étant situés dans ou hors de la zone source. Six seuils de teneur ont été utilisés pour définir la zone source, entre 400 et 20 ppm.

Le modèle géostatistique montre des tendances qui sont conformes à celles observées sur les données de validation avec des teneurs simulées qui évoluent généralement dans le même sens que celles mesurées. L'amplitude des variations est cependant mal reproduite. Le modèle tend en effet à produire des teneurs faibles à moyennes supérieures aux données de validation et des teneurs fortes inférieures comme le montrent les diagrammes de dispersion entre valeurs simulées et données de validation.

Les diagrammes de justesse et précision constituent le mode d'évaluation qui très certainement illustre le mieux les décalages entre le modèle géostatistique et les données de validation. Ils attestent que les teneurs en PCE et TCE simulées sont fortement décalées par rapport à celles observées sur les données de validation. Le PCE mais aussi le TCE étant les deux composants qui dominent, la mauvaise prédiction de ces teneurs se traduit nécessairement par des erreurs de classification des échantillons de validation comme situés dans ou hors de la zone source.

Les simulations erronées s'expliquent en partie au moins par l'aperçu partiel et incomplet que le jeu de données restreint fournit sur les teneurs en composants présentes dans les sols. Le modèle géostatistique vise à reproduire des distributions de teneurs déduites de ces données et qui ne correspondent pas à la réalité. Sauf à étudier davantage l'incertitude sur les distributions de teneur et à prendre en compte cette incertitude dite de modèle dans les simulations, le modèle géostatistique ne peut réussir à reproduire convenablement un phénomène spatial dont il ne connaît pas complètement ou justement les caractéristiques. Des solutions existent pour mieux définir l'incertitude spatiale à partir de jeux de données réduits.¹² Elles font appel cependant à des notions avancées qui sortent du cadre de ces études de démonstration et nécessitent un travail d'analyse et de modélisation plus approfondi et donc demandeur de temps.

A ce stade de la démonstration, se pose la question de savoir quelle est la part d'erreur des prédictions du modèle géostatistique due aux caractéristiques statistiques et spatiales incorrectes tirées du jeu de données restreint. Pour y répondre, il a été choisi de construire un second modèle géostatistique (amélioré) à partir de modèles de distribution et de variogramme déduits du jeu de données complet, tout en conservant les 49 échantillons du jeu de données restreint pour simuler les teneurs. La présentation de ce modèle amélioré et des résultats obtenus fait l'objet du paragraphe suivant.

¹² Des techniques de ré-échantillonnage des données (*bootstrap* en anglais) permettent par exemple de quantifier l'incertitude relative aux caractéristiques statistiques d'un jeu de données.

7. AMELIORATION DU MODELE GEOSTATISTIQUE AVEC DES STATISTIQUES COMPLETES

7.1 Améliorations apportées au modèle

La comparaison des données de validation avec celles du modèle géostatistique a montré que les deux jeux de données ne correspondaient pas à des distributions de teneurs identiques (voir §6.2). Ces décalages se retrouvent logiquement sur les simulations géostatistiques qui produisent des teneurs faibles à moyennes supérieures aux données de validation et des teneurs fortes inférieures (§6.3.2).

Afin d'évaluer l'impact des modèles de distribution et de variogramme, tirés du jeu de données restreint, sur la prédiction des données de validation, il a été décidé de relancer la simulation conjointe des teneurs en introduisant des modèles statistiques établis à partir de l'ensemble des données disponibles (données du modèle géostatistique initial et données de validation). Pour le conditionnement des simulations, seul le jeu de données restreint a bien sûr été utilisé (49 échantillons), les teneurs correspondant aux données de validation devant continuer à être simulées pour en vérifier les prédictions.

Les nouveaux modèles statistiques ainsi obtenus présentent les caractéristiques suivantes.

1. Les histogrammes des teneurs et les résumés statistiques correspondants (Figure 53 et Tableau 17) révèlent de très grandes proportions de valeurs sous les limites de quantification pour les trois composants. Ces proportions varient de 15 à 50% des données selon le composant.
2. Les biais d'échantillonnage observés séparément sur les jeux de données restreint et de validation continuent d'exister après regroupement des données comme l'attestent les diagrammes de dispersion de la Figure 54. Les poids varient entre 0.4 et 4.5 pour tous les composants. Ils confirment l'échantillonnage préférentiel des sols les plus pollués mis en évidence précédemment (§4.1.3 et 6.2). Sauf indication contraire, les statistiques relatives au jeu de données complet portent sur des données dégroupées.
3. Comme conséquence de la grande proportion de données sous les limites de quantification pour les trois composants, les teneurs en PCE n'ont pas pu être utilisées à nouveau pour extrapoler les teneurs en TCE non définies puis celles en DCE (§4.1.6). Le problème est d'autant plus complexe que les données de PCE ou TCE sous la limite de quantification ne correspondent pas nécessairement elles-mêmes à des teneurs en TCE ou DCE sous une limite de quantification (Figure 55). En conséquence, la transformation gaussienne des teneurs a été opérée différemment selon une approche qui présente l'avantage d'être simple, systématique et théoriquement rigoureuse. Cette approche a été répétée pour chaque composant. Elle comprend les étapes suivantes, l'objectif étant d'obtenir pour les trois composants des données transformées qui peuvent être simulées avec des fonctions aléatoires gaussiennes centrées et réduites (moyennes de 0 et écarts-types de 1).
 - Transformation gaussienne des données de teneur du composant considéré. Cette transformation produit une variable transformée avec un pic de valeurs à la limite de quantification (Figure 56, colonne de gauche). Ces données transformées présentent cependant toutes une moyenne différente de zéro et des écarts-types (*standard-deviation* en anglais) inférieurs 1.
 - Correction de la valeur transformée correspondant à la limite de quantification pour ramener la moyenne des valeurs transformée à zéro (Figure 56, colonne de droite). Cette correction ramène toutes les moyennes à zéro et permet de mieux caler le modèle gaussien (courbes en bleu sur les histogrammes). Les écarts-types se retrouvent augmentés mais toujours inférieurs à 1.
 - La correction précédente conduit à des écarts-types augmentés mais toujours inférieurs à 1. Pour les ramener à 1, des facteurs correctifs peuvent être calculés pour tous les composants

comme l'inverse de ces écarts-types. Ces facteurs ont servi dans un premier temps à réajuster les modèles de variogramme tirés des données transformées (voir ci-après item 6).

4. Les corrélations entre teneurs en composants (Figure 57 et Tableau 18) restent conformes à celles calculées à partir du jeu de données restreint (0.9 entre PCE et TCE, 0.75 entre TCE et DCE). Ces coefficients de corrélation ont dû être diminués par la suite, à 0.7 entre PCE et TCE et à 0.5 entre TCE et DCE, pour amener les simulations à reproduire convenablement les corrélations entre composants (effet d'amélioration des corrélations par les simulations géostatistiques).
5. Les modèles de variogramme calés sur les variogrammes expérimentaux des données transformées sont présentés à la Figure 58. Ils sont de la forme suivante où h désigne la distance entre points et C_1 et C_2 la contribution des deux structures sphériques :

$$\text{Plan horizontal : } \gamma(h) = C_1 \text{ Sph}_{25}(h) + C_2 \text{ Sph}_{25}(h)$$

$$\text{Direction verticale : } \gamma(h) = C_1 \text{ Sph}_1(h) + C_2 \text{ Sph}_{10000}(h)$$

Par rapport au jeu de données restreint (voir §4.1.8), l'effet de pépité a été supprimé, ce type de modèle correspondant mieux à une pollution des sols par des composants organiques qui ont migré depuis une zone source en laissant derrière eux des teneurs qui varient de façon continue. Les portées de corrélation restent les mêmes et les contributions des deux structures sphériques ont été réajustées.

6. Les modèles de variogrammes finalement retenus pour simuler conjointement les teneurs des trois composants sont décrits dans le **Tableau 19**. Comparés à ceux calés sur les variogrammes expérimentaux (Figure 58), des corrections ont été apportées aux contributions (seuils) des structures pour réussir à simuler, à partir des données transformées, des fonctions aléatoires qui tendent à suivre une loi gaussienne centrée et réduite. Les corrélations entre composants, utilisés pour déduire les variogrammes croisés, ont dû être ajustées également pour reproduire convenablement les corrélations observées entre composants. Ces corrections ont été opérées en deux étapes. Les facteurs correctifs déduits des écarts-types calculés sur les données transformées (voir précédemment item 3) ont d'abord été appliqués aux contributions des structures pour simuler une première fois les teneurs. Ces facteurs ont ensuite été ajustés ainsi que les corrélations entre composants jusqu'à obtenir des simulations des variables transformées qui tendent à reproduire des écarts-types de 1 et les corrélations entre composants souhaitées. Des modèles satisfaisants ont été obtenus avec les modifications suivantes : facteurs correctifs de 1.17 pour PCE, 0.97 pour TCE et 1.0 pour DCE, coefficients de corrélation linéaire de 0.7 entre TCE et PCE et de 0.5 entre DCE et TCE.
7. Les corrections précédentes ont eu pour effet de décentrer les distributions simulées (moyennes différentes de 0). Ces décalages ont été corrigés simplement en spécifiant en entrée des simulations des moyennes globales qui compensent les décalages (soit, pour chaque composant, une moyenne globale égale à l'opposé du décalage observé).

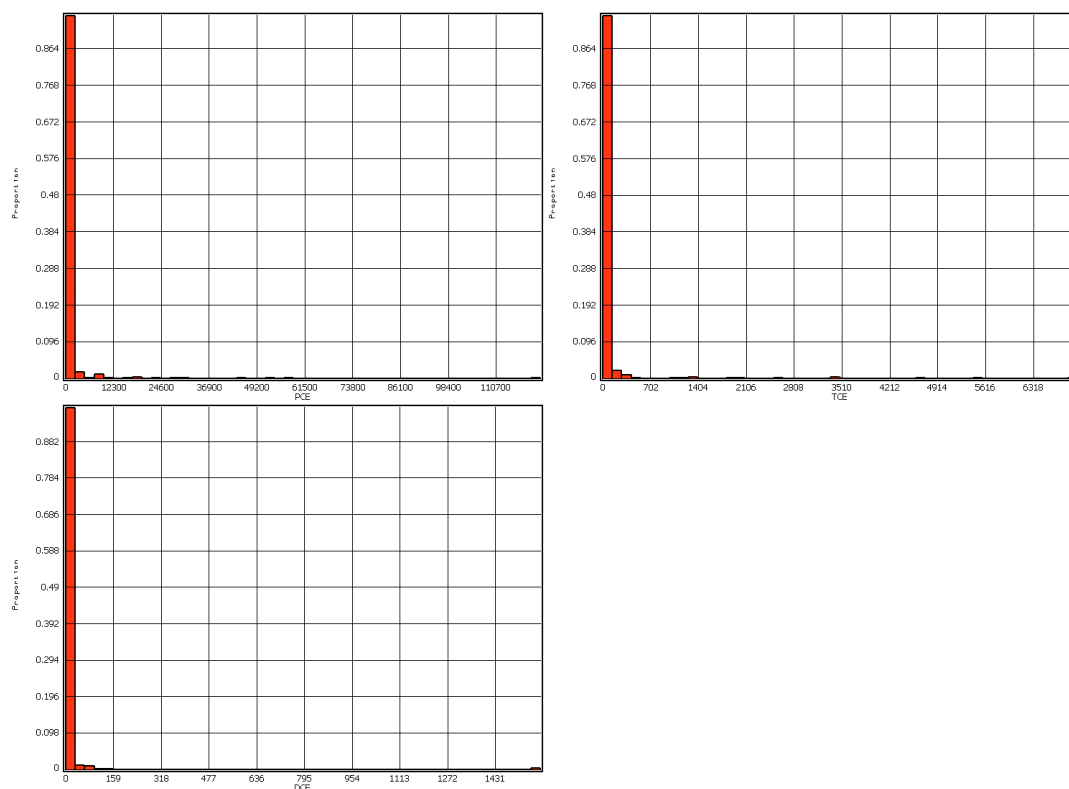


Figure 53 : Histogrammes des teneurs en PCE (en haut à gauche), TCE (en haut à droite) et DCE (en bas à gauche) pour l'ensemble des données du modèle géostatistique et de validation (unité : ppm).

Tableau 17 : Résumés statistiques des teneurs en PCE, TCE et DCE pour l'ensemble des données du modèle géostatistique et de validation (unité : ppm).

	Mesures	Mesures > LD	Moyenne	Ecart-type	Coeff. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LQ)
DCE	287	148	7.88	88.23	11.20	0.1	1580	0.1	0.1	0.32	<0.15
PCE	287	242	853.69	6232.52	7.30	0.05	122000	0.05	0.45	3.21	<0.05
TCE	287	170	61.30	453.48	7.40	0.05	6960	0.05	0.05	0.24	<0.05

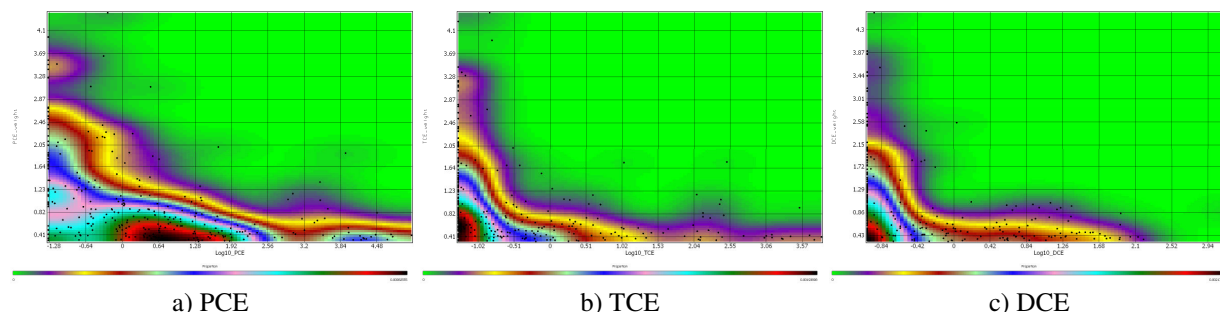


Figure 54 : Diagrammes de dispersion entre poids de dégroupage (vertical) et données de teneur (horizontal en échelle log) pour l'ensemble des données du modèle géostatistique et de validation. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures).

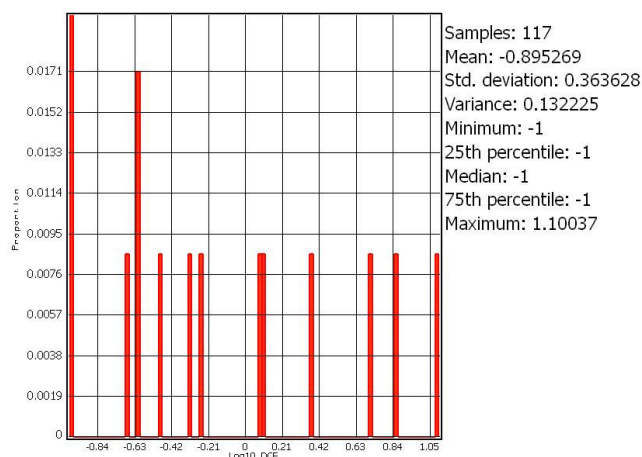


Figure 55 : Histogrammes en échelle log et résumés statistiques des teneurs en DCE correspondant à des échantillons où la teneur en TCE est sous la limite de quantification.

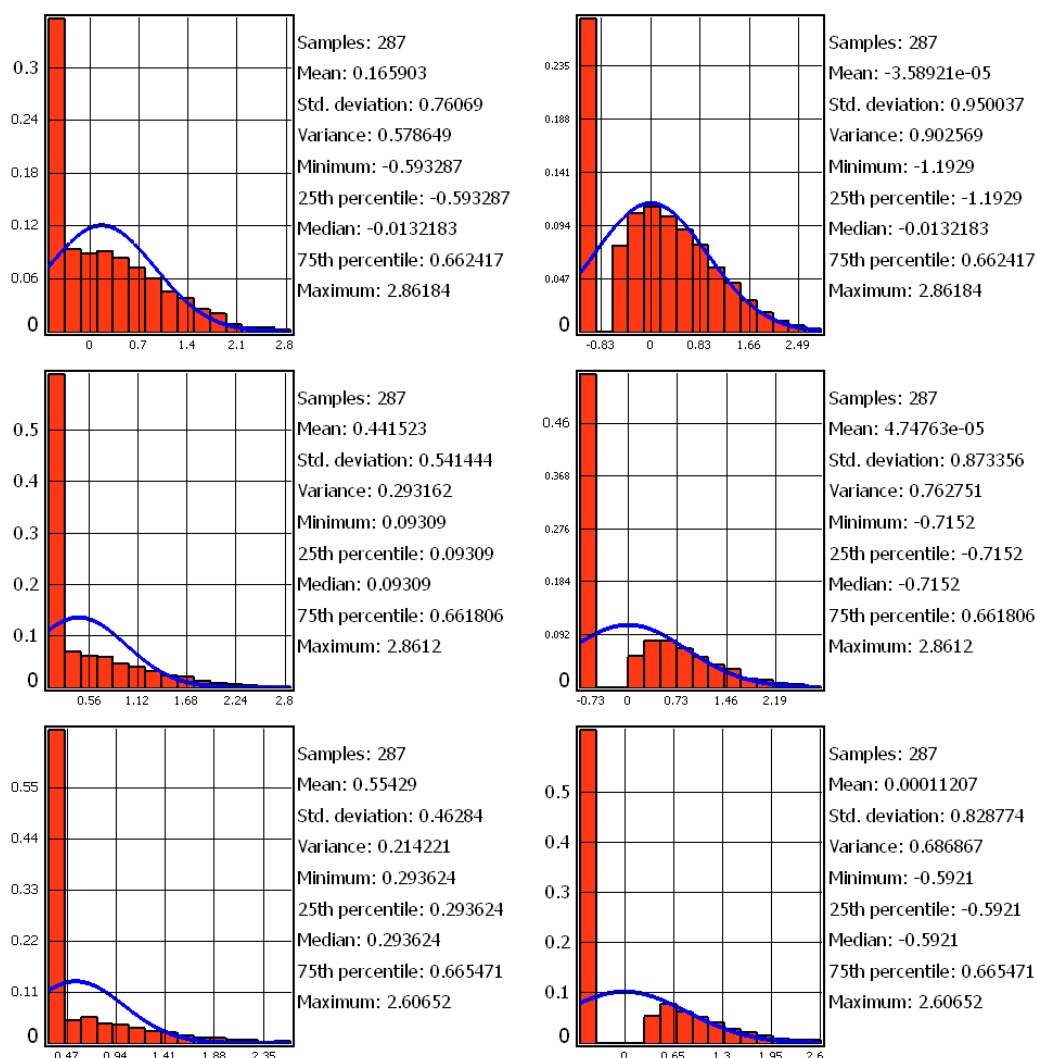


Figure 56 : Histogrammes et statistiques des teneurs en PCE (en haut), TCE (au centre) et DCE (en bas) après transformation gaussienne de l'ensemble des données du modèle géostatistique et de validation (colonne de gauche) puis après correction de la valeur transformée correspondant à la limite de quantification pour centrer les distributions sur 0 (colonne de droite). La courbe bleue indique le modèle gaussien calé sur chaque distribution.

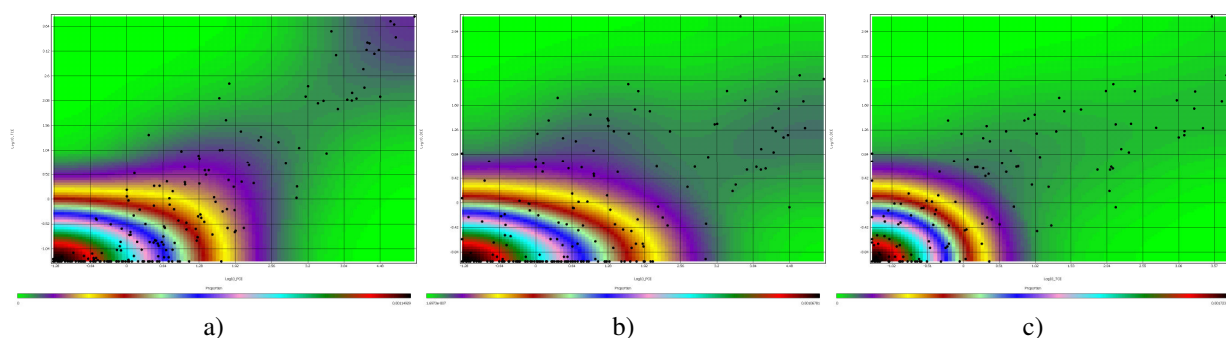


Figure 57 : Diagrammes de dispersion en échelle log entre a) TCE et PCE, b) DCE et PCE et c) DCE et TCE pour l'ensemble des données du modèle géostatistique et de validation. Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés avec échelle de couleur).

Tableau 18 : Coefficients de corrélation linéaire et de rang entre la transformée Log10 des teneurs de chacun des polluants pour l'ensemble des données (jeux de données complet) et sans les données sous les limites de quantification (LQ).

	Ensemble des données	Sans données sous LQ
PCE – TCE	Données : 287 Linéaire : 0.88 Rang : 0.83	Données : 169 Linéaire : 0.88 Rang : 0.80
PCE – DCE	Données : 287 Linéaire : 0.62 Rang : 0.59	Données : 144 Linéaire : 0.48 Rang : 0.44
TCE – DCE	Données : 287 Linéaire : 0.82 Rang : 0.80	Données : 136 Linéaire : 0.74 Rang : 0.73

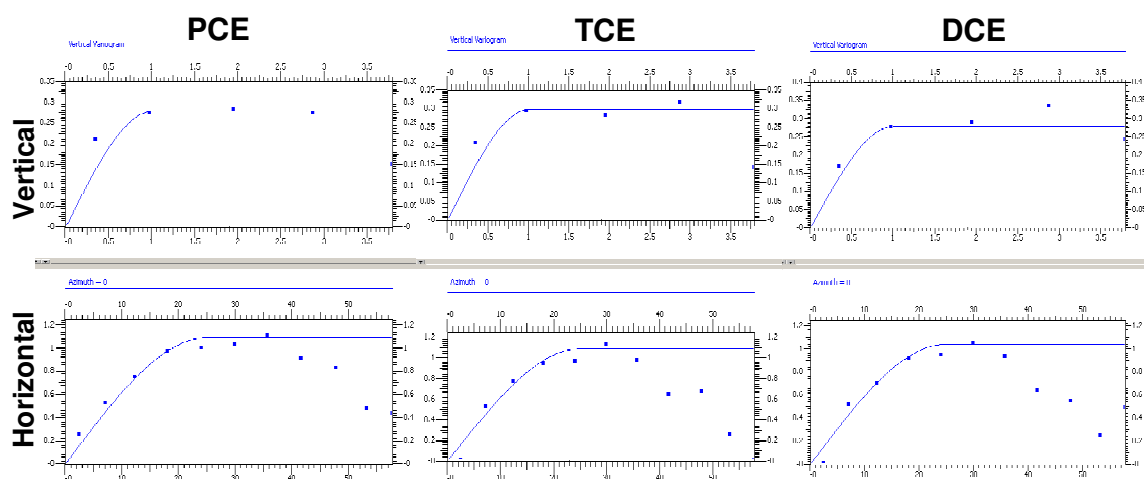


Figure 58 : Variogrammes expérimentaux (points) et modèles calés (lignes) pour le PCE, TCE et DCE (données transformées, profondeur en coordonnée verticale) établis à partir de l'ensemble des données du modèle géostatistique et de validation.

Tableau 19 : Modèles de variogrammes directs et croisés calés sur les variogrammes expérimentaux, à partir de l'ensemble des données du modèle géostatistique et de validation, et corrigés pour que les simulations reproduisent convenablement les distributions de valeurs et les corrélations entre composants souhaitées (voir item 6 p 89).

Composant	Structure 1	Structure 2
	Modèle : sphérique Portée horizontale : 25 m Portée verticale : 1 m Seuil (C_1)	Modèle : sphérique Portée horizontale : 25 m Portée verticale : 10 000 m Seuil (C_2)
Variogrammes directs		
PCE	0.33	0.96
TCE	0.29	0.78
DCE	0.28	0.76
Variogrammes croisés		
TCE – PCE	0.20	0.55
DCE – TCE	0.14	0.38

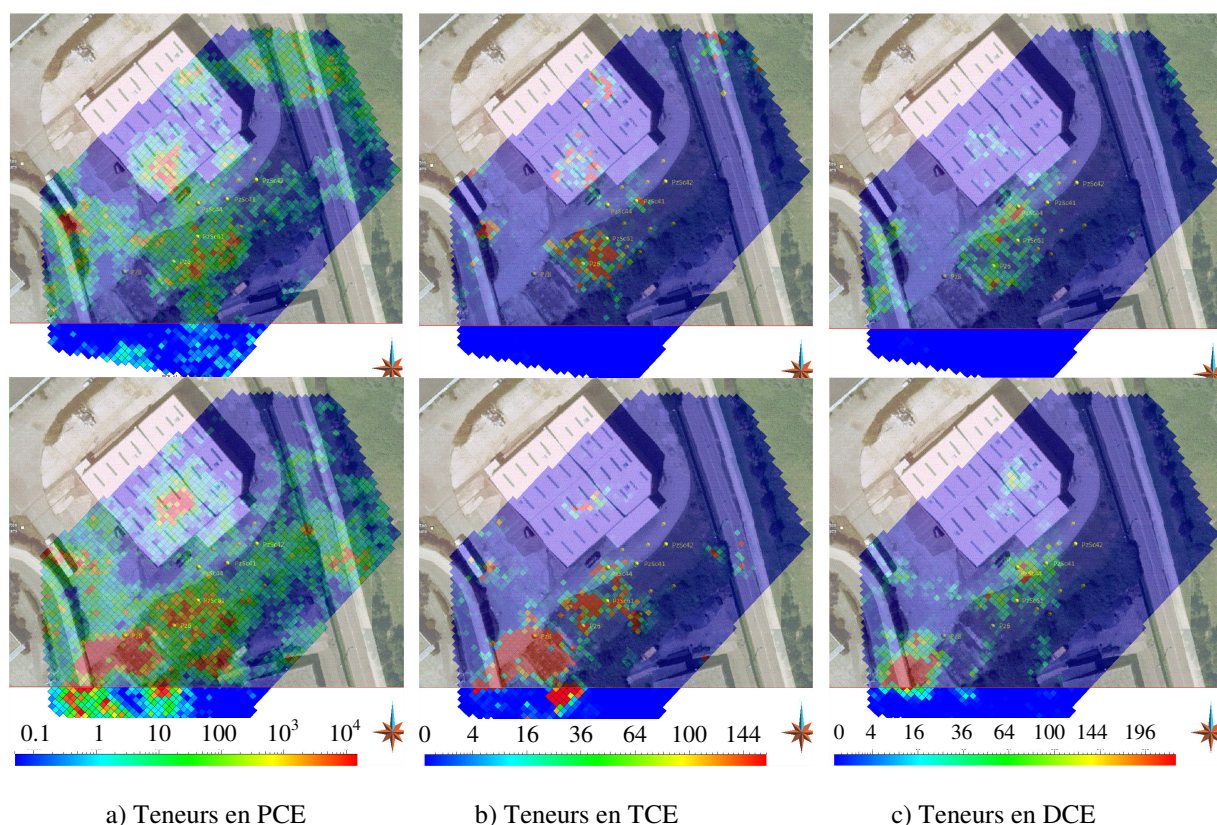


Figure 59 : Coupes horizontales de deux simulations conjointes des teneurs pour les trois polluants (unité : ppm). Les sondages affichés sont ceux d'où proviennent les données de teneur utilisées pour conditionner les simulations. Modèle géostatistique amélioré.

7.2 Résultats du modèle géostatistique amélioré

7.2.1 Simulation des teneurs

Comme pour le premier modèle géostatistique (§5.1.1), cent simulations conjointes des teneurs en PCE, TCE et DCE ont été générées. La Figure 59 présente deux simulations particulières correspondant au même plan de coupe horizontal que celui de la Figure 35 (§5.1.1). Visuellement, les nouvelles simulations ressemblent à celles du premier modèle géostatistique tout en étant différentes. Il en est de même de la variabilité des teneurs d'une réalisation à l'autre. Un tel résultat n'est pas surprenant car les deux modèles géostatistiques reposent sur le même jeu de données restreint pour conditionner les simulations (données de 49 échantillons).

7.2.2 Incertitude sur le volume de la zone source

Les histogrammes et les résumés statistiques des volumes de la zone source calculés à partir des réalisations, pour les différents seuils de définition de 400 à 20 ppm, sont regroupés à la Figure 60. Comparés à ceux du premier modèle géostatistique (§5.2.2), ces volumes apparaissent relativement similaires, bien que légèrement plus faibles pour le seuil de 20 ppm. L'incertitude sur ces volumes est en revanche un peu plus importante avec le modèle géostatistique amélioré comme l'attestent par exemple les écarts-types plus grands.

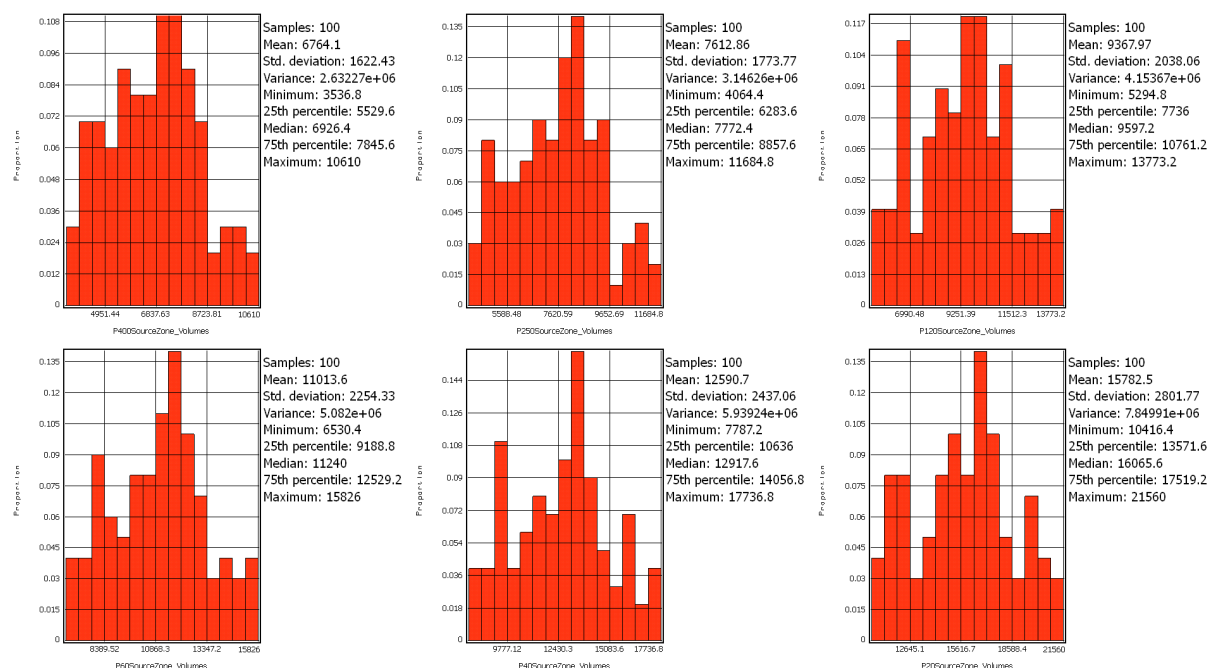


Figure 60 : Histogramme des volumes en m³ de la zone source calculés sur les 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique amélioré.

7.2.3 Incertitude spatiale sur la position de la zone source

Comme pour le modèle géostatistique de base, l'incertitude sur la position de la zone source peut être étudiée à partir de cartes de probabilité pour la somme des teneurs en PCE, TCE et DCE de dépasser le seuil de définition considéré (voir §5.2.3).

La Figure 61 et la Figure 62 montrent des coupes horizontales et verticales de la carte de probabilité de présence de la zone source calculée à partir des 100 simulations pour un seuil de teneur totale de 20 ppm. Ces cartes sont équivalentes à celles de la Figure 38 et de la Figure 39 du premier modèle géostatistique. Elles montrent une zone de grandes probabilités plus étendue (zone en orange et rouge) et qui signifie qu'un plus grand nombre de mailles peut être identifié avec confiance comme appartenant à la zone source.

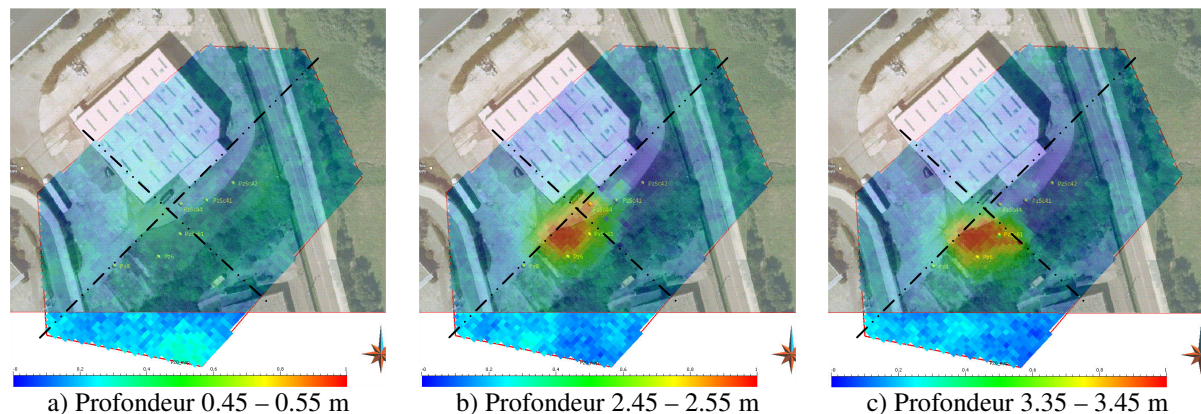


Figure 61 : Coupes horizontales de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE. Modèle géostatistique amélioré.

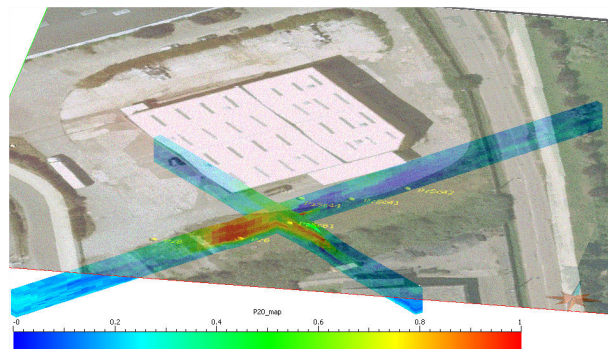


Figure 62 : Coupes verticales vues en perspective de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE. Modèle géostatistique amélioré.

Les histogrammes et les résumés statistiques des probabilités de présence de la zone source pour les différents seuils de définition se trouvent à la Figure 63. Ces résultats statistiques sont à comparer à ceux de la Figure 40. Ils confirment la proportion croissante de mailles qui peuvent être classées avec confiance comme situées dans ou hors de la zone source, proportion qui est d'autant plus grande que le seuil de définition de la zone source est petit. En appliquant les mêmes critères de classification que ceux utilisés au paragraphe §5.2.3, ces tendances se traduisent de la façon suivante.

- Les mailles que l'on peut classer comme étant hors de la zone source avec un niveau de confiance de 80% représentent 95.65% du volume pour le seuil de 400 ppm et 68.5% pour le seuil de 20 ppm. Ces volumes sont à comparer à ceux du modèle géostatistique initial : 96.77% du volume classé avec confiance hors de la zone source pour le seuil de 400 ppm et 44.4% pour le seuil de 20 ppm. Pour un même niveau de confiance de 80%, le modèle géostatistique amélioré permet ainsi d'identifier un volume de sols hors de la zone source qui est équivalent pour le seuil de 400 ppm et beaucoup plus grand pour le seuil de 20 ppm. La plus grande capacité du modèle géostatistique amélioré à prédire avec confiance les sols situés hors de la zone source à 20 ppm s'explique par les distributions de teneurs qui sont différentes de celles du modèle initial (voir §6.2, Figure 45). La proportion de teneurs faibles à moyenne est en effet supérieure pour le modèle amélioré établi à partir de l'ensemble des données, ce qui conduit à une probabilité plus faible de voir la somme des teneurs en PCE, TCE et DCE supérieures à 20 ppm.
- Les mailles que l'on peut classer comme situées dans la zone source avec le même niveau de confiance de 80% forment 0.04% du volume pour le seuil de 400 ppm et 1.3% pour le seuil de 20 ppm. Le volume de la zone source étant très petit, ces proportions restent très faibles mais en augmentation importante par rapport au premier modèle géostatistique (0.01% du volume pour le seuil de 400 ppm et 0.2% pour le seuil de 20 ppm).

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

- Les mailles incertaines, qui n'entrent pas dans une des deux classes précédentes, représentent 4.31% des mailles pour le seuil de 400 ppm et 30.2% pour celui de 20 ppm. Pour le seuil de 20 ppm, l'incertitude spatiale se retrouve ainsi logiquement beaucoup plus réduite (55.44% du volume classé comme incertain dans le modèle initial).

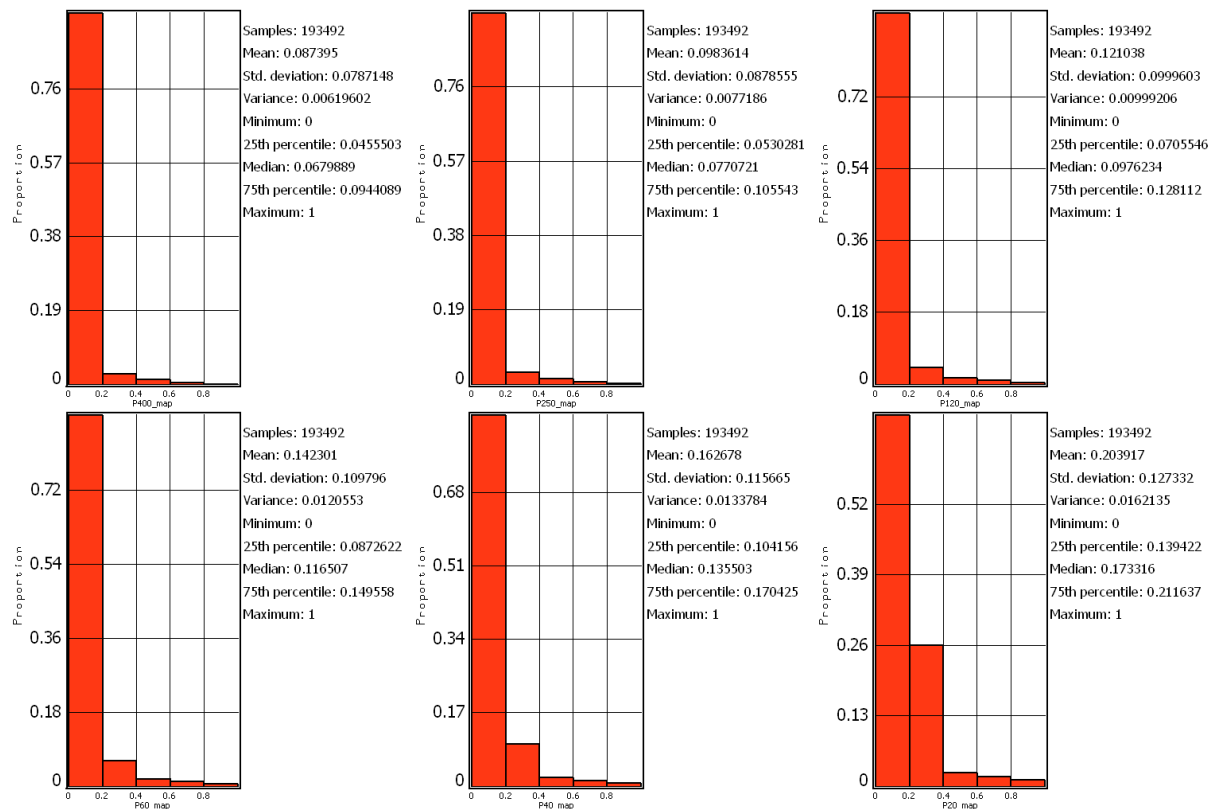


Figure 63 : Histogrammes et résumés statistiques des probabilités de présence de la zone source calculées à partir des 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique amélioré.

La visualisation 3D de la position la plus probable de la zone source, pour un seuil de 20 ppm et une probabilité de présence de la zone source supérieure à 0.8 (plus de 80% de chances), est illustrée à la Figure 64. Sur les 16 000 m³ que la zone source peut représenter en moyenne, 993 m³ peuvent être positionnés avec un niveau de confiance de 80% de chances, contre 153 m³ pour le premier modèle géostatistique. Il ressort également de cette figure que les deux positions qui apparaissaient déconnectées précédemment (Figure 41.a) font bien partie de la même zone source.

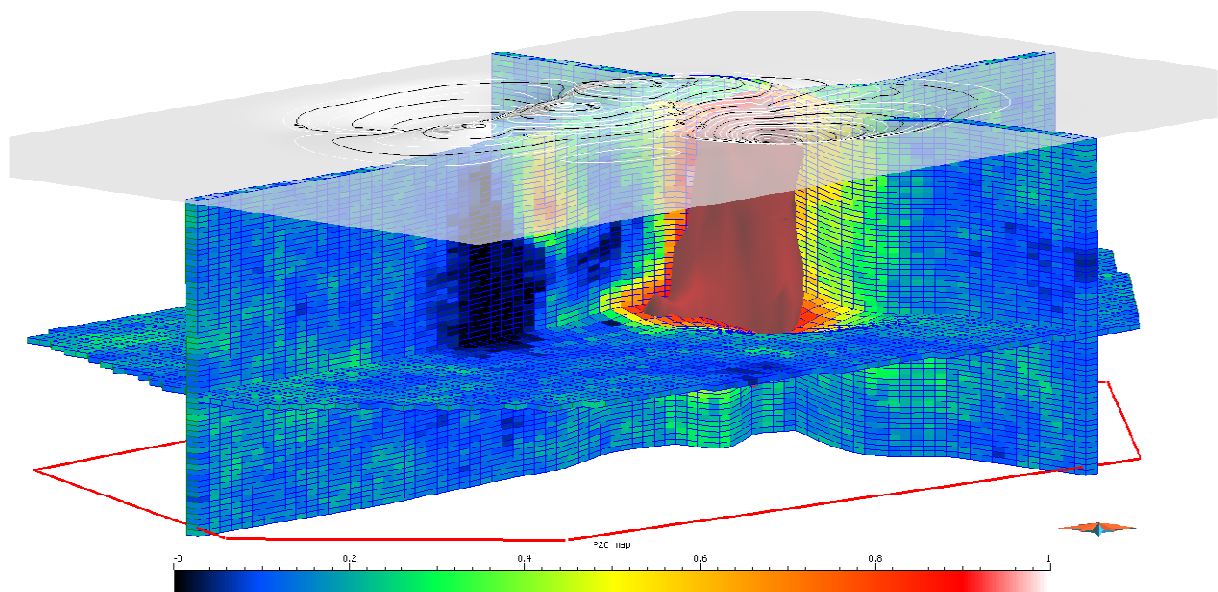


Figure 64 : Visualisations 3D de la position la plus probable de la zone source à 20 ppm si on ne choisit que les mailles pour lesquelles la probabilité de dépassement du seuil critique est supérieure ou égale à 0.8 (volume : 993 m3). La distorsion verticale de l'image est de 7 pour 1. Modèle géostatistique amélioré.

7.3 Confrontation du modèle avec les données de validation

7.3.1 Comparaison entre teneurs simulées et teneurs mesurées

Comme pour le modèle géostatistique initial, les résultats suivants ont été produits pour comparer les données de validation aux valeurs simulées.

- Diagrammes de dispersion entre valeurs simulées et données de validation (Figure 65).
- Diagrammes de dispersion entre moyennes des valeurs simulées et données de validation (Figure 66).
- Diagrammes de dispersion entre médianes des valeurs simulées et données de validation (Figure 67).

Les coefficients de corrélation correspondant aux différents diagrammes de dispersion sont regroupés dans le **Tableau 20**. Tous ces résultats sont à confronter à ceux du paragraphe §6.3.2.

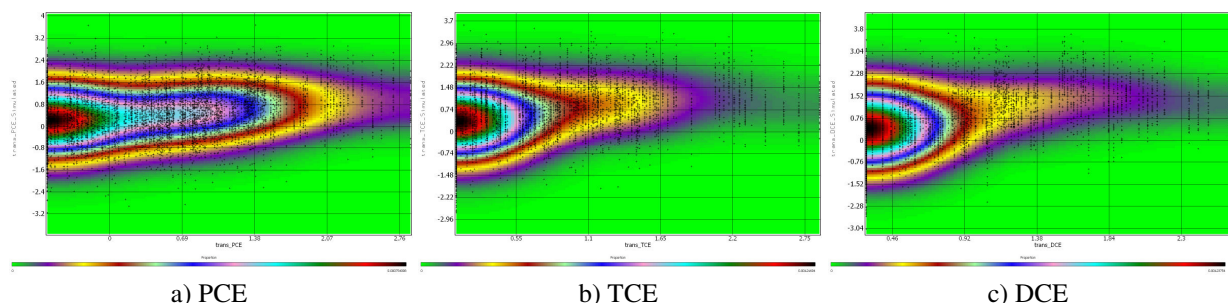


Figure 65 : Diagrammes de dispersion entre les valeurs simulées et l'ensemble des données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures). Modèle géostatistique amélioré.

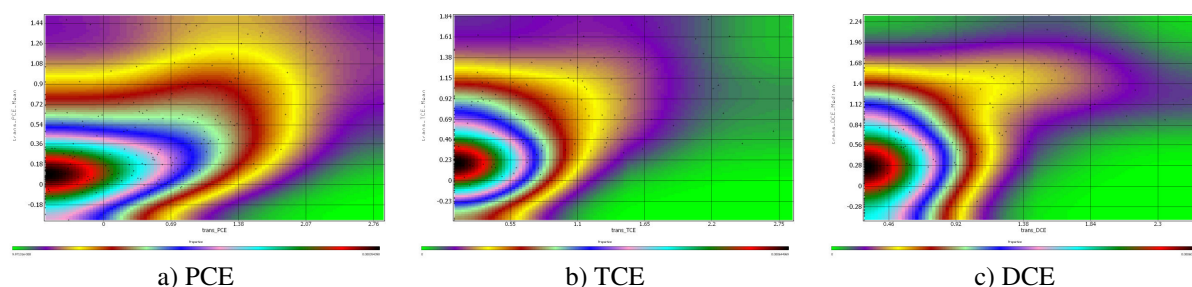


Figure 66 : Diagrammes de dispersion entre les moyennes des valeurs simulées et l'ensemble des données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures). Modèle géostatistique amélioré.

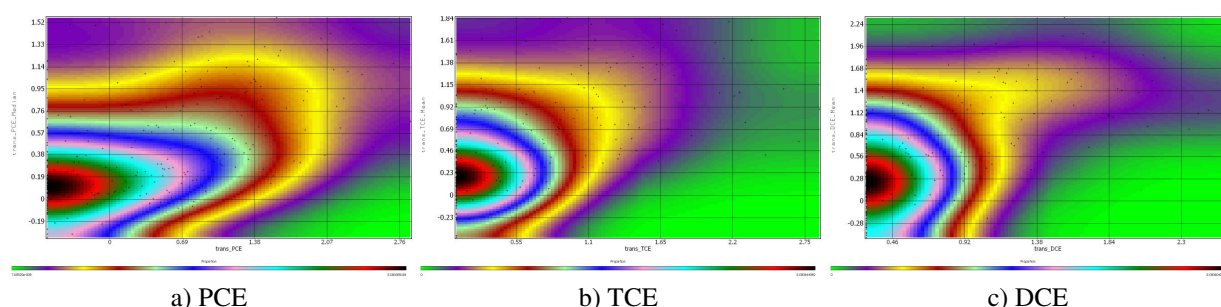


Figure 67 : Diagrammes de dispersion entre les médianes des valeurs simulées et l'ensemble des données de validation pour les trois composants (échelles = variables transformées). Les cartes de fond renseignent sur la densité de probabilité (histogrammes bidimensionnels lissés associés à une échelle de couleur arc-en-ciel pour faire ressortir les structures). Modèle géostatistique amélioré.

Tableau 20 : Coefficients de corrélation linéaire et de rang entre l'ensemble des données de validation (variables transformées) et les valeurs simulées ou des statistiques centrales tirées de ces valeurs. Modèle géostatistique amélioré.

	Valeurs simulées	Moyenne des valeurs simulées	Médiane des valeurs simulées
PCE	Données : 23 800 Linéaire : 0.26 Rang : 0.25	Données : 238 Linéaire : 0.52 Rang : 0.54	Données : 238 Linéaire : 0.51 Rang : 0.53
TCE	Données : 23 800 Linéaire : 0.33 Rang : 0.36	Données : 238 Linéaire : 0.56 Rang : 0.64	Données : 238 Linéaire : 0.56 Rang : 0.64
DCE	Données : 23 800 Linéaire : 0.41 Rang : 0.41	Données : 238 Linéaire : 0.61 Rang : 0.61	Données : 238 Linéaire : 0.61 Rang : 0.61

Les diagrammes de dispersion permettent de vérifier que les simulations tendent toujours à produire des valeurs simulées plus élevées aux endroits où les données de validation sont plus grandes. La présence de ces données à des endroits où l'incertitude locale peut varier considérablement d'une donnée à l'autre continue de rendre complexe l'analyse des prédictions du modèle géostatistique à partir de ces graphiques.

Les meilleures prédictions du modèle géostatistique amélioré semblent cependant se confirmer à travers les corrélations entre les données de validation et les moyennes ou médianes des valeurs simulées qui

sont en forte augmentation : corrélations de rang de 0.54, 0.64 et 0.61 pour le PCE, le TCE et le DCE, contre 0.42, 0.44 et 0.36 avant.

7.3.2 Evaluation de la précision et justesse du modèle amélioré

Les diagrammes de précision et justesse tirés du modèle géostatistique amélioré sont présentés à la Figure 68. Ils sont à comparer à ceux de la Figure 50 du §6.3.3.

Ces diagrammes montrent que les nouvelles simulations prédisent avec justesse les teneurs des trois composants mesurées sur les données de validation mais que l'incertitude sur les teneurs en TCE et DCE est moindre que celle attendue. La très grande proportion de données sous la limite de quantification de ces deux composants peut en partie au moins expliquer ce résultat.

Ces diagrammes confirment aussi que les modèles de distribution et de variogramme établis à partir de l'ensemble des données suffisent à mieux prédire les teneurs des trois composants avec les mêmes données de conditionnement que celles du modèle géostatistique initial (jeu de données restreint sur 49 échantillons). Ces aspects sont discutés dans les conclusions relatives au modèle géostatistique amélioré (§7.4).

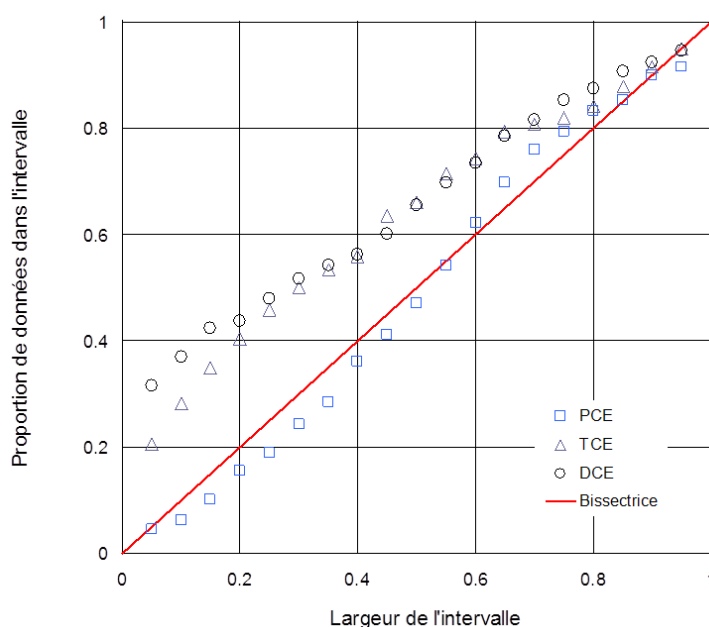


Figure 68 : Diagrammes de précision et justesse calculé sur les données de validation non extrapolées pour les trois composants (les valeurs simulées sous les seuils de quantification ont été placées au seuil). Modèle géostatistique amélioré.

7.3.3 Prédictions du modèle amélioré par rapport à la zone source

Pour juger des prédictions du modèle géostatistique amélioré en termes de classification des données de validation par rapport à la zone source, il a été suivi la même approche que celle appliquée au modèle géostatistique initial (voir détails au §6.3.3) en générant les résultats statistiques suivants.

1. Figure 69 : histogrammes des probabilités de dépassement des seuils à l'endroit des données de validation en différenciant les données dans et hors de la zone source pour les six seuils de définition de la zone source.
2. Tableau 21 : proportion de données de validation mal classées et probabilité d'erreur de classification moyenne correspondante pour différents intervalles de confiance et les six seuils de définition de la zone source.
3. Figure 70 : diagrammes des erreurs relatives entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne.

Il peut être relevé les points suivants qui tendent à confirmer les meilleures prédictions du modèle géostatistique amélioré.

- L'augmentation de la proportion de données de classification classées comme situées dans la zone source est plus marquée que précédemment (Figure 69 par rapport à Figure 51) et s'accompagne de probabilités de dépassement des seuils supérieures pour tous les seuils comme déjà observé plus haut (voir §7.2.3).
- Les proportions de données de validation mal classées sont presque toutes inférieures à celles du modèle géostatistique initial (Tableau 21 par rapport à Tableau 16) et confirment les meilleures prédictions du modèle géostatistique amélioré. Ces proportions sont conformes aux attentes (probabilités d'erreur de classification moyennes) pour les niveaux de confiance de 50 et 80% mais continuent d'être supérieures pour les niveaux de confiance les plus grands (90 et 95%).
- Les améliorations par rapport au modèle géostatistique initial sont confirmées graphiquement par les diagrammes des erreurs relatives (Figure 70 par rapport à Figure 52).

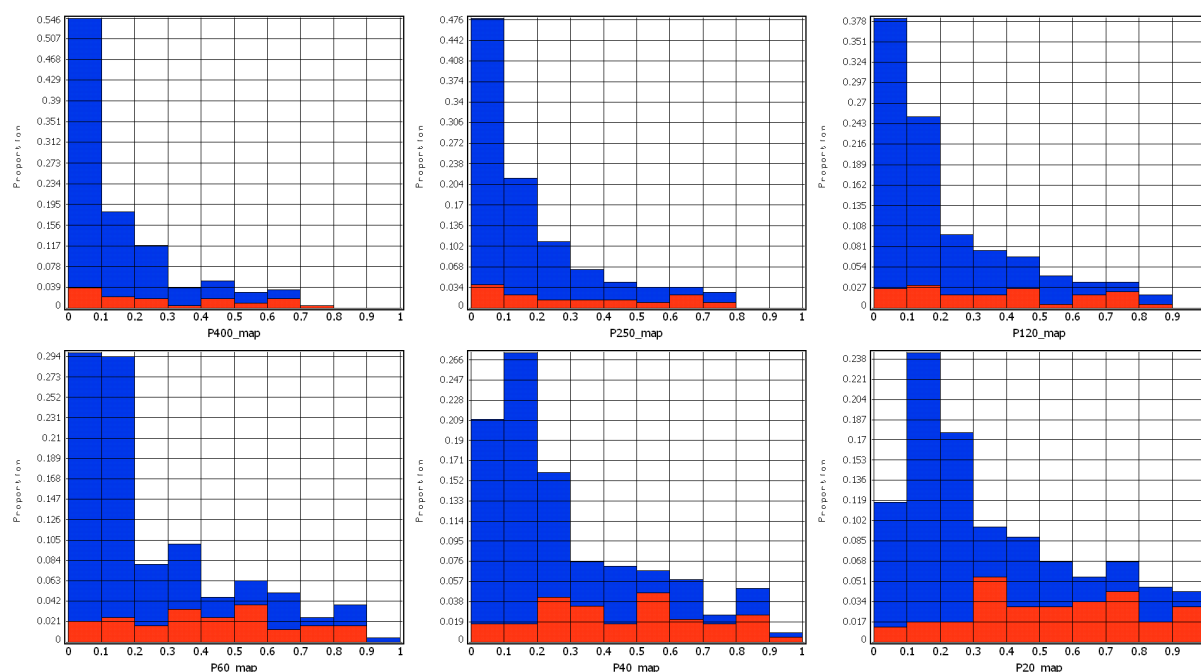


Figure 69 : Histogrammes des probabilités de dépassement des seuils, à l'endroit des données de validation pour les seuils de 400, 250, 1200, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Différentiation en rouge et en bleu des données situées respectivement dans et hors de la zone source pour le seuil de teneur considéré. Modèle géostatistique amélioré.

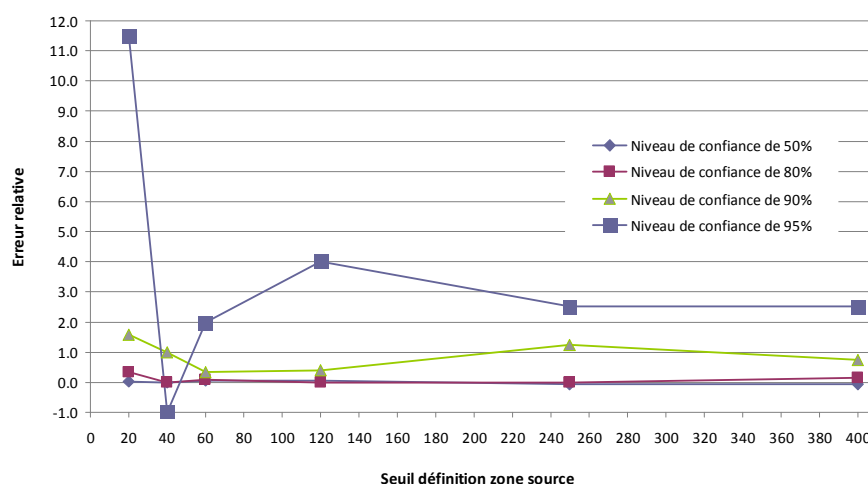


Figure 70 : Diagrammes des erreurs relatives entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne des données classées selon la valeur du seuil critique (en ppm). Modèle géostatistique amélioré.

Tableau 21 : Correspondance entre la proportion de données de validation mal classées et la probabilité d'erreur de classification moyenne des données classées. Calculs répétés pour des différents intervalles de confiance et différents seuils de définition de la zone source. Modèle géostatistique amélioré.

Seuil	Nombre données classées	Proportion de données mal classées (%)	Probabilité d'erreur moyenne (%)
Niveau de confiance de 50%			
400 ppm	238	13%	14%
250 ppm	238	15%	16%
120 ppm	238	19%	18%
60 ppm	238	21%	20%
40 ppm	238	22%	22%
20 ppm	238	25%	24%
Niveau de confiance de 80%			
400 ppm	170	8%	7%
250 ppm	164	8%	8%
120 ppm	152	9%	9%
60 ppm	147	12%	11%
40 ppm	123	11%	11%
20 ppm	104	16%	12%
Niveau de confiance de 90%			
400 ppm	115	7%	4%
250 ppm	105	9%	4%
120 ppm	76	7%	5%
60 ppm	64	8%	6%
40 ppm	42	12%	6%
20 ppm	34	18%	7%
Niveau de confiance de 95%			
400 ppm	61	7%	2%
250 ppm	46	7%	2%
120 ppm	31	10%	2%
60 ppm	23	9%	3%
40 ppm	16	0%	3%
20 ppm	4	25%	2%

7.4 Conclusions sur le modèle amélioré

Les conclusions suivantes peuvent être tirées du modèle géostatistique amélioré.

1. L'amélioration des modèles de distribution et de variogramme, sans pour autant augmenter les données de conditionnement des simulations géostatistiques, a suffi à mieux prédire les données de validation.
2. Les meilleures performances du modèle géostatistique amélioré ont pu être vérifiées de différentes façons : à partir des corrélations entre les données de validation et les teneurs simulées (moyennes et médianes), de diagrammes de justesse et précision et de la classification des données de validation comme étant situées dans ou hors de la zone source pour différents seuils de définition de la zone source.
3. Cet aspect de l'étude de démonstration illustre l'extrême importance de l'analyse exploratoire et statistique des données qui doit être menée avec grand soin et réflexion pour établir les caractéristiques statistiques et spatiales importantes du phénomène étudié et que le modèle géostatistique devra tâcher de reproduire.
4. Dans le cas présent, la solution de facilité a été d'utiliser le jeu de données complet pour améliorer les modèles statistiques. En pratique, hors d'un cadre de démonstration, le modèle géostatistique aurait pu être amélioré en prenant plus de soin à définir ces modèles, en faisant certaines hypothèses sur la continuité spatiale (effets de pépité nuls compte tenu du type de pollution des sols), en ajustant davantage les simulations pour reproduire correctement les distributions de valeurs souhaitées, en vérifiant la justesse et précision des simulations directement à partir des 49 données (validation croisée).

Après avoir défini des modèles de distribution et de variogramme qui conviennent mieux pour décrire la répartition spatiale des teneurs dans les sols, il a été étudié, comme dernière amélioration possible du modèle géostatistique, l'apport de données complémentaires comme unique moyen de réduire les incertitudes spatiales et produire des estimations plus précises sur le volume de la zone source et sur sa position. Les données de validation ont ainsi été utilisées pour compléter les données de conditionnement des simulations. Ces résultats sont présentés dans la suite.

8. MODELISATION GEOSTATISTIQUE COMPLETE AVEC L'ENSEMBLE DES DONNEES

8.1 Caractéristiques du modèle

Afin de mener à terme la démonstration de l'application de la géostatistique sur ce cas d'étude, il a été simulé un dernier modèle géostatistique le plus complet possible. Ce modèle repose sur les modèles de distribution et de variogrammes déduits de l'ensemble des données de diagnostic et de validation du modèle amélioré précédent (voir §7) mais utilise également toutes ces données comme données de conditionnement (287 teneurs mesurées pour chaque composant). Les objectifs sont doubles.

1. Produire un modèle de référence sur l'état de pollution des sols.
2. Démontrer l'apport des données de validation comme données complémentaires susceptibles de réduire l'incertitude spatiale.

8.2 Résultats du modèle géostatistique complet

8.2.1 Simulation des teneurs

Comme pour les deux modèles précédents, cent simulations conjointes des teneurs en PCE, TCE et DCE ont été générées. La Figure 71 présente deux simulations particulières sur le même plan de coupe horizontal que ceux de la Figure 59 (§7.2.1) et de la Figure 35 (§5.1.1). Visuellement, les nouvelles simulations continuent de ressembler aux précédentes, les distributions de teneurs simulées étant les mêmes que celles du modèle géostatistique amélioré. A partir de ces seules réalisations, il n'est ainsi pas possible de juger de l'impact du nombre plus grand de données de conditionnement.

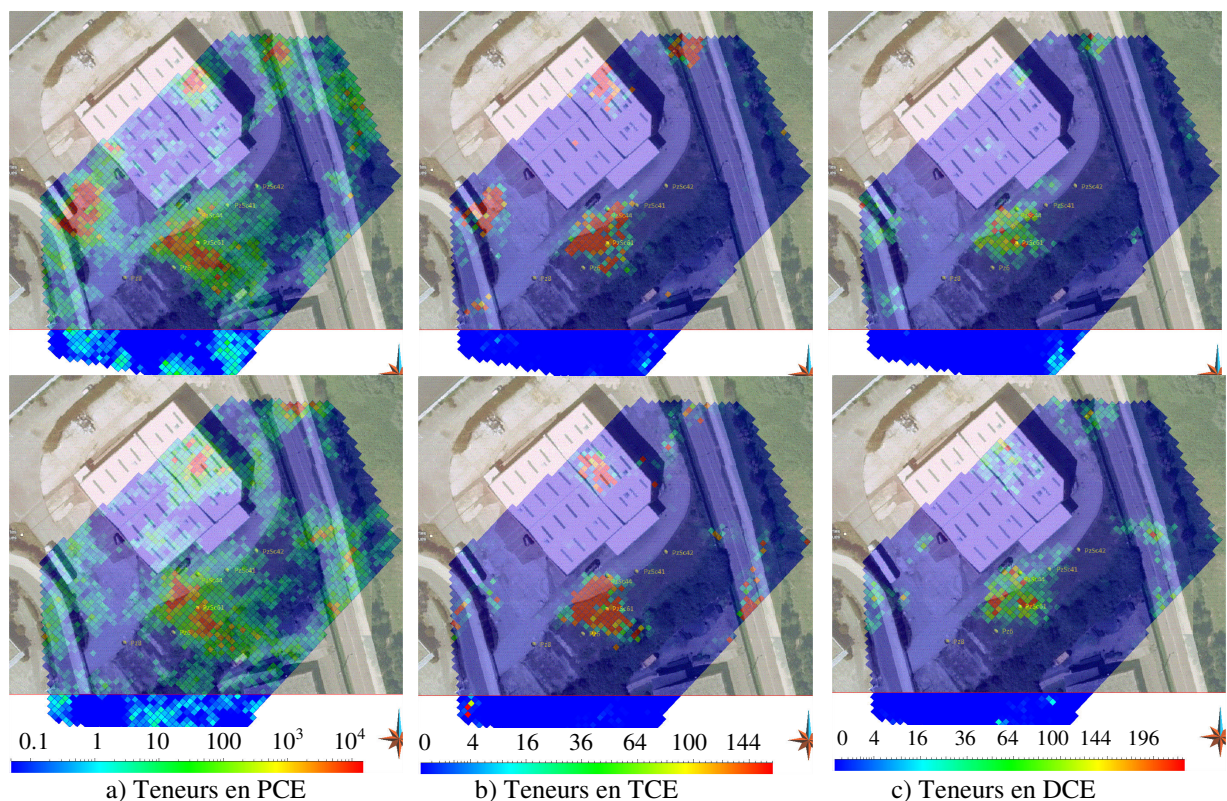


Figure 71 : Coupes horizontales de deux simulations conjointes des teneurs pour les trois polluants (unité : ppm). Les sondages affichés sont ceux d'où proviennent les données de teneur utilisées pour conditionner les simulations. Modèle géostatistique complet.

8.2.2 Incertitude sur le volume de la zone source

La Figure 72 présente les histogrammes et les résumés statistiques des volumes de la zone source calculés à partir des réalisations pour les différents seuils de définition de 400 à 20 ppm. Comparés à ceux du modèle géostatistique amélioré, qui repose sur les mêmes modèles de distribution et de variogramme (§7.2.2) mais n'utilise que 49 échantillons de sols comme données de conditionnement, ces volumes se révèlent 13 à 19% plus petits, la différence étant d'autant plus importante que le seuil de définition de la zone source est élevé.

Comme attendu, le plus grand nombre de données de conditionnement (287 pour chaque composant) conduit à diminuer l'incertitude sur les volumes de la zone source. Ces volumes peuvent ainsi être estimés plus précisément comme le montrent par exemple les écarts-types qui sont environ 20% inférieurs pour tous les seuils de définition.

Ce gain de précision peut cependant paraître modeste comparé au nombre de données qui est multiplié par presque 6 par rapport aux modèles géostatistiques précédents. Ce résultat s'explique par la position des données dont la plupart sont situées dans la partie centrale du site. Beaucoup de mailles de la grille de simulation se trouvent ainsi éloignées des données et donc fortement incertaines. En éliminant ces mailles, dont on sait qu'elles ne sont pas affectées par la pollution des sols et n'ont pas été explorées, il peut être évalué plus justement l'apport des nouvelles données (données de validation) sur la précision d'estimation des volumes. Le gain de précision (diminution de l'incertitude) apparaît bien meilleur comme cela est illustré au §9.2.

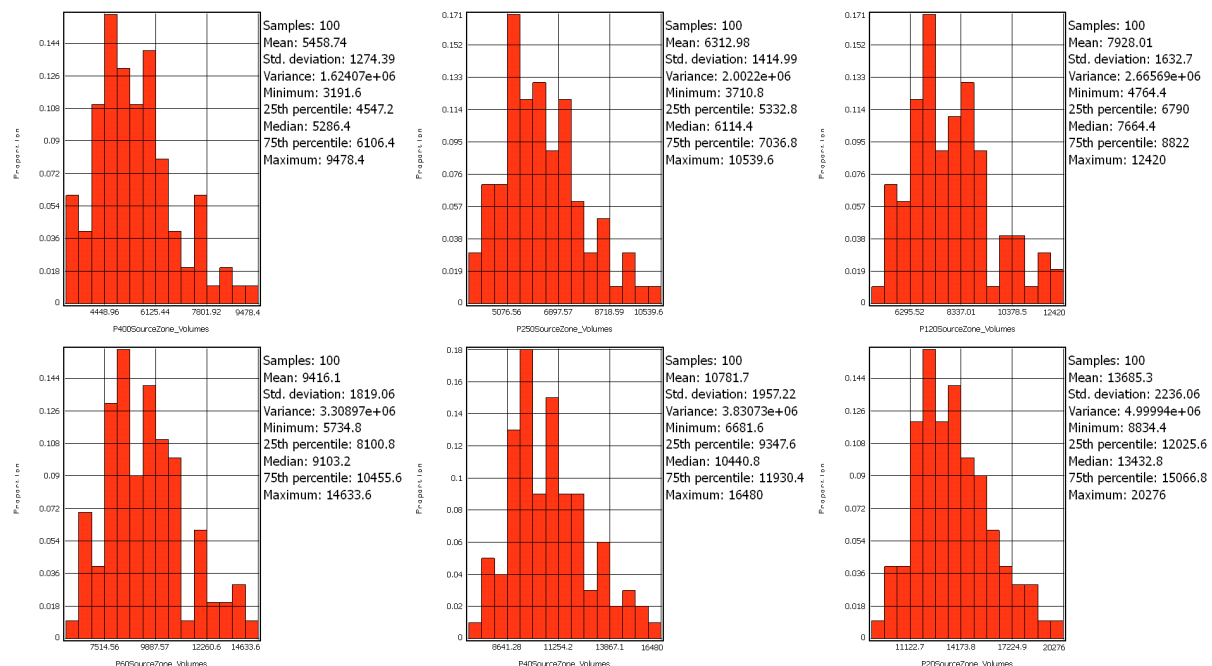


Figure 72 : Histogramme des volumes en m³ de la zone source calculés sur les 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique complet.

8.2.3 Incertitude spatiale sur la position de la zone source

De la même façon que pour les modèles géostatistiques précédents (voir §7.2.3 et §5.2.3), l'incertitude sur la position de la zone source est étudiée à partir de cartes de probabilité pour la somme des teneurs en PCE, TCE et DCE de dépasser le seuil de définition considéré.

La Figure 73 et la Figure 74 montrent les coupes horizontales et verticales de la carte de probabilité de présence de la zone source calculée à partir des 100 simulations pour un seuil de teneur totale de 20 ppm. Ces plans de coupe sont identiques à ceux du modèle géostatistique amélioré de la Figure 61 et

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

de la Figure 62. Sur ces nouvelles cartes, les régions de grandes et faibles probabilités, en rouge et bleu foncé aux profondeurs de 2.45–2.55 m et 3.35–3.45 m, apparaissent plus distinctivement que précédemment, tandis que la région incertaine, en vert et bleu clair, voit son extension fortement réduite en périphérie de la région de fortes probabilités. Ce résultat traduit la plus grande proportion de mailles qui peuvent être reconnues avec confiance comme étant situées dans ou hors de la zone source. Les mailles incertaines, entre celles de grandes et faibles probabilités, forment une région de transition dans laquelle on peut s'attendre à trouver les limites de la zone source.

Des mailles incertaines continuent d'exister également sur tout le pourtour du site. Comme indiqué précédemment (§8.2.2) et détaillé plus loin (§9.1), l'incertitude sur l'état de pollution de ces mailles est simplement dû à leur éloignement par rapport aux données, aucune source de pollution n'étant attendu à l'endroit de ces mailles qui ont été exclues des campagnes de reconnaissance.

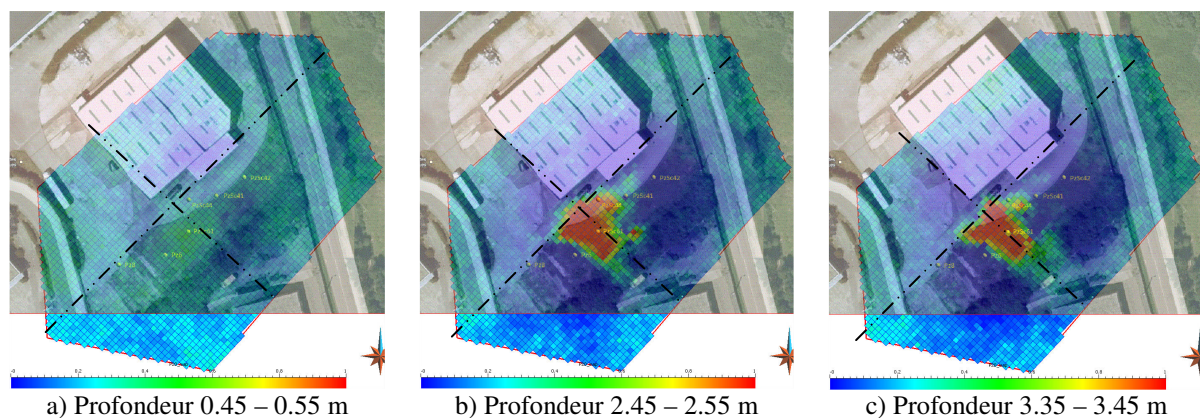


Figure 73 : Coupes horizontales de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE. Modèle géostatistique complet.

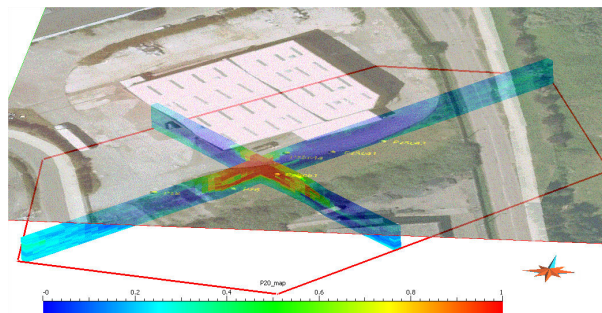


Figure 74 : Coupes verticales vues en perspective de la probabilité de présence de la zone source calculée à partir des 100 simulations conjointes pour un seuil critique de 20 ppm appliqué à la somme des teneurs en PCE, TCE et DCE. Modèle géostatistique complet.

Les histogrammes et les résumés statistiques des probabilités de présence de la zone source calculées pour les seuils de 400 à 20 ppm sont regroupés dans la Figure 75. Comparés à ceux du modèle géostatistique amélioré (Figure 63), ces histogrammes révèlent une plus grande proportion de mailles avec des probabilités inférieures à 0.2 ou supérieures à 0.8 et confirment donc que le modèle géostatistique complet prédit plus précisément la position de la zone source (diminution de la proportion de mailles incertaines). Le gain en précision du modèle est d'autant meilleur que le seuil de définition de la zone source est petit et que donc son volume est grand. En appliquant les mêmes critères de classification que ceux utilisés au paragraphe §7.2.3, ces tendances se résument comme suit.

- Les mailles que l'on peut classer comme étant hors de la zone source avec un niveau de confiance de 80% représentent 97.8% du volume pour le seuil de 400 ppm et 66.3% pour le seuil de 20 ppm. Ces volumes sont à comparer à ceux du modèle géostatistique amélioré conditionné par les

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

données de 49 échantillons : 95.65% du volume classé avec confiance hors de la zone source pour le seuil de 400 ppm et 68.5% pour le seuil de 20 ppm. Pour un même niveau de confiance de 80%, les 238 données complémentaires ne permettent ainsi pas de classer davantage de sols comme étant situés hors de la zone source. Un tel résultat témoigne de données complémentaires inadaptées à une meilleure délimitation des sols situés hors de la zone source, et ce quel que soit le seuil de définition considéré. Il est conforme aux conclusions de l'analyse des pondérateurs calculés sur les données de validation et qui révèle un échantillonnage préférentiel des sols pollués (voir §6.2, Figure 43).

- Les mailles que l'on peut classer comme situées dans la zone source avec le même niveau de confiance de 80% forment 0.44% du volume pour le seuil de 400 ppm et 1.7% pour le seuil de 20 ppm. Le volume de la zone source étant très petit, ces proportions continuent d'être très faibles mais sont en forte progression par rapport au modèle géostatistique amélioré (0.04% du volume pour le seuil de 400 ppm et 1.3% pour le seuil de 20 ppm).
- Les mailles incertaines, qui n'entrent pas dans une des deux classes précédentes, représentent 1.77% des mailles pour le seuil de 400 ppm et 32% pour celui de 20 ppm. En accord avec les résultats précédents, elles sont en diminution pour le seuil de 400 ppm (4.31% du volume pour le modèle amélioré) et relativement identiques pour celui de 20 ppm (30.2% du volume pour le modèle amélioré). Les données du projet ATTENA n'ont pas été positionnées pour réduire les incertitudes spatiales mais pour étudier finement la migration de produits organo-halogénés dans les sols. Dans ces conditions, il n'est pas surprenant que l'intégration des données de validation dans le modèle géostatistique ait peu d'impact sur les incertitudes spatiales.

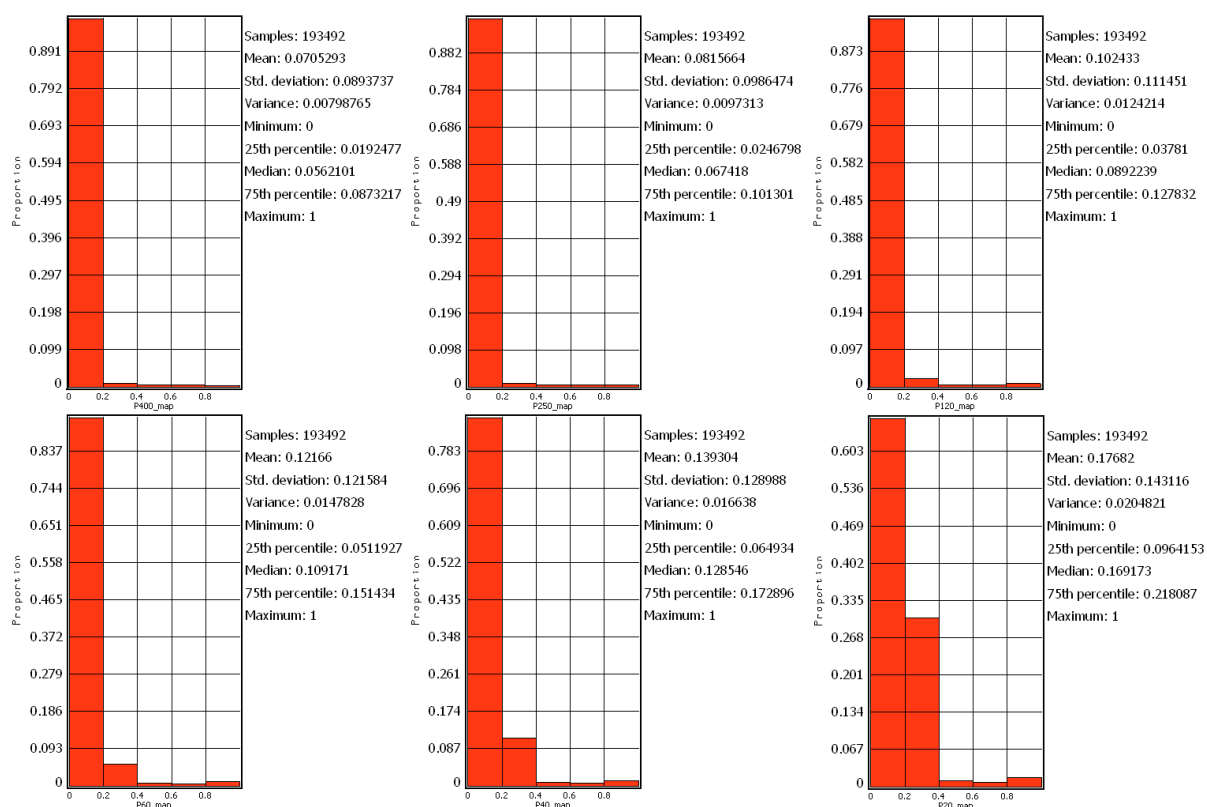


Figure 75 : Histogrammes et résumés statistiques des probabilités de présence de la zone source calculées à partir des 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique complet.

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

La Figure 76 fournit la visualisation 3D de la position la plus probable de la zone source, pour un seuil de 20 ppm et une probabilité de présence de la zone source supérieure à 0.8 (plus de 80% de chances). Elle est à comparer à la Figure 64 du modèle géostatistique amélioré. Sur les 13 700 m³ que la zone source peut représenter en moyenne, 1 346 m³ peuvent être positionnés avec un niveau de confiance de 80% de chances, contre 993 m³ pour le modèle géostatistique amélioré conditionné par 49 échantillons. La position la plus probable apparaît ainsi plus affinée et révèle quelques mailles déconnectées des autres mais dont on peut penser que la séparation est simplement due à un défaut d'information (manque de données localement pour reconnaître précisément la géométrie complexe de la zone source à cet endroit). **Comme indiqué à plusieurs reprises précédemment, l'impact réduit des données complémentaires de validation sur la délimitation de la zone à 20 ppm s'explique par le positionnement inadapté de ces données qui sont situées préférentiellement dans des sols pollués.**

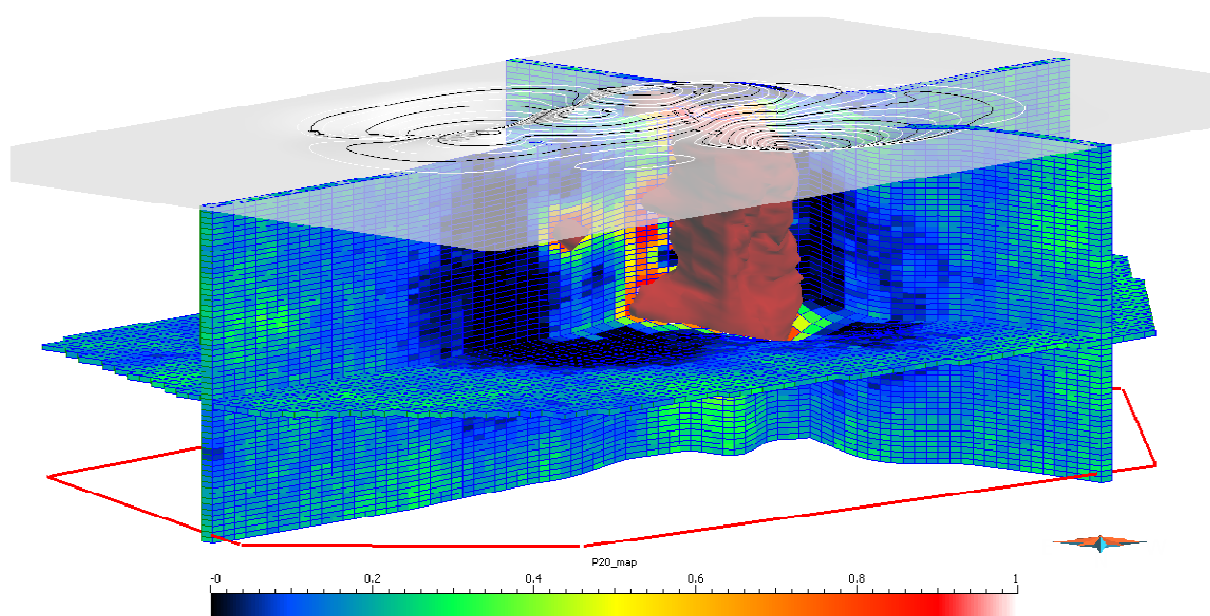


Figure 76 : Visualisations 3D de la position la plus probable de la zone source à 20 ppm si on ne choisit que les mailles pour lesquelles la probabilité de dépassement du seuil critique est supérieure ou égale à 0.8 (volume : 1 346 m³) (Distorsion verticale = 7 pour 1). Modèle géostatistique complet.

9. SYNTHÈSE DES VOLUMES DE LA ZONE SOURCE ESTIMÉS ET LOCALISÉS À PARTIR DES TROIS MODÈLES GÉOSTATISTIQUES

Les trois modèles géostatistiques fournissent des estimations du volume de la zone source qui varient plus ou moins d'un modèle à l'autre en fonction du seuil de définition de la source. Ces modèles permettent également de classer avec confiance certains sols comme étant situés dans ou hors de la zone source. La proportion de sols qui peuvent être classés ainsi avec confiance varie également entre les modèles géostatistiques. Elle est étroitement liée à l'incertitude locale prédite par chaque modèle et qui dépend :

- des modèles statistiques (distributions, corrélations et variogrammes) introduits dans le modèle géostatistique à partir des données disponibles : sur 49 échantillons pour le modèle géostatistique de référence, sur 287 échantillons pour les modèles géostatistiques amélioré et complet,
- du nombre et de la position des données utilisées pour conditionner les simulations : sur 49 échantillons pour les modèles géostatistiques de référence et amélioré, sur 287 échantillons pour le modèle géostatistique complet,

Le choix du domaine d'étude n'est pas sans conséquence non plus sur la capacité du modèle géostatistique à prédire l'état de pollution des sols. Dans le cas présent, le domaine recouvre l'ensemble du site et intègre des sols éloignés de tous les sondages de reconnaissance. Par manque de données, l'état de pollution de ces sols situés sur le pourtour du site est nécessairement incertain, bien qu'il y ait de bonnes raisons de croire qu'ils ne sont pas affectés par la pollution. L'élimination de ces zones en se concentrant sur un domaine restreint autour de la région investiguée permet de mieux juger de l'incertitude spatiale prédite par les modèles géostatistiques par rapport aux données disponibles.

Ce dernier chapitre récapitule les résultats de volumes obtenus avec les trois modèles géostatistiques afin de mettre en évidence l'apport de chacun par rapport aux autres. Y sont présentés d'abord les volumes calculés sur l'ensemble du site et qui ont fait l'objet des chapitres précédents, puis ceux situés à l'intérieur d'un domaine restreint en éliminant les sols trop incertains en périphérie du site.

9.1 Estimation des volumes de la zone source sur l'ensemble du domaine

Le **Tableau 22** et le **Tableau 23** fournissent un récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour les deux seuils de définition de la zone source de 20 et 400 ppm. Les pourcentages de mailles « Incertaines » indiqués dans les tableaux correspondent aux mailles qui ne peuvent pas être classées comme situées dans ou hors de la zone source avec un niveau de confiance de 80%.

En comparant les trois modèles géostatistiques générés successivement, les résultats importants à retenir peuvent être résumés comme suit.

9.1.1 Modèle amélioré comparé au modèle initial

- **Amélioration apportée au modèle géostatistique**
 - Modèles de distribution et de variogramme établis à partir d'un beaucoup plus grand nombre de données en incluant les données de validation.
- **Conséquences sur les volumes de la zone source estimés**
 - Volume moyen inférieur d'environ 10% pour le seuil de 20 ppm.
 - Sans effet pour le seuil de 400 ppm.
 - La différence de volumes pour le seuil de 20 ppm est due à l'amélioration des modèles de distribution des teneurs qui révèlent une proportion de données faibles à moyennes plus grande qu'initialement estimée (voir §6.2).

- **Conséquences sur le positionnement de la zone source**
 - Une plus grande proportion de sols peut être classée avec confiance comme située dans la zone source : 6.29% contre 0.89% pour le seuil de 20 ppm, 0.44% contre 0.07% pour le seuil de 400 ppm.
 - Ces proportions demeurent cependant très faibles sans l'intégration de nouvelles données de conditionnement pour lever les incertitudes spatiales (données de conditionnement issues de 49 échantillons pour les deux modèles géostatistiques).
- **Conséquences sur l'incertitude spatiale**
 - Pour le seuil de 20 ppm : le pourcentage de mailles incertaines est considérablement réduit (30.2% contre 55.44%).
 - Pour le seuil de 400 ppm : légère augmentation du pourcentage de mailles incertaines qui reste très petit (4.31% contre 3.22%). Il est à noter que cette faible proportion de mailles incertaines s'explique par le plus grand nombre de mailles classées comme très certainement saines quand le seuil est augmenté de 20 à 400 ppm.
 - La diminution du pourcentage de mailles incertaines pour le seuil de 20 ppm est une autre conséquence de l'amélioration des modèles de distribution des teneurs (proportion plus grande de données faibles à moyennes).

9.1.2 Modèle complet comparé au modèle amélioré

- **Améliorations apportées au modèle géostatistique**
 - Intégration des 238 données de validation comme données de conditionnement des simulations (multiplication par presque 6 des données de conditionnement).
- **Conséquences sur les volumes de la zone source estimés**
 - Volume moyen inférieur d'environ 13% pour le seuil de 20 ppm.
 - Volume moyen inférieur de presque 20% pour le seuil de 400 ppm.
 - Les différences de volumes s'expliquent nécessairement par le meilleur conditionnement des simulations qui est la seule différence entre les deux modèles géostatistiques (données sur 287 échantillons contre 49). Certaines données complémentaires situées à l'extérieur de la zone source doivent se trouver à des endroits où l'incertitude est grande sans ces données.
- **Conséquences sur le positionnement de la zone source**
 - Comme attendu, une plus grande proportion de sols peut être classée avec confiance comme située dans la zone source.
 - Cette augmentation est modeste pour le seuil de 20 ppm (9.84% contre 6.29%) mais plus conséquente pour le seuil de 400 ppm (6.17% contre 0.44%) pour lequel le volume moyen de la zone source est beaucoup plus petit.
 - Dans tous les cas, ces proportions continuent d'apparaître très faibles et ce malgré la multiplication par presque 6 des données de conditionnement. L'étendue du domaine d'étude explique en partie ce résultat comme cela est démontré dans la suite (§9.2). La position des données de conditionnement complémentaires (données de validation) complète l'explication. Le choix de l'emplacement de ces données n'a en effet pas été dicté par un objectif de délimitation de la zone source mais par des considérations scientifiques pour étudier finement la physique de migration de composants organo-halogénés dans les sols.
- **Conséquences sur l'incertitude spatiale**
 - Pour le seuil de 20 ppm : le pourcentage de mailles incertaines est en légère augmentation (32.0% contre 30.2%).

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

- Pour le seuil de 400 ppm : le pourcentage de mailles incertaines, déjà petit précédemment, est divisé par presque 2.5 (1.77% contre 4.31%).
- Comme pour le positionnement avec confiance des sols situés dans la zone source, l'impact faible ou inexistant des nombreuses données complémentaires du modèle géostatistique résulte de l'étendue du domaine d'étude et du positionnement inadapté de ces données qui se trouvent situées préférentiellement dans des sols pollués.

Tableau 22 : Récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour un seuil de 20 ppm en considérant l'ensemble du domaine.

	Volume moyen (en m ³)	Ecart-type sur le volume moyen	Proportion du volume de la zone source localisé avec confiance	Pourcentage de mailles « Incertaines »
Modèle de référence	17 118	2 168	0.89%	55.4%
Modèle amélioré	15 783	2 802	6.29%	30.2%
Modèle complet	13 685	2 236	9.84%	32.0%

Tableau 23 : Récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour un seuil de 400 ppm en considérant l'ensemble du domaine.

	Volume moyen (en m ³)	Ecart-type sur le volume moyen	Proportion du volume de la zone source localisé avec confiance	Pourcentage de mailles « Incertaines »
Modèle initial	6 749	1 201	0.07%	3.22%
Modèle amélioré	6 764	1 622	0.44%	4.31%
Modèle complet	5 459	1 274	6.17%	1.77%

9.2 Estimation des volumes de la zone source sur un domaine restreint

Afin de mieux démontrer l'apport des modèles géostatistiques successifs, les volumes et positions de la zone source ont été étudiés à l'intérieur d'un domaine restreint qui ignore les mailles incertaines situées sur le pourtour du site, loin des sondages de reconnaissance où aucune source de pollution n'est attendue. La Figure 77 montre les limites du domaine restreint et la grille réduite correspondante.

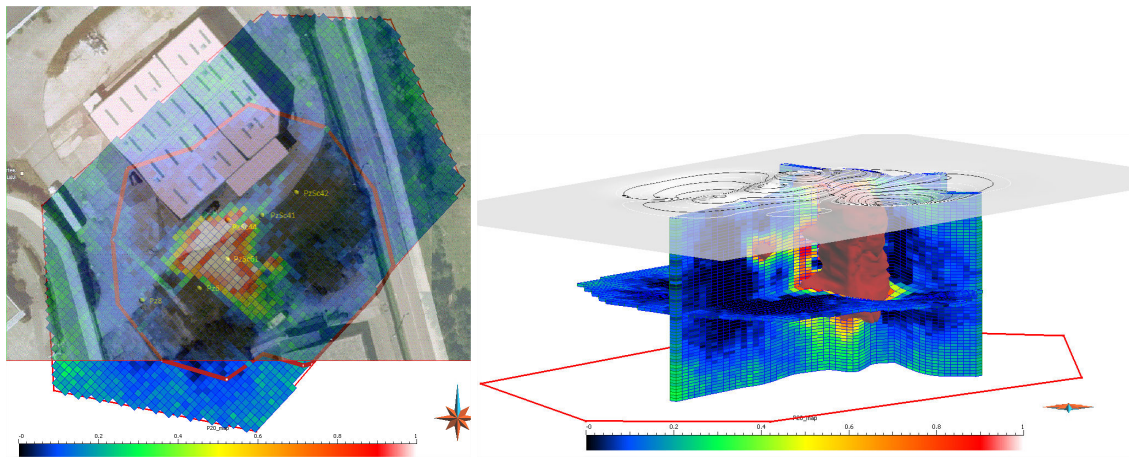
Le domaine restreint occupe un volume de 36 673 m³. Comparé aux 77 397 m³ du domaine d'étude complet, cela signifie que plus de la moitié des mailles étudiées précédemment se trouvent dans des zones mal renseignées où l'incertitude est grande sur l'état de pollution des sols. Les conséquences de cette incertitude sur les calculs de volume dépend alors à la fois du seuil de définition considéré pour la zone source et des distributions de teneurs à simuler déduites des données disponibles.

Les statistiques sur les volumes de la source ont ainsi été recalculées à l'intérieur du domaine restreint pour les trois modèles géostatistiques : modèle de référence conditionné par 49 données (Figure 78), modèle amélioré conditionné par les mêmes 49 données (Figure 79), modèle complet conditionné par 287 = 49 + 238 données (Figure 80). Ces histogrammes sont à comparer à ceux obtenus sur le domaine d'étude complet : Figure 37 pour le modèle de référence, Figure 60 pour le modèle amélioré, Figure 72 pour le modèle complet.

Le domaine restreint est supposé contenir la seule et unique zone source identifiée à partir des données, les mailles exclues de ce domaine étant toutes situées en périphérie du site où aucune source de pollution n'est susceptible d'exister. On peut donc s'attendre à ce que les volumes de sols pollués calculés sur le domaine d'étude complet se retrouvent en grande partie dans le domaine restreint. Tel

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

n'est pas le cas. Avec un volume divisé par deux, le domaine restreint est le siège d'un volume de zone source qui lui-aussi est divisé par 2 pour le modèle géostatistique de référence, par 1.8 pour le modèle géostatistique amélioré et par 2.5 pour le modèle géostatistique complet, et ce quel que soit le seuil de définition de la source considéré. Plus ou moins de la moitié des volumes de zone source prédits par les trois modèles géostatistiques se trouvent ainsi localisés dans les mailles incertaines (mal renseignées) du pourtour du site. L'élimination de ces mailles permet de mieux juger de l'impact des améliorations apportées au modèle géostatistique sur l'estimation des volumes de la zone source et de leur positionnement.



a) Délimitation du domaine restreint.

b) Grille réduite contenue dans le domaine restreint.

Figure 77 : Délimitation du domaine restreint et de la grille réduite associée utilisée pour estimer les volumes de la zone source.

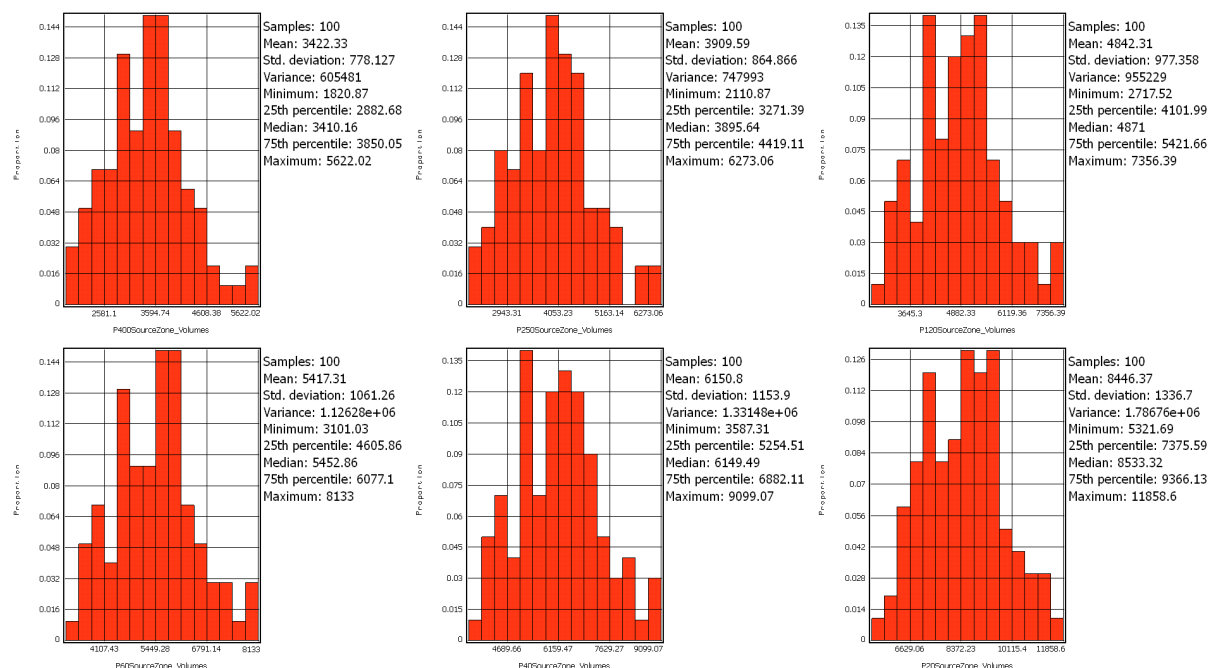


Figure 78 : Histogramme des volumes en m^3 de la zone source calculés sur les 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique initial limité au domaine restreint.

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

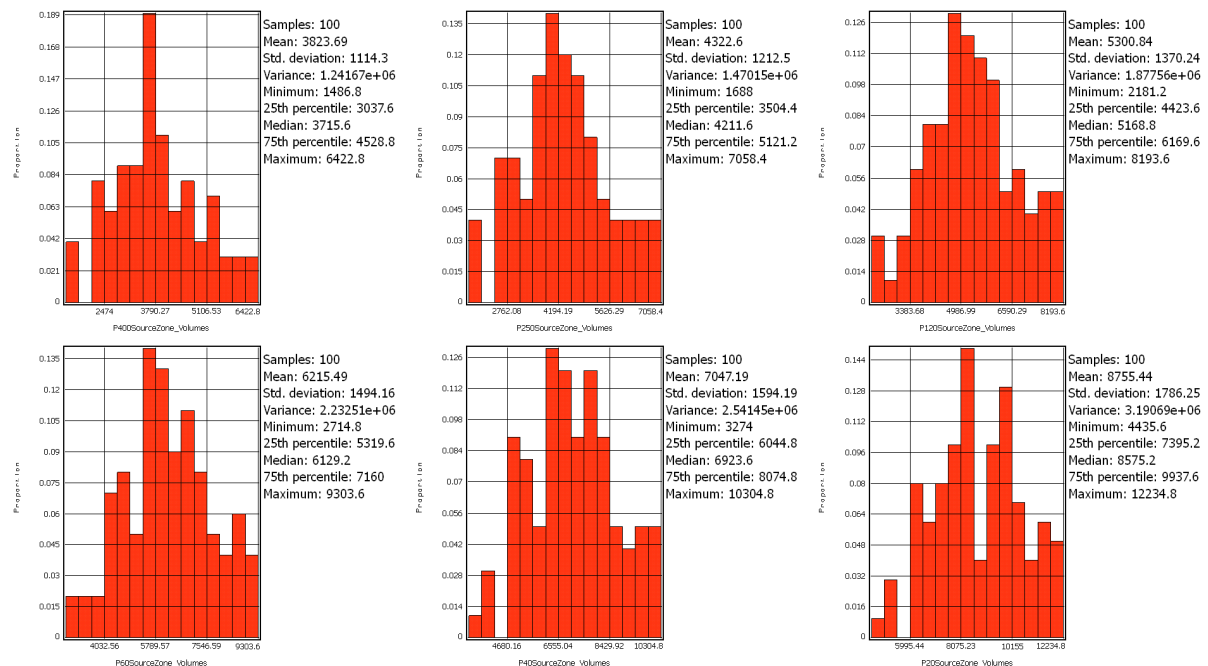


Figure 79 : Histogramme des volumes en m³ de la zone source calculés sur les 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique amélioré limité au domaine restreint.

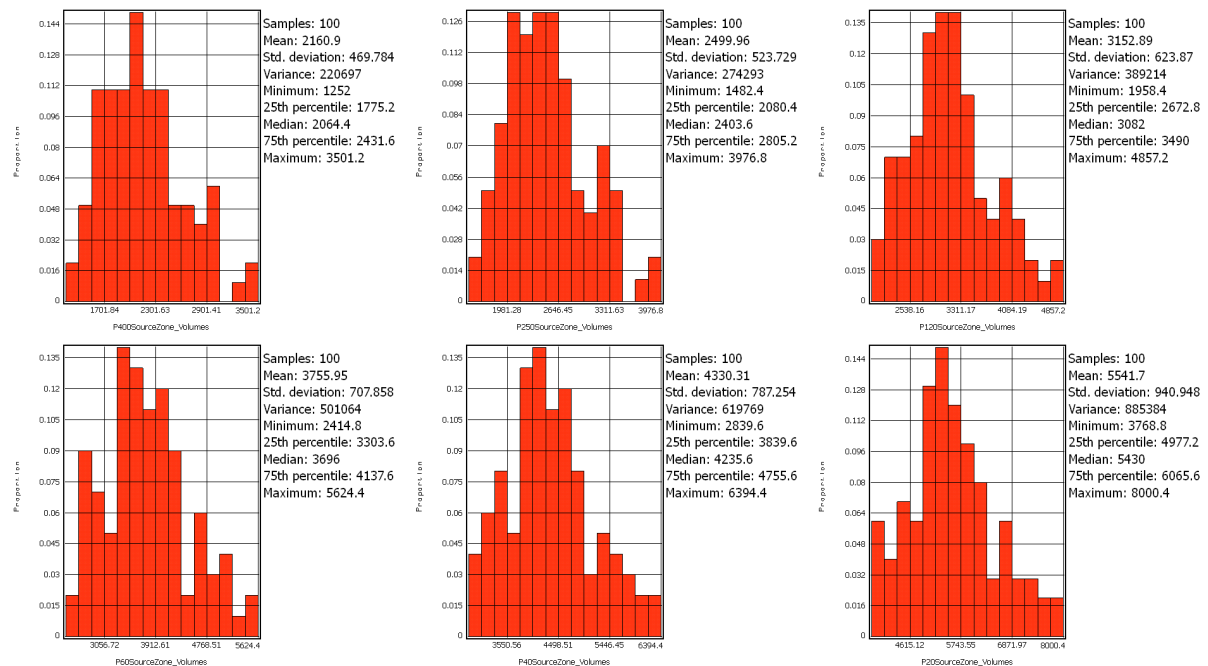


Figure 80 : Histogramme des volumes en m³ de la zone source calculés sur les 100 simulations conjointes des trois composants en appliquant à la somme des teneurs les seuils de 400, 250, 120, 60, 40 et 20 ppm (de gauche à droite et de haut en bas). Modèle géostatistique complet limité au domaine restreint.

De la même façon que pour le domaine d'étude complet (§9.1), le **Tableau 24** et le **Tableau 25** fournissent un récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour les deux seuils de définition de la zone source de 20 et 400 ppm en se limitant au domaine d'étude restreint. Les pourcentages de mailles « Incertaines » indiqués dans les tableaux correspondent aux mailles qui ne peuvent pas être classées comme situées dans ou hors de la zone source avec un niveau de confiance de 80%.

En comparant les trois modèles géostatistiques générés successivement, les résultats importants à retenir peuvent être résumés comme suit.

9.2.1 Modèle amélioré comparé au modèle de référence

- **Amélioration apportée au modèle géostatistique**
 - Modèles de distribution et de variogramme établis à partir d'un beaucoup plus grand nombre de données en incluant les données de validation.
- **Conséquences sur les volumes de la zone source estimés**
 - Le volume moyen est supérieur de 4% pour le seuil de 20 ppm et 10% pour celui de 400 ppm.
 - Ce résultat confirme que la diminution du volume observée sur le domaine d'étude complet pour le seuil de 20 ppm (§9.1) provient des mailles incertaines situées en périphérie du site.
 - L'estimation à la hausse des volumes de la zone source résulte uniquement des corrections ou améliorations apportées aux modèles de distribution et de variogramme.
- **Conséquences sur le positionnement de la zone source**
 - Une plus grande proportion de sols peut être classée avec confiance comme située dans la zone source : 11.3% contre 1.8% pour le seuil de 20 ppm, 0.77% contre 0.14% pour le seuil de 400 ppm.
 - Ces proportions, en forte progression par rapport aux statistiques sur le domaine d'étude complet, demeurent cependant très faibles sans l'intégration de nouvelles données de conditionnement pour lever les incertitudes spatiales (données de conditionnement issues de 49 échantillons pour les deux modèles géostatistiques).
- **Conséquences sur l'incertitude spatiale**
 - Pour le seuil de 20 ppm : le pourcentage de mailles incertaines est considérablement réduit (39.7% contre 54.1%).
 - Pour le seuil de 400 ppm : le pourcentage de mailles incertaines augmentent mais reste très petit (9.09% contre 6.79%).
 - Comme pour le domaine d'étude complet, la diminution du pourcentage de mailles incertaines pour le seuil de 20 ppm est une conséquence de l'amélioration des modèles de distribution des teneurs (proportion plus grande de données faibles à moyennes).

9.2.2 Modèle complet comparé au modèle amélioré

- **Améliorations apportées au modèle géostatistique**
 - Intégration des 238 données de validation comme données de conditionnement des simulations (multiplication par presque 6 des données de conditionnement).
- **Conséquences sur les volumes de la zone source estimés**
 - Le volume moyen est réduit d'environ 40% pour les deux seuils de 20 et 400 ppm.
 - Les différences de volumes s'expliquent à nouveau par le meilleur conditionnement des simulations qui est la seule différence entre les deux modèles géostatistiques (données sur 287 échantillons contre 49). Comme pour le domaine d'étude complet, certaines données

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

complémentaires situées à l'extérieur de la zone source doivent se trouver à des endroits où l'incertitude est grande sans ces données.

- Les diminutions de volumes sont plus importantes en pourcentage que celles observées pour le domaine d'étude complet du fait de l'élimination des nombreuses mailles incertaines situées en périphérie du site.
- **Conséquences sur le positionnement de la zone source**
 - Comme attendu, une beaucoup plus grande proportion de sols peut être classée avec confiance comme située dans la zone source : 24.3% contre 11.3% pour le seuil de 20 ppm, 15.6% contre 0.77% pour le seuil de 400 ppm.
 - Bien que ces pourcentages soient en net progression par rapport au domaine d'étude complet, ils continuent de paraître modestes en regard du nombre de données de conditionnement multiplié par presque 6. Le domaine restreint permettant de faire abstraction des mailles incertaines situées en périphérie du site, ces améliorations modérées s'expliquent essentiellement par le positionnement inadapté des 238 données complémentaires dont l'objectif n'était pas de mieux délimiter la zone source mais d'étudier la migration de composants organo-halogénés dans les sols.
- **Conséquences sur l'incertitude spatiale**
 - Contrairement au domaine d'étude complet, le pourcentage de mailles incertaines diminue considérablement pour les deux seuils de définition de la source : 16.4% contre 39.7% pour le seuil de 20 ppm, 3.66% contre 9.09% pour le seuil de 400 ppm.
 - Malgré une répartition non optimale des 238 données complémentaires, le modèle géostatistique complet permet de réduire beaucoup l'incertitude spatiale. Un meilleur choix d'emplacement de ces données, en prenant en compte les zones incertaines établies par le modèle géostatistique précédent, permettrait de réduire davantage l'incertitude spatiale avec certainement un plus petit nombre de données complémentaires.

Tableau 24 : Récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour un seuil de 20 ppm en considérant le domaine restreint de 36 673 m³.

	Volume moyen (en m ³)	Ecart-type sur le volume moyen	Proportion du volume de la zone source localisé avec confiance	Pourcentage de mailles « Incertaines »
Modèle initial	8 446	1 337	1.8%	54.1%
Modèle amélioré	8 755	1 786	11.3%	39.7%
Modèle complet	5 542	941	24.3%	16.4%

Tableau 25 : Récapitulatif de l'incertitude sur les volumes de la zone source et sur leur localisation pour les trois modèles géostatistiques et pour un seuil de 400 ppm en considérant le domaine restreint de 36 673 m³.

	Volume moyen (en m ³)	Ecart-type sur le volume moyen	Proportion du volume de la zone source localisé avec confiance	Pourcentage de mailles « Incertaines »
Modèle initial	3 422	778	0.14%	6.79%
Modèle amélioré	3 824	1 114	0.77%	9.09%
Modèle complet	2 161	470	15.6%	3.66%

10. CONCLUSIONS

Le cas du site 1bis du projet R&D ATTENA a permis de mettre en œuvre une approche géostatistique pour étudier la zone source d'une pollution des sols par des composés organo-halogénés. La zone source correspond aux sols où la somme des teneurs en PCE, TCE et DCE est supérieure à un seuil donné qui, dans la pratique, dépend des propriétés du sol. Plusieurs seuils de 20 à 400 ppm ont été étudiés, le volume de la zone source étant d'autant plus petit et donc difficile à délimiter que le seuil est grand.

Les données de ce cas d'étude proviennent du projet de recherche ATTENA que les partenaires du projet ont aimablement accepté de mettre à la disposition de GeoSiPol pour en faire une étude de démonstration. Sans disposer de données de dépollution pour ce site qui n'a pas encore été réhabilité, les nombreuses données disponibles ont pu être décomposées en deux jeux de données, un restreint pour construire le modèle géostatistique et l'autre de validation pour vérifier les prédictions sur les teneurs et sur la présence de la zone source.

Trois modèles géostatistiques ont finalement été générés, chacun complétant le précédent. Le premier modèle a été établi entièrement avec le jeu de données restreint. Pour le second, les modèles de distribution et de variogramme ont été estimés plus justement en prenant en compte l'ensemble des données disponibles ; en revanche, seul le jeu de données restreint a continué d'être utilisé pour conditionner les simulations des teneurs en composants. Le troisième modèle est le plus complet en intégrant l'ensemble des données comme données de conditionnement. Les deux premiers modèles ont pu être confrontés aux données de validation. Le troisième a servi à étudier l'apport des données complémentaires de validation sur l'estimation des volumes et de l'emplacement de la zone source. Il fournit également un modèle de référence le plus précis possible.

Cette étude peut être perçue comme une étude de démonstration à double titre. D'une part, comme souhaité pour ces travaux de démonstration, elle a permis de confronter des modèles géostatistique à des données de validation pour juger de la justesse et de la précision des prédictions des modèles et pour trouver des explications aux différences observées. D'autre part, elle a été l'occasion d'aborder différents aspects méthodologiques sur l'application de la géostatistique aux sites et sols pollués et sur les conséquences possibles de certains choix méthodologiques sur la qualité des résultats obtenus.

Il peut être retenu de cette étude de démonstration qu'elle a été menée pour apporter des informations et des éléments de réponses sur les aspects suivants.

1. Pertinence de la géostatistique pour estimer le volume et la localisation d'une zone source ou plus généralement de sols pollués (§5).
2. Validation des prédictions des modèles géostatistiques et interprétation des écarts (§6 et 9).
3. Importance des analyses statistiques et variographiques sur la qualité des modèles géostatistiques (§7).
4. Importance des données complémentaires pour réduire les incertitudes spatiales (§8).
5. Complexité des études géostatistiques et façon de les aborder (§4).

L'ensemble des travaux réalisés pour mener à bien l'étude de démonstration a été détaillé au paragraphe §2. Ces travaux sont plus nombreux et variés que ceux qui auraient dû être réellement effectués dans le cadre d'une étude géostatistique habituelle de la pollution du site. Le Tableau 26 résume les étapes de l'étude de démonstration et les replace par rapport à une étude géostatistique classique.

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

Tableau 26 : Principales étapes réalisées pour l'étude de démonstration et pertinence de ces étapes dans le cadre d'une étude géostatistique classique.

Étape	Commentaire
1. Etapes préliminaires 1.1. Compilation des données disponibles 1.2. Sélection des polluants « d'intérêt » pour l'étude	Etapes nécessaires dans lesquelles les données disponibles sont regroupées et mises dans un format qui convient, des données aberrantes peuvent être identifiées pour être corrigées ou écartées, les composants analysés dans les sols et qui sont des polluants potentiels sont identifiés.
2. Réalisation de l'étude géostatistique 2.1. Analyse exploratoire des données 2.2. Modélisation géostatistique des teneurs en polluants 2.3. Analyse des résultats	Etapes habituelles d'une étude géostatistique avec l'analyse exploratoire et statistique des données, nécessaire pour établir les modèles statistiques à introduire dans le modèle géostatistique (distributions, variogrammes), la simulation géostatistique des polluants sélectionnés (simulation conjointe si les polluants sont corrélés), l'analyse des résultats pour s'assurer de la conformité du modèle avec les données et les modèles statistiques établis. Particularités de l'étude de démonstration Il a été présenté dans le détail les conséquences et la prise en compte possible du biais d'échantillonnage (échantillonnage préférentiel de zones polluées) et de la grande proportion de données de teneur sous la limite de détection.
3. Validation de l'étude géostatistique 3.1. Confrontation du modèle à des données complémentaires	Une telle étape peut se produire si des données nouvelles viennent compléter celles utilisées pour générer le modèle géostatistique. Ces données peuvent provenir soit d'une campagne de reconnaissance complémentaire, soit de la dépollution du site. Dans tous les cas, il est souhaitable que l'étude géostatistique réalisée précédemment serve à définir au mieux le nombre et l'emplacement des données complémentaires en fonction des objectifs recherchés (réduction de l'incertitude spatiale, contrôle de la dépollution).
4. Amélioration du modèle géostatistique 4.1. Etablissement de modèles statistiques plus justes à partir de l'ensemble des données 4.2. Simulation géostatistique des teneurs en polluants uniquement à partir des modèles statistiques améliorés 4.3. Analyse des résultats 4.4. Confrontation du modèle géostatistique amélioré aux données de validation 4.5. Comparaison avec le modèle géostatistique de base	Si des données complémentaires sont disponibles, il peut être souhaitable de mettre à jour le modèle géostatistique en intégrant ces nouvelles données, que ce soit pour produire un modèle plus précis après une reconnaissance complémentaire ou pour juger de l'état de pollution d'un site après dépollution. La mise à jour du modèle s'effectue alors en deux temps : 1) vérification et amélioration possible des modèles statistiques à partir de l'ensemble des données disponibles (tâche 4.1 de l'étude de démonstration), 2) génération d'un nouveau modèle géostatistique qui intègre à la fois les modèles statistiques améliorés et l'ensemble des données (tâche 5.1). Particularités de l'étude de démonstration Contrairement à une étude classique, le modèle géostatistique a été amélioré en deux temps lors de l'intégration des données de validation : • dans la tâche 4.2, en ne prenant d'abord en compte que les modèles statistiques améliorés, • dans la tâche 5.1, en rajoutant le conditionnement sur l'ensemble des données disponibles (données initiales et de validation).
5. Modélisation géostatistique complète 5.1. Simulation géostatistique des teneurs en polluants à partir des modèles statistiques améliorés et de l'ensemble des données disponibles pour conditionner les simulations 5.2. Comparaison avec les modèles géostatistiques de base et amélioré simulés avec le jeu de données restreint	De ces deux modèles géostatistiques améliorés, il a pu être étudié la part d'amélioration due aux modèles statistiques corrigés et le gain de précision obtenu en intégrant les données de validation pour conditionner les simulations. Cette façon de procéder se justifie ici dans le cadre de l'étude de démonstration mais n'aurait pas d'intérêt pratique dans une étude géostatistique classique.

Annexes site « ATTENA »

11. ANNEXE A : RESUMES STATISTIQUE DE L'ENSEMBLE DES TENEURS MESUREES DANS LES SOLS

Tableau 27 : Résumés statistiques des composants mesurés moins de 10 fois.

	Mesures	Mesures > LD	Moyenne	Ecart-type	Coeff. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LQ)
Acetone	7	0	10	0	0	10	10	10	10	10	<10
Butanol_1	7	0	10	0	0	10	10	10	10	10	<10
Butanol_iso	7	0	10	0	0	10	10	10	10	10	<10
Butanol_sec	7	0	10	0	0	10	10	10	10	10	<10
Butanol_tert	7	0	10	0	0	10	10	10	10	10	<10
Ethanol	7	0	10	0	0	10	10	10	10	10	<10
MEC	7	0	10	0	0	10	10	10	10	10	<10
MIC	7	0	10	0	0	10	10	10	10	10	<10
Methanol	7	4	44.9	17.47	1.031	10	120	10	23	100	<10
Propanol_1	7	0	10	0	0	10	10	10	10	10	<10
Propanol_2	7	0	10	0	0	10	10	10	10	10	<10

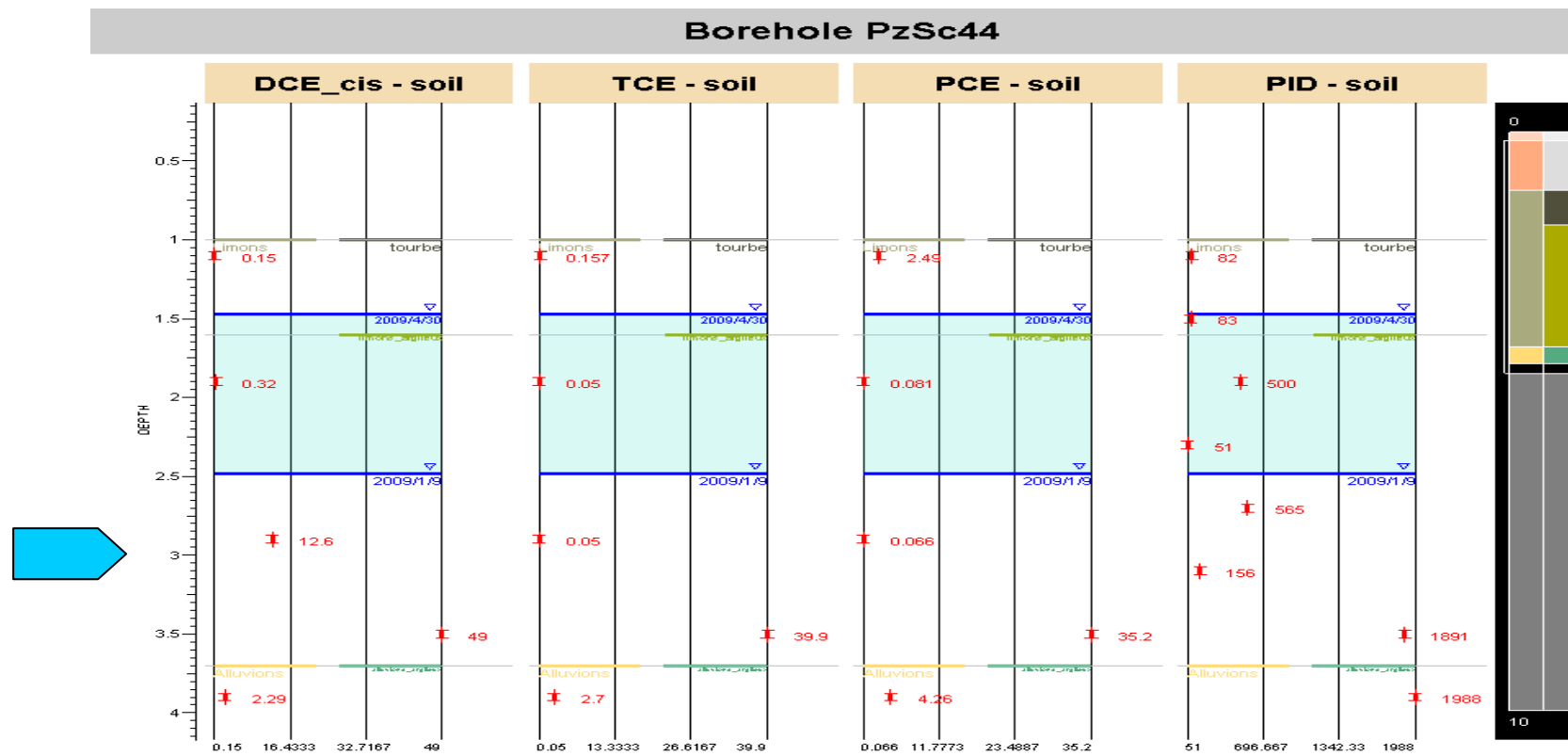
Tableau 28 : Résumés statistiques des composants mesurés entre 10 et 100 fois.

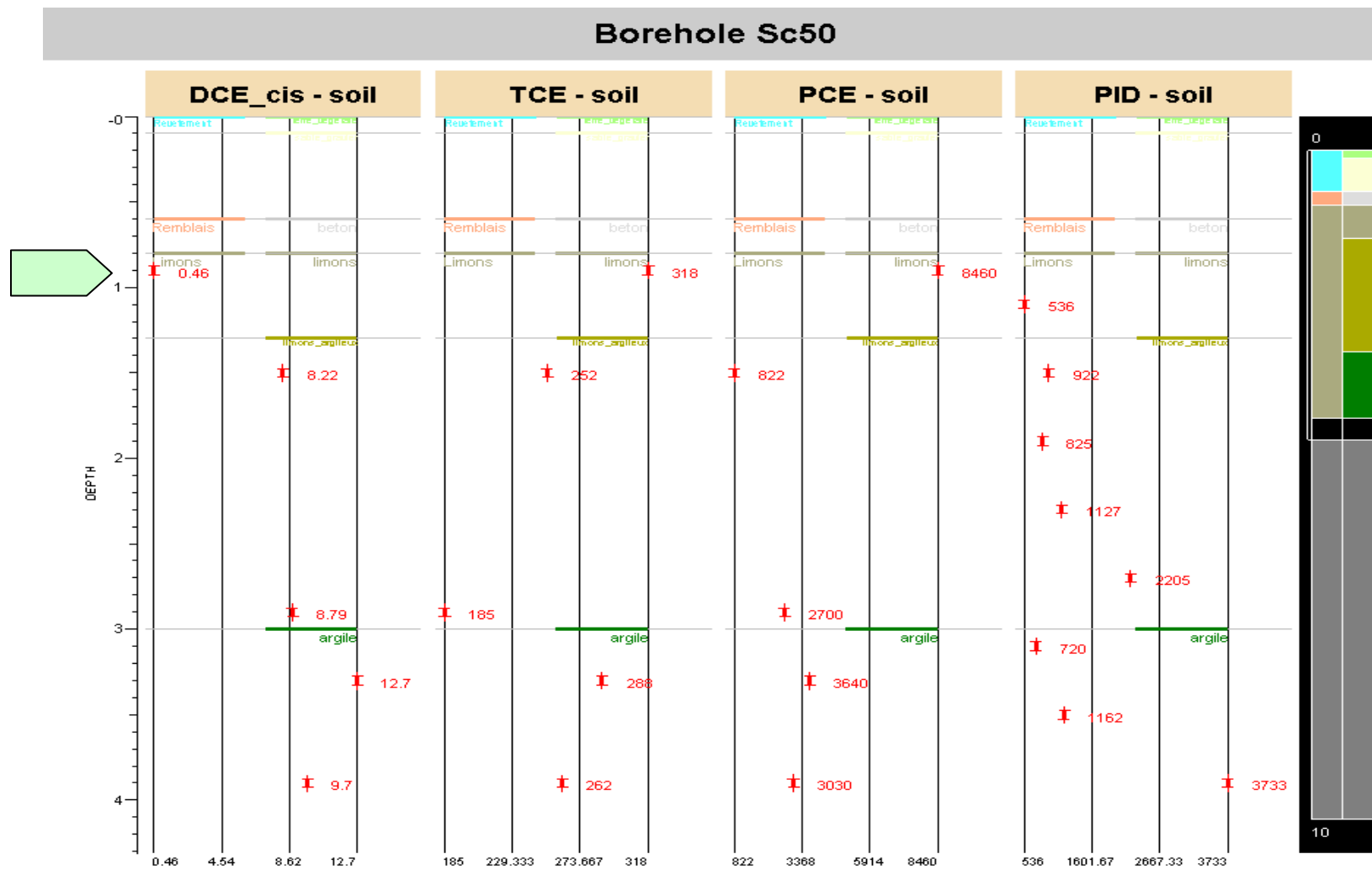
	Mesures	Mesures > LD	Moyenne	Ecart-type	Coeff. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LD)
CB	48	3	0.12	0.014	0.861	0.1	0.79	0.1	0.1	0.1	<0.1
Chlorotoluene_2	48	0	0.1	0	0	0.1	0.1	0.1	0.1	0.1	<0.1
Chlorotoluene_4	48	0	0.1	0	0	0.1	0.1	0.1	0.1	0.1	<0.1
DCB_14	48	17	0.7	0.412	4.073	0.1	19.8	0.1	0.1	0.24	<0.1
DCB_12	49	17	0.88	0.298	2.366	0.01	13	0.1	0.1	0.7	<0.1
DCB_13	49	9	0.15	0.025	1.17	0.01	1.09	0.1	0.1	0.1	<0.1
TCB_123	49	22	3.9	2.902	5.207	0.04	142	0.1	0.1	0.96	<0.1
TCB_124	49	22	30.6	27.93	6.389	0.03	1370	0.1	0.1	1.16	<0.1
HCT_C10_C12	63	52	205	168.3	6.504	0.02	10476	0.26	1.12	5.77	<0.02
HCT_C12_C16	63	52	17.7	10.69	4.785	0.02	630.11	0.26	1.05	4.61	<0.02
HCT_C16_C20	63	52	7.29	1.744	1.899	0.02	63.96	0.58	1.33	5.73	<0.02
HCT_C20_C24	63	52	88.6	38.17	3.422	0.02	1697.4	1.89	4.66	13.9	<0.02
HCT_C24_C28	63	52	286	126.3	3.501	0.02	5506.71	2.95	8.97	22.9	<0.02
HCT_C28_C32	63	52	199	84.62	3.38	0.02	3627.27	3.52	10.5	26.2	<0.02
HCT_C32_C36	63	52	62.7	23.84	3.016	0.02	1025.82	2.34	7.09	21.6	<0.02
HCT_C36_C40	63	52	24.3	8.135	2.663	0.02	360.39	0.72	3.55	10.6	<0.02
HCT_C10_C40	78	52	725	266.5	3.248	15	12300	15	28.3	97.7	<15
MeC5_C10	79	4	4.97	1.531	2.74	2	111	2	2	2	<39
MeC5_C8	79	0	1.07	0.026	0.219	1	2.4	1	1	1	<2.4
MeC8_C10	79	4	3.95	1.544	3.473	1	111	1	1	1	<36.7

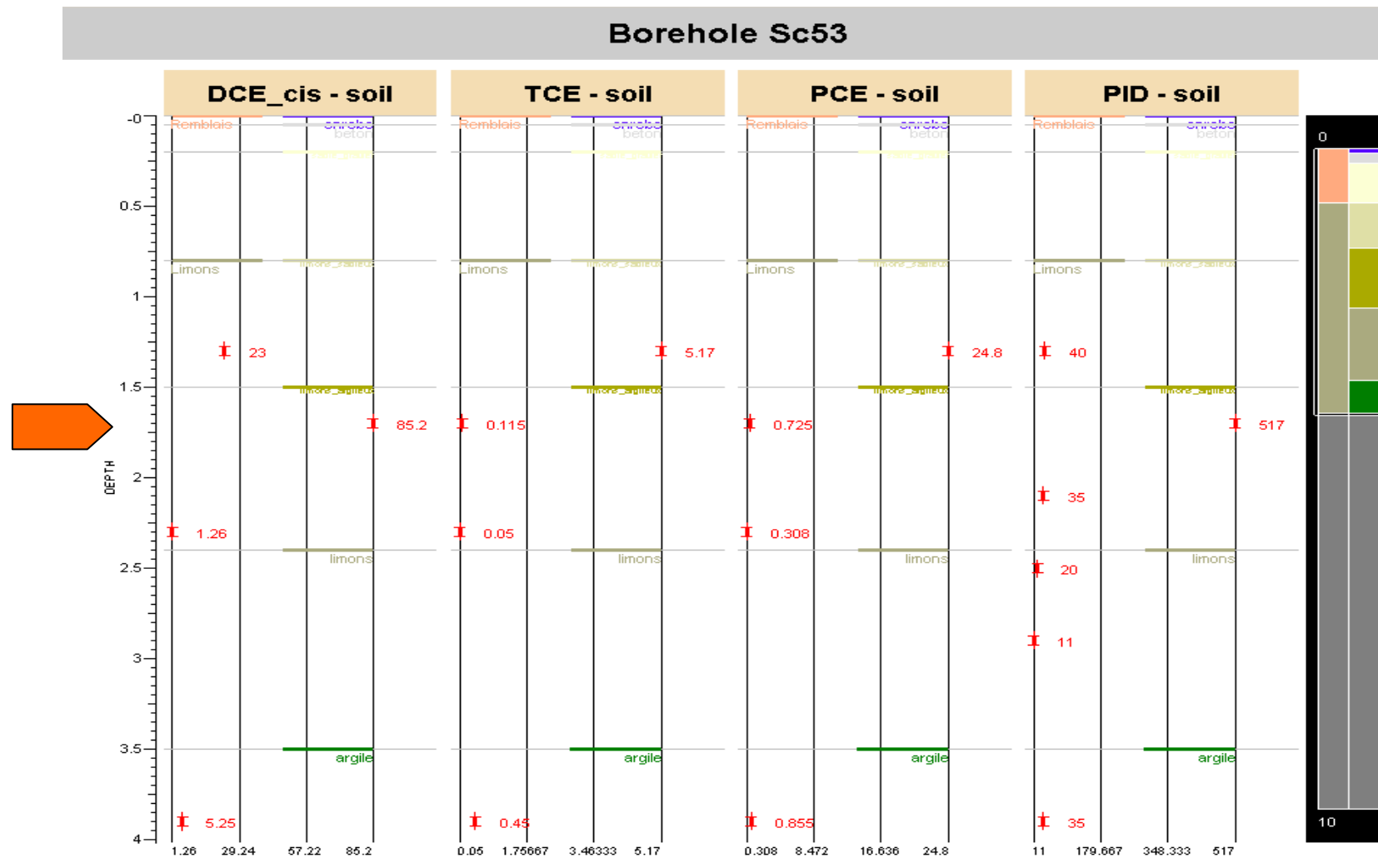
Tableau 29 : Résumés statistiques des composants mesurés plus de 100 fois, classés selon le nombre de leurs mesures au-dessus de la limite de quantification des appareils de mesure.

	Mesures	Mesures > LD	Moyenne	Ecart-type	Coeff. de variation	Minimum	Maximum	25ème Percentile	Médiane	75 ème Percentile	Limite de quantification (LQ)
CA	280	0	2.07	0.016	0.126	2	4.7	2	2	2	<4.7
CM	280	0	2.07	0.016	0.126	2	4.7	2	2	2	<4.7
PCA_1112	280	0	0.1	0	0	0.1	0.1	0.1	0.1	0.1	<0.1
CCl4	282	0	0.05	2E-04	0.066	0.05	0.094	0.05	0.05	0.05	<0.094
CF	282	0	0.1	5E-04	0.078	0.1	0.19	0.1	0.1	0.1	<0.19
DCA_12	282	0	0.05	2E-04	0.066	0.05	0.094	0.05	0.05	0.05	<0.094
TCA_112	282	0	0.21	0.002	0.126	0.2	0.47	0.2	0.2	0.2	<0.47
DCE_11	282	1	0.12	0.016	2.262	0.1	4.52	0.1	0.1	0.1	<0.15
PCA_1122	280	3	0.21	0.003	0.237	0.2	0.7	0.2	0.2	0.2	<0.47
PeCA	280	7	0.83	0.284	5.701	0.2	54.3	0.2	0.2	0.2	<0.47
DCA_11	282	14	0.13	0.022	2.888	0.1	6.3	0.1	0.1	0.1	<0.15
TCA_111	282	16	0.41	0.286	11.64	0.1	80.8	0.1	0.1	0.1	<0.15
DCE_trans	282	20	0.12	0.006	0.856	0.1	1.41	0.1	0.1	0.1	<0.15
HCA	280	22	2.31	1.217	8.823	0.2	297	0.2	0.2	0.2	<0.47
DCM	282	54	0.13	0.013	1.687	0.05	2.03	0.05	0.06	0.1	<2.03
CV	282	56	0.52	0.282	9.162	0.02	78.3	0.02	0.02	0.02	<0.038
DCE_cis	282	118	10.7	5.685	8.949	0.1	1580	0.1	0.1	0.81	<83.9
TCE	282	148	92.5	35.91	6.521	0.05	6960	0.05	0.06	0.65	<0.065
PCE	282	224	1167	519.5	7.473	0.05	122000	0.08	0.99	10.9	<0.051

12. ANNEXE B : LOGS DES SONDAGES OU SE SITUENT LES DONNEES SINGULIERES







13. ANNEXE C : PRESENTATION DE METHODES ET MESURES STATISTIQUES UTILISEES DANS L'ETUDE ATTENA

13.1 Statistiques bidimensionnelles

Le coefficient de corrélation de Pearson mesure l'intensité de la relation linéaire qui existe entre deux variables. Il est calculé par la formule suivante :

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

où ρ_{XY} est le coefficient de corrélation, σ_{XY} la covariance et σ_X^2 et σ_Y^2 les variances pour les variables aléatoires X et Y . Le coefficient de corrélation est toujours compris entre -1 et 1 . Un coefficient de corrélation positif indique que les deux variables sont plus ou moins directement proportionnelles, un coefficient de corrélation négatif signifie que les variables tendent à être inversement proportionnelles.

Dans le cas où les variables ne suivent pas une distribution Gaussienne ou que la relation entre les variables n'est pas linéaire, le coefficient de corrélation de rang de Spearman (r_s) s'avère plus indiqué que celui de Pearson pour étudier les corrélations entre variables.

Le coefficient de corrélation de rang de Spearman est défini par

$$r_s = 1 - 6 \sum_{i=1}^N \frac{d_i^2}{N(N^2 - 1)}$$

où d_i est la différence de rang entre les variables pour l'observation i et N est le nombre d'observations.

Le coefficient de corrélation de Spearman mesure le degré d'association entre les valeurs faibles (respectivement fortes) des deux variables.

13.2 Diagramme quantile-quantile

Les diagrammes quantile-quantile (q-q plots en anglais) permettent de comparer graphiquement deux distributions de données entre elles. Le principe est illustré sur la figure ci-dessous. Il consiste à calculer des quantiles identiques pour les deux distributions et de les tracer sur un diagramme de dispersion avec des axes définis à l'identique. Si les deux distributions sont égales, les points de calcul doivent s'aligner le long de la bissectrice. Les écarts par rapport à la bissectrice peuvent s'interpréter de différentes façons et révéler en particulier des différences de forme entre les distributions.

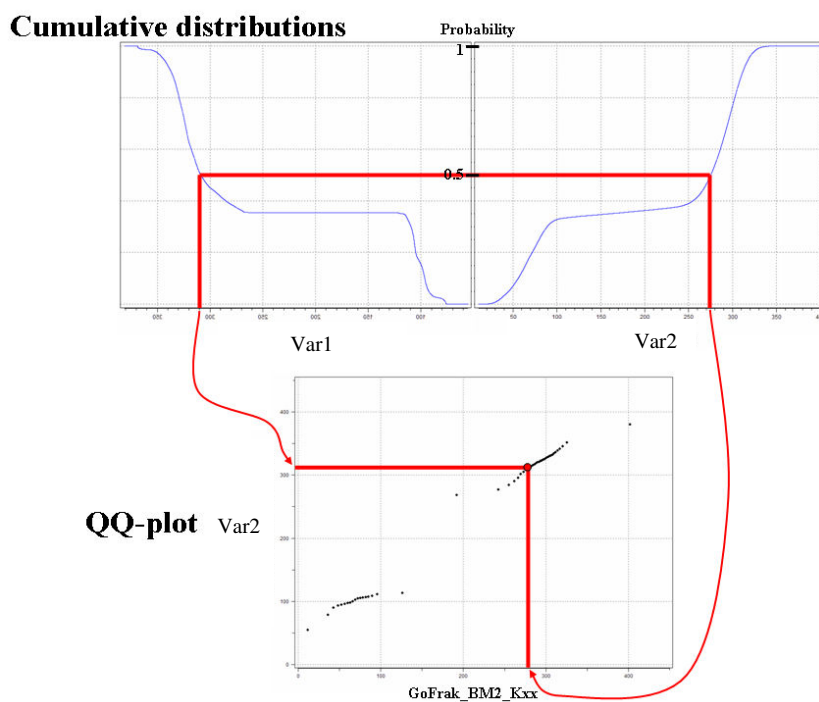


Illustration du principe de calcul d'un q-q plot.

14. ANNEXE D : PRINCIPE DE L’EVALUATION DE LA JUSTESSE ET DE LA PRECISION D’UN MODELE DE SIMULATION GEOSTATISTIQUE

La simulation géostatistique (stochastique) conduit à simuler en tout nœud de la grille une série de valeurs dont la distribution détermine l’incertitude locale au point de simulation. Cette distribution dépend des données disponibles au voisinage du point et de leur répartition.

En effectuant la simulation géostatistique après avoir supprimé une donnée du jeu de données, l’incertitude locale prédite par le modèle peut être estimée à l’endroit de la donnée supprimée et être comparée à la valeur connue de la donnée. En supprimant tour à tour chaque donnée et en répétant l’exercice pour chaque donnée supprimée, il est possible d’évaluer plus complètement la capacité du modèle à reproduire les données. Ce type d’exercice s’apparente à de la validation croisée replacée dans le cadre de la simulation stochastique.

Le modèle sera jugé satisfaisant s’il vérifie les deux conditions suivantes.

1. Les valeurs simulées sont justes, c’est-à-dire qu’elles contiennent et se situent suffisamment près des données.
2. Les valeurs simulées sont précises, c’est-à-dire que l’incertitude sur la prédiction des données est conforme aux incertitudes locales simulées.

La méthode proposée par Deutsch (1996) et reprise par Shmaryan *et al.* (1998) vise ainsi à vérifier la justesse et la précision d’un modèle de simulation géostatistique. La méthode consiste à vérifier la proportion de données qui se trouvent à l’intérieure d’intervalles de confiance de taille croissante déduits des distributions locales de valeurs simulées à l’endroit des données. La simulation géostatistique génère des valeurs et une incertitude satisfaisantes si les proportions de données retrouvées dans les intervalles de confiance sont en rapport avec les tailles de ces intervalles.

Les intervalles de confiance sont déduits de la distribution de valeurs simulées de la façon suivante.

On note :

1. N le nombre de données total.
2. $Z(\mathbf{u}_\alpha)$ la variable aléatoire simulée à l’endroit de la données α .
3. $F_\alpha(z|N-1) = \text{Prob}\{Z(\mathbf{u}_\alpha) \leq z\}$ la distribution locale correspondante estimée à partir des $N-1$ autres données.

L’intervalle de confiance à P % centré sur la distribution $F_\alpha(z|N-1)$ se définit comme l’intervalle $[q_{0.5-p/2}(\alpha); q_{0.5+p/2}(\alpha)]$ où :

- $p = P/100$ est la probabilité associée au pourcentage P ,
- $q_{0.5-p/2}(\alpha)$ est le quantile tel que $F_\alpha(q_{0.5-p/2}(\alpha)|N-1) = \text{Prob}\{Z(\mathbf{u}_\alpha) \leq q_{0.5-p/2}(\alpha)\} = 0.5-p/2$.
- $q_{0.5+p/2}(\alpha)$ est le quantile tel que $F_\alpha(q_{0.5+p/2}(\alpha)|N-1) = \text{Prob}\{Z(\mathbf{u}_\alpha) \leq q_{0.5+p/2}(\alpha)\} = 0.5+p/2$.

La méthode de vérification de la justesse et de la précision d’un modèle de simulation géostatistique peut donc se résumer comme suit.

1. Un nombre suffisant de simulations est effectué en supprimant tour à tour chacune des données comme indiqué précédemment.
2. De ces simulations il est déduit, à l’endroit de chaque donnée α , la distribution locale des valeurs simulées à partir des autres données.
3. De ces distributions locales il est déduit des intervalles de confiance centrés sur les distributions et de tailles croissantes (par ex. 10%, 20%, ..., 90%).

4. Pour chaque taille d'intervalle de confiance (pourcentage P), il est calculé la proportion p^* de données qui tombe à l'intérieur de l'intervalle de confiance à P % qui leur est associé.
5. La comparaison entre intervalles de confiance et données se fait en traçant le diagramme de dispersion de la proportion p^* des données en fonction de la taille de l'intervalle de confiance exprimée en probabilité $p = P/100$.
6. La simulation est juste si les points du graphique sont situés au-dessus de la première bissectrice, c'est-à-dire que les valeurs simulées sont suffisamment proches des données.
7. La simulation est précise si les points sont proches de cette bissectrice, c'est-à-dire que les incertitudes locales sont cohérentes avec les taux de prédiction des données.

Dans le cadre de l'étude de démonstration réalisée sur les données du site ATTENA, ce type de méthode a été utilisé pour évaluer les performances du modèle géostatistique, établi à partir du jeu de données restreint, pour prédire les données de validation. Les simulations ont été générées en utilisant uniquement le jeu de données restreint. Les valeurs simulées aux positions des données de validation ont servi pour calculer les intervalles de confiance qui ont été comparés aux valeurs de ces données.

Références bibliographiques

1. C. V. Deutsch, Direct assessment of local accuracy and precision, Proceedings of the Fifth International Geostatistics Congress, pp 115-125, ed. Baafi and Schofield, Wollongong, Sept. 1996, Kluwer.
2. L. E. Shmaryan, C. V. Deutsch et A. G. Journel, A new Markov model for cokriging and simulation, *SCRF Report N°11 Vol 2*, 1998.

Étude géostatistique de la contamination d'une friche industrielle : le site T.

Fournisseur du cas d'étude :	Arcelor Real Estate France
Réalisation de l'étude :	Mines ParisTech
Contrôle qualité :	GEOVARIANCES
Rédacteurs :	Gaëlle Le Loc'h, Jean-Paul Chilès
Outil logiciel utilisé :	Isatis® (GEOVARIANCES)
Relecteurs :	Claire Fauchoux, Nicolas Jeannée

TABLE DES MATIERES

1. RESUME.....	129
2. INTRODUCTION	129
3. ANALYSE CRITIQUE DES DONNEES.....	130
3.1 EXAMEN DIRECT DES DONNEES	130
3.1.1 Campagnes de mesures	130
3.1.2 Examen des fosses	133
3.1.3 Statistiques	133
3.1.4 Recherche des zones de forte concentration	135
3.1.5 Etude de la classification des types de terrain	136
3.1.6 Etude du support des échantillons	138
4. ANALYSE VARIOGRAPHIQUE ET ESTIMATION.....	143
4.1 ETUDE DES DONNEES PONCTUELLES (HAP).....	143
4.1.1 Variogrammes expérimentaux.	143
4.1.2 Ajustement des variogrammes	144
4.1.3 Validation croisée.	145
4.1.4 Cartographie : estimation par krigeage	146
4.1.5 Changement de coordonnée verticale : utilisation de la profondeur	147
4.1.6 Coordonnée verticale : profondeur. Estimation de bloc	151
4.2 ETUDE DES DONNEES NON PONCTUELLES (HAP)	154
4.3 UTILISATION DE TOUTES LES DONNEES (HAP).....	155
4.3.1 Estimation ponctuelle	155
4.3.2 Utilisation de la profondeur comme coordonnée verticale	156
4.4 ETUDE DU PLOMB ET DU ZINC	159
4.4.1 Variogrammes et ajustement	159
4.4.2 Validation croisée multivariable	161
4.4.3 Cartographie : estimation par cokrigeage	162
5. ETUDE DU LOGARITHME TRANSLATE DE LA CONCENTRATION EN HAP.....	163
5.1 ESTIMATION DU LOGARITHME TRANSLATE ET DU LOGARITHME NATUREL DES HAP.	163
5.1.1 Calcul du logarithme traduit	163
5.1.2 Analyse variographique du logarithme traduit	163
5.1.3 Validation croisée	164
5.1.4 Estimation et calcul de l'intervalle de confiance conventionnel. Dépassement de seuil.	164
5.2 KRIGEAGE LOGNORMAL ? ETUDE DE LA DISTRIBUTION DU LOGARITHME NATUREL DES HAP.....	165
6. ETUDE DU LOGARITHME TRANSLATE DU PLOMB ET DU ZINC	166
7. VALIDATION SUR LES DONNEES NOUVELLEMENT FOURNIES	168
8. CONCLUSION	168

Etude géostatistique de la contamination d'une friche industrielle : le site T.

1. RESUME

L'ancienne cokerie du site T a fait l'objet de plusieurs campagnes d'échantillonnage dans le but de délimiter et quantifier la contamination du sol. La présente étude reprend les données de deux campagnes, LECES et GEOSTOCK. L'examen des données met en évidence une difficulté majeure pour l'étude de la pollution : alors que le sol est extrêmement hétérogène, aucune information n'a été transmise concernant les méthodes d'échantillonnage et le volume des échantillons prélevés. Or ces informations sont fondamentales car dans un milieu hétérogène comme celui-ci de nombreuses décisions méthodologiques dépendent des méthodes d'échantillonnage et du volume des échantillons.

En l'absence de ces informations, nous nous sommes attachés à présenter à chaque étape les divers choix qui s'offraient à nous et avons investigué les conséquences de chacun d'eux.

In fine, les ajustements de modèles effectués sur ce jeu de donnée étant insatisfaisants, nous avons choisi de les utiliser uniquement dans un but de présentation méthodologique illustrant les diverses étapes d'une estimation des concentrations : étude critique des données, calculs des variogrammes expérimentaux, ajustement des modèles, estimation, calcul des intervalles de confiance conventionnels. Nous avons également montré comment il était possible de s'aider de variables auxiliaires, soit mesurées conjointement pour les variables Plomb + Zinc, soit obtenues par calcul à partir de la variable d'intérêt dans le cas des HAP : variable brute, logarithme translaté, indicatrice du quantile 50°%.

Par contre, les cartes de probabilités de dépassement de seuil qui auraient pu permettre la délimitation des zones contaminées nécessitent l'utilisation de méthodes très sensibles à la qualité de la modélisation (simulations) et n'ont donc pas été calculées.

L'intérêt premier de ce site pour une étude de démonstration n'est malgré tout pas de montrer les méthodes géostatistiques à utiliser, mais de permettre d'expliquer en quoi les informations sur les méthodes d'échantillonnage sont fondamentales pour une étude de géostatistique de qualité.

Les données supplémentaires que nous espérions pour valider la méthodologie nous ont été transmises tardivement. La mauvaise qualité de la modélisation étant déjà connue, les résultats d'une validation, qu'elle soit favorable ou non, ne peuvent pas être significatifs et cette validation n'a donc pas été effectuée.

2. INTRODUCTION

Le site T est une ancienne cokerie, démantelée en 1973 et 1974.

Plusieurs campagnes de mesures ont été effectuées en vue de l'analyse de la pollution des sols. Pour cette étude de démonstration, nous disposons des résultats des campagnes effectuées entre 1991 et 1995. Lors de ces campagnes, 166 échantillons ont été prélevés dans 97 fosses. D'autres données seraient disponibles. Elles nous ont été transmises (tardivement) pour une possible validation mais n'ont pas été utilisées : une validation n'a de sens que lorsqu'on est raisonnablement confiant dans la qualité de la modélisation. Nous verrons que ce n'est pas le cas pour ce site.

Les analyses ont montré la présence d'Hydrocarbures Aromatiques Polycycliques (HAP) et de métaux lourds (Plomb et Zinc en particulier).

Ce site a déjà fait l'objet de plusieurs études géostatistiques [M. Garcia 1996], [F. Balabdaoui et al 1999]

La présente étude comporte les étapes suivantes :

- analyse critique des données
- analyse variographique et modélisation
- estimation, calculs des intervalles de confiance conventionnels

Elle aurait dû comporter aussi une étape de simulations et calculs de probabilités de dépassement de seuils. Mais ces calculs sont très sensibles aux paramètres de la modélisation. Comme nous verrons que ces derniers sont sujets à caution, cette étape n'a pas été effectuée.

3. ANALYSE CRITIQUE DES DONNEES

Pour cette étude de démonstration, nous nous limitons à l'étude des 3 premiers mètres du sol et sous-sol : il y a bien quelques échantillons (17) entre les profondeurs 3 m et 5 m, mais pas suffisamment pour espérer obtenir une modélisation et une cartographie de qualité au-delà de 3 m. S'il est toujours numériquement possible d'extrapoler, cette pratique est à éviter au maximum : même avec une bonne modélisation des 3 m de surface, rien ne dit que le modèle obtenu reste valable dans la zone située immédiatement au dessous : la nature des terrains reste-t-elle la même ? L'infiltration d'éventuels contaminants s'est-elle effectuée de la même façon ? D'autre part, même avec un comportement qui reste identique, l'incertitude sur l'estimation devient beaucoup plus élevée lorsque les données utilisables diminuent aussi fortement.

3.1 Examen direct des données

L'examen direct des données a consisté en l'étude :

- de la localisation des échantillons, en tenant compte de l'historique des mesures,
- de la comparaison des campagnes de mesure,
- des premières statistiques sur les données,
- des zones de forte concentration (recherche de ces zones),
- de la classification des types de terrain,
- du support des mesures (taille et forme des échantillons prélevés).

3.1.1 Campagnes de mesures

Les données dont nous disposons proviennent de deux campagnes de mesure, prélevées et analysées par des organismes différents. La plus ancienne est la campagne LECES, et la plus récente la campagne GEOSTOCK. La Figure 2 montre l'implantation des fosses de chacune des campagnes.

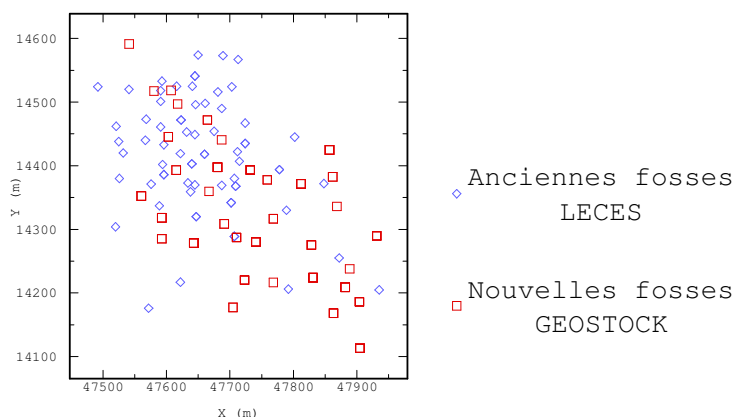


Figure 1. Carte d'implantation des fosses. Bleu : fosses anciennes LECES, rouge : fosses récentes GEOSTOCK.

La plupart des fosses GEOSTOCK ont été implantées dans la zone sud-est où les fosses LECES étaient peu nombreuses. Par contre, il y a peu de fosses GEOSTOCK dans la zone où les fosses LECES sont densément implantées. Cela complique la comparaison des campagnes de mesures : si une différence statistique est constatée, s'expliquera-t-elle par une implantation différente ou par des procédés de prélèvements et d'analyses différents ? Ou encore par des fluctuations statistiques normales quand on étudie des échantillons peu nombreux (peu d'échantillons LECES dans la zone sud-est, peu d'échantillons GEOSTOCK dans la zone nord-ouest) ? Enfin, elle pourrait tout simplement être due à des niveaux de pollution différents.

Néanmoins, l'examen des histogrammes des concentrations en HAP, plomb et zinc ne montre pas de différences manifestes entre les deux campagnes (Figure 3). Ces histogrammes laissent donc penser que les données des deux campagnes pourront être mélangées. Cela devra être vérifié ultérieurement.

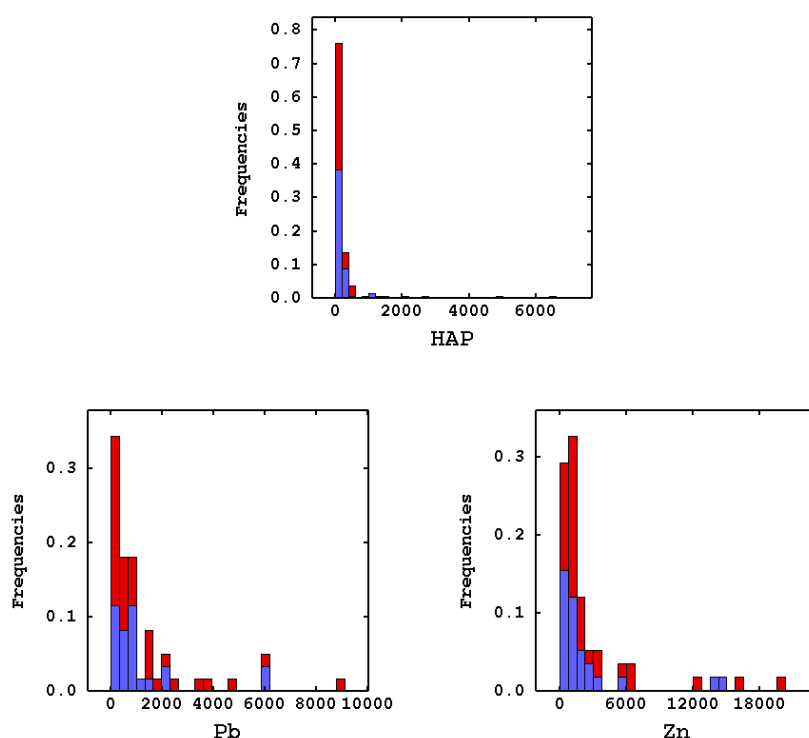


Figure 2. Histogrammes des concentrations en HAP (gauche), plomb (centre) et zinc (droite). Bleu : fosses anciennes LECES, rouge : fosses récentes GEOSTOCK. Unités : ppm

Les cartes d'implantation des mesures des différents polluants (Figure 3) montrent une répartition du même type, mesures plus resserrées dans la zone nord-ouest et plus éparées dans la zone sud-ouest. Par contre, l'échantillonnage est visiblement moins dense pour les métaux lourds.

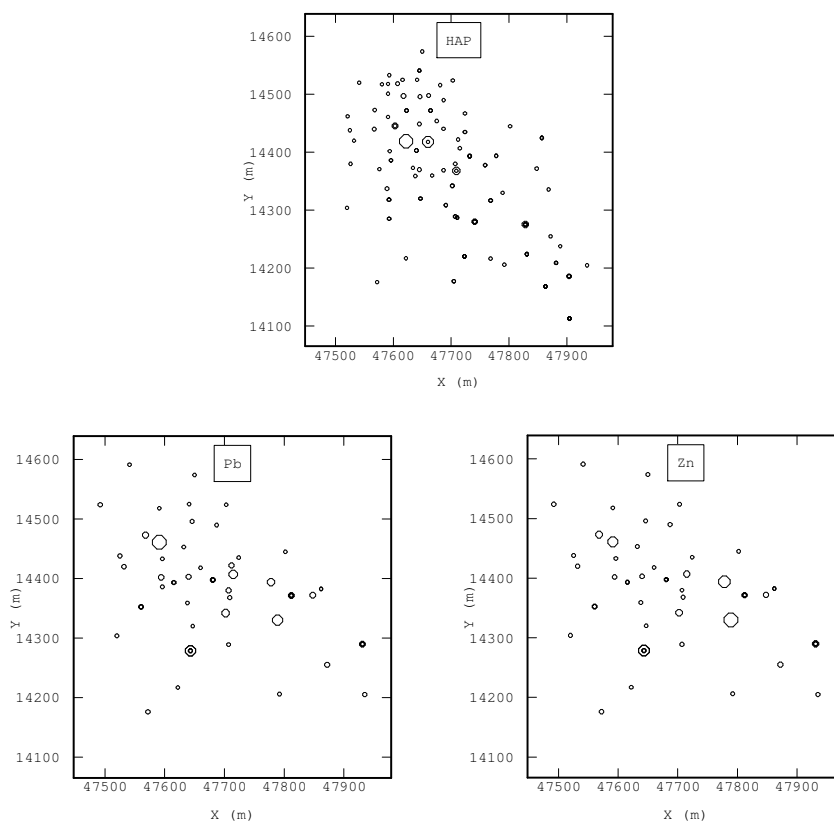


Figure 3. Cartes d'implantation des échantillons par type de polluant : haut HAP, gauche plomb, droite zinc

3.1.2 Examen des fosses

L'examen des fosses creusées en vue de prélèvements a montré que le sous-sol de la zone d'étude est extrêmement hétérogène : présence de plusieurs types de graviers plus ou moins grossiers, de goudrons en couche, d'alluvions. On peut noter la présence de murs et morceaux de bétons, qui peuvent éventuellement constituer une frontière brutale entre les types de terrain (Figure 4).

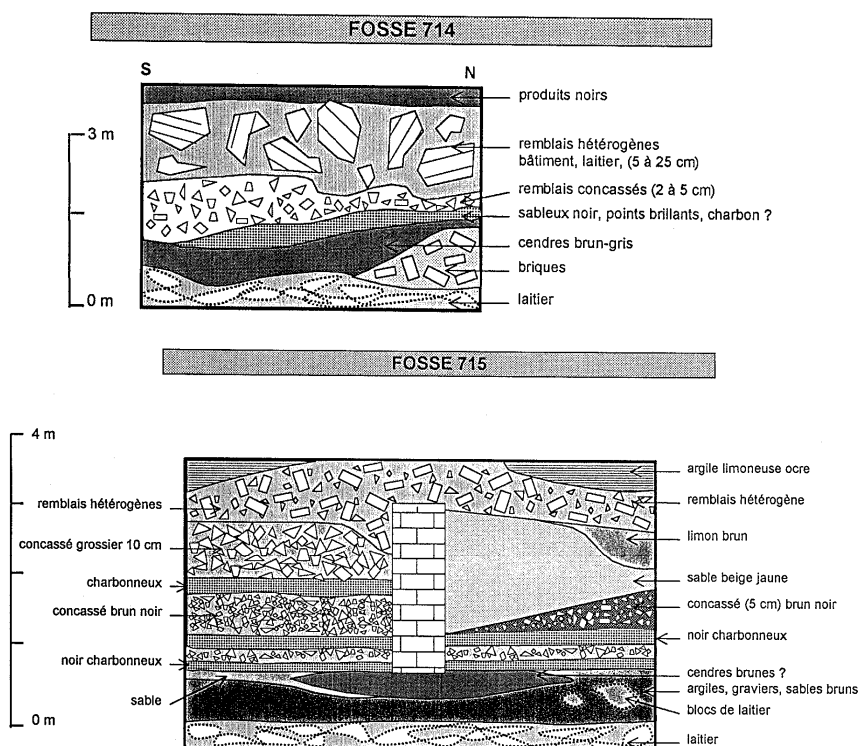


Figure 4. Exemples de fosses hétérogènes (D'après Thiry et al. 2000)

3.1.3 Statistiques

Les statistiques de base (nombre d'échantillons, minimum, maximum, moyenne, variance, écart type et écart type relatif, ainsi que les corrélations entre les variables d'étude) ont tout d'abord été calculées sur l'ensemble des échantillons (Tableaux 1 et 2).

Tableau 1. Statistiques de base sur les 166 échantillons bruts (maximum, moyenne et écart-type arrondis à l'entier le plus proche).

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	138	0,10	6 500	243	761	3,25
Pb	61	19,00	8 913	1 230	1 748	1,42
Zn	58	49,00	20 000	2 752	4 231	1,54

Les seuils de contamination sont de 200ppm pour les HAP, 1 300ppm pour le plomb, 3 000ppm pour le zinc. 76% des mesures sont en dessous de ce seuil pour les HAP, 72% pour le plomb et 79% pour le zinc. Par contre, les valeurs maximales sont très supérieures aux seuils de contamination : de l'ordre de 30 fois plus élevées pour les HAP, 7 fois plus élevées pour le plomb et pour le zinc. Cela semble indiquer une contamination forte mais localisée.

La moyenne globale est proche du seuil de contamination pour les trois polluants étudiés.

Comme il avait été vu sur les cartes d'implantation, les analyses n'ont pas été effectuées sur tous les échantillons pour les différents polluants. 28 échantillons n'ont pas de mesure de HAP, 105 n'ont pas de mesure de plomb et 108 n'ont pas de mesure de zinc. En particulier, bien que le nombre de mesures de plomb et de zinc soit bien plus faible que le nombre de mesures de HAP, seuls 33 échantillons ont été analysés à la fois pour les HAP et le plomb, dont 30 seulement pour les trois contaminants, HAP, plomb et zinc. Par contre, tous les échantillons analysés pour le zinc l'ont également été pour le plomb (3 échantillons analysés pour le plomb ne l'ont pas été pour le zinc).

Tableau 2. Matrice de corrélation (166 échantillons bruts). Entre parenthèses, nombre d'échantillons communs aux deux polluants, ayant permis le calcul du coefficient de corrélation.

	HAP	Pb	Zn
HAP	1,00	-0,04 (33)	-0,08 (30)
Pb	-0,04 (33)	1,00	0,89 (58)
Zn	-0,08 (30)	0,89 (58)	1,00

Le coefficient de corrélation entre plomb et zinc est élevé, 0,89. Comme le montre le nuage de corrélation (Figure 5), cela signifie que les concentrations du plomb et du zinc évoluent conjointement. L'apport de ces deux contaminants s'est probablement fait conjointement (ce qui est fréquent pour le plomb et le zinc). Par contre, les HAP ne sont quasiment pas corrélés aux métaux lourds (coefficients de -0,04 et -0,08), comme le montre les nuages de corrélation entre métaux lourds et HAP (Figure 6). Les contaminations sont probablement d'origines différentes ou ont été lessivées par des processus différents.

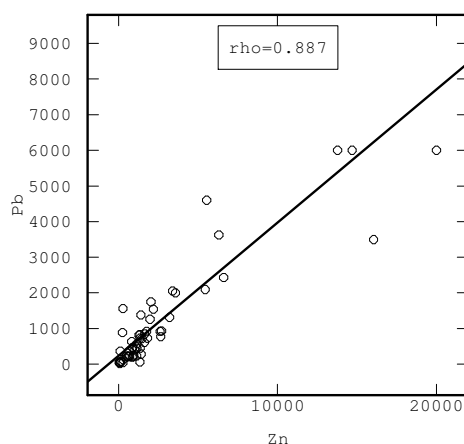


Figure 5. Nuage de corrélation plomb-zinc et droite de régression linéaire.

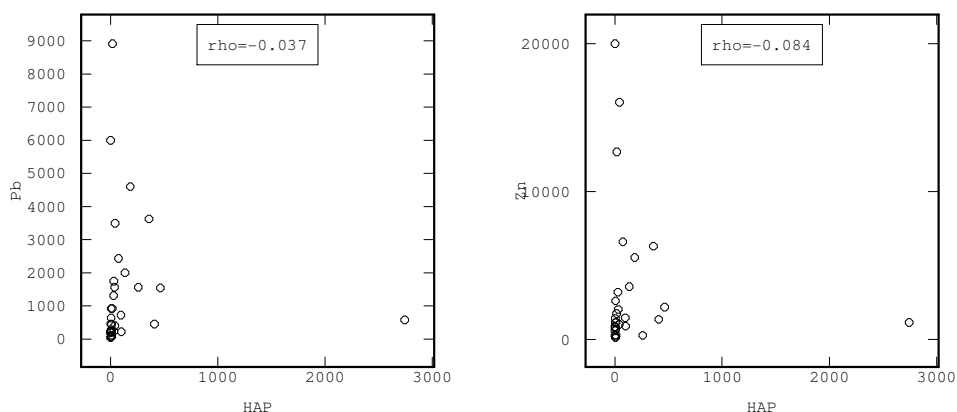


Figure 6. Nuages de corrélation plomb-HAP (gauche) et zinc-HAP (droit).

3.1.4 Recherche des zones de forte concentration

La figure 7 montre, pour chacun des polluants, la carte des échantillons en projection sur un plan horizontal (haut) et en projection sur une section verticale orientée est-ouest. Les échantillons dont la mesure est supérieure au seuil de contamination sont représentés par des carrés rouges. La taille des symboles est proportionnelle à la valeur mesurée.

On voit sur cette figure que, comme le laissaient prévoir les coefficients et nuages de corrélation, les échantillons contaminés sont approximativement les mêmes pour le plomb et le zinc, les contaminations possédant les mêmes caractéristiques statistiques. En particulier, on constate une contamination de surface importante. On retrouve également quelques échantillons très contaminés à une profondeur d'environ 3 m. Entre ces deux zones, il n'y a pas d'échantillon contaminé en zinc, et seulement trois échantillons contaminés en plomb mais montrant une concentration proche du seuil de contamination. Sur la projection horizontale, on voit que les échantillons de concentration très élevée sont assez disséminés.

Par contre, le comportement de la contamination en HAP est assez différent. Il y a une contamination nette mais relativement modérée dans la zone centrale, mais les très fortes contaminations sont regroupées autour de la zone $x = 14\ 400\text{ m}$, $y = 47\ 600\text{ m}$. La projection sur la section verticale montre que la contamination est présente partout dans les 3 premiers mètres de profondeur

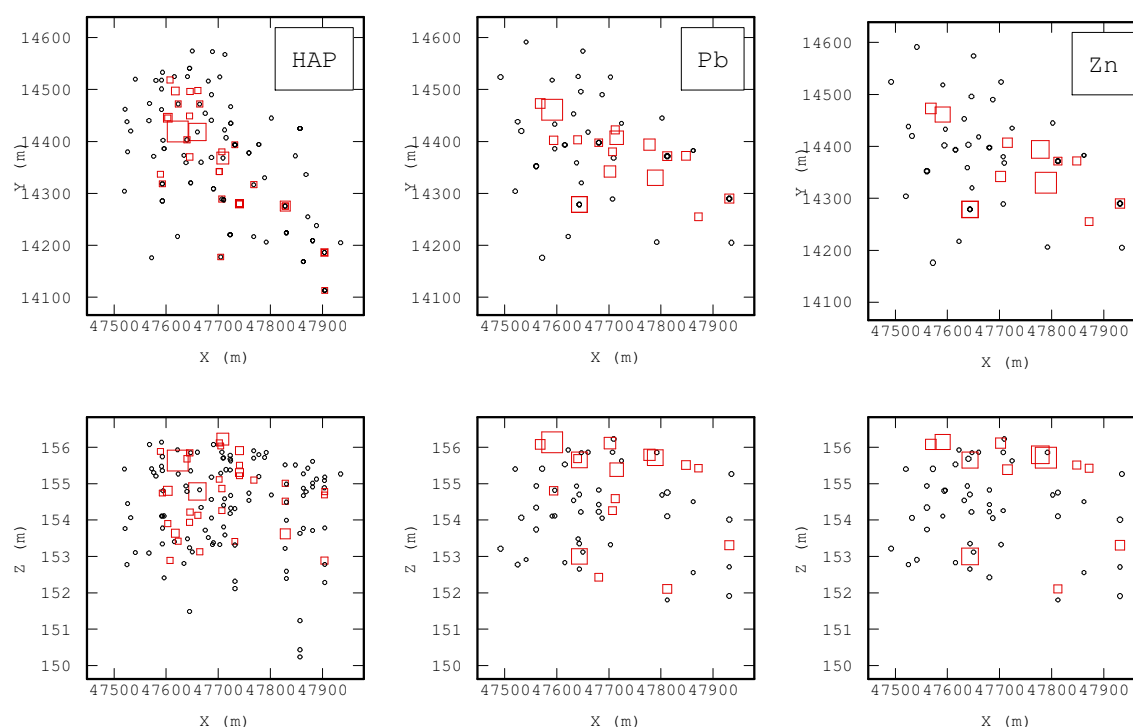


Figure 7. Cartographie des échantillons pour les différents polluants. Haut : projection sur un plan horizontal, bas projection sur une section verticale est-ouest. Gauche HAP, centre plomb, droite zinc. Rouge : échantillons dépassant le seuil de contamination.

Il sera donc préférable d'étudier séparément les concentrations en HAP d'un côté, en plomb et zinc de l'autre. De plus, il est sans doute inutile d'essayer de modéliser un lien statistique et spatial entre ces deux types de contamination.

3.1.5 Etude de la classification des types de terrain

Sept types de terrains ont été décrits sur les échantillons :

- remblai argileux,
- remblai à granulométrie fine,
- remblai divers + blocs,
- remblai de laitier granulé,
- remblai de laitier induré,
- alluvions,
- goudrons.

La figure 8 représente la carte des différents types de remblai (au centre), en projection sur un plan horizontal (haut) ou sur une section verticale est-ouest (bas). Les images de gauche représentent l'implantation des fortes concentrations en HAP, pour comparaison. Comme il est prévisible, la plus forte concentration en HAP correspond à l'unique échantillon de type goudron (sur lequel les métaux lourds n'ont pas été analysés). L'image de droite représente la concentration en HAP en fonction de la profondeur. On voit que le seuil de contamination (trait rouge) est dépassé à toutes les profondeurs jusqu'à 3 m. En dessous de cette profondeur il n'est plus dépassé, mais très peu d'échantillons ont été analysés (seules sept fosses ont atteint les alluvions). Cette disparition de la contamination aux profondeurs supérieures à 3m est-elle réelle, ou n'est-ce qu'un artéfact lié à un échantillonnage insuffisant ? Il ne nous est pas possible de trancher ici. Il est possible que la contamination soit effectivement limitée aux trois premiers mètres, mais le nombre d'échantillons réduits ne permet pas de l'affirmer avec certitude.

Le terrain « remblai divers + blocs » amène à se poser d'autres questions : dans ce type de terrain, comment les échantillons ont-ils été prélevés ? Les prélèvements contiennent-ils réellement des blocs, ou ont-ils été prélevés à côté de ceux-ci, dans la fraction fine du terrain ? Et si les blocs ont été prélevés, ont-ils été ensuite broyés avec le reste de l'échantillon pour homogénéisation, ou l'analyse proprement-dite n'a-t-elle porté que sur la fraction fine ? Nous verrons que la réponse à cette question est essentielle pour déterminer le vrai support des échantillons. Etant donné l'extrême hétérogénéité des fosses, les échantillons n'ont-ils pas été choisis aux endroits où l'échantillonnage semblait plus facile ? Si c'est le cas, cela pourrait remettre en cause la représentativité statistique de l'échantillonnage.

Augmenter le support en agglomérant les fosses lisserait probablement les mesures, mais au prix d'une incertitude encore plus grande sur la localisation des échantillons et d'une diminution de leur nombre. De plus, l'estimation ne pourrait se faire de façon raisonnable que sur des très grands blocs, au moins aussi grands que la distance séparant les fosses regroupées. Si le but est de cartographier des blocs suffisamment petit pour guider le processus de dépollution, ça ne permettrait pas d'y répondre.

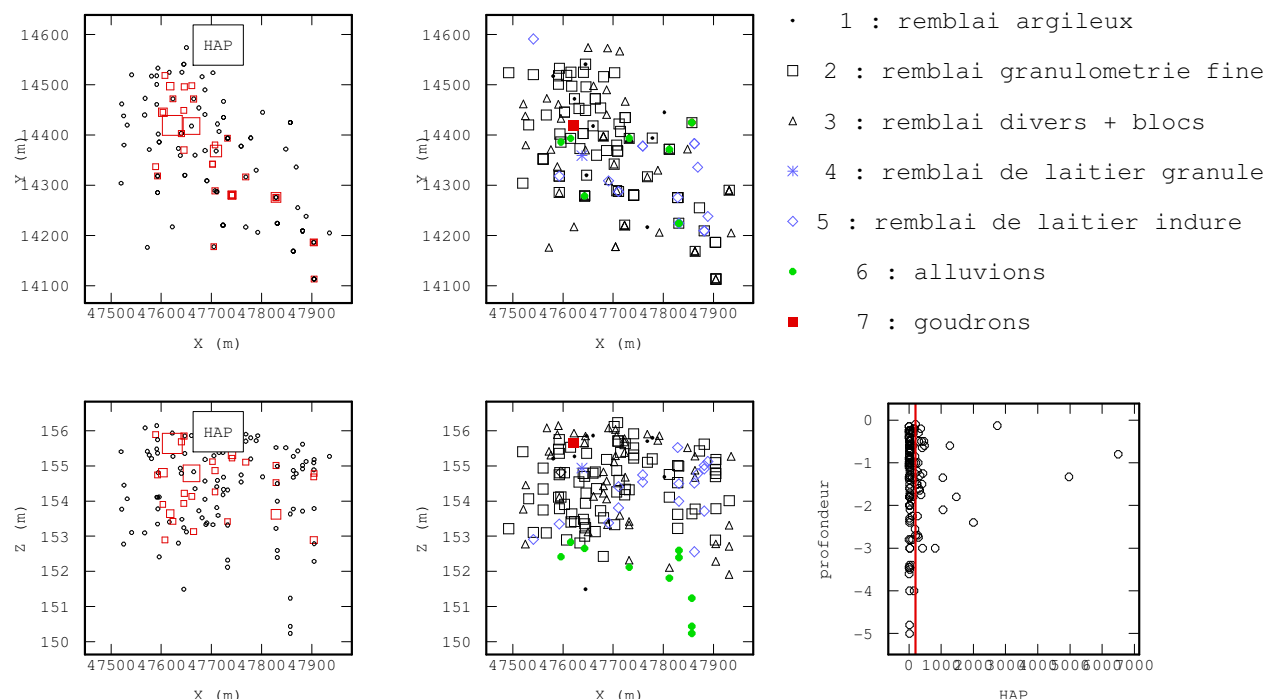


Figure 8. Cartes des concentrations en HAP (gauche) et cartes des divers types de terrain (centre). Haut : projection sur un plan horizontal ; bas : projection sur un plan vertical est-ouest. Droite : représentation de la concentration en HAP en fonction de la profondeur.

La figure 9 montre le lien entre terrain et profondeur. Les alluvions (losanges verts) se trouvent en profondeur comme il est normal. Les remblais argileux (points noirs), assez peu présents, sont plutôt proches de la surface même si on en trouve à près de 3 m de profondeur. Les autres terrains présents en quantité notable sont présents dans toute la hauteur bien échantillonnée.

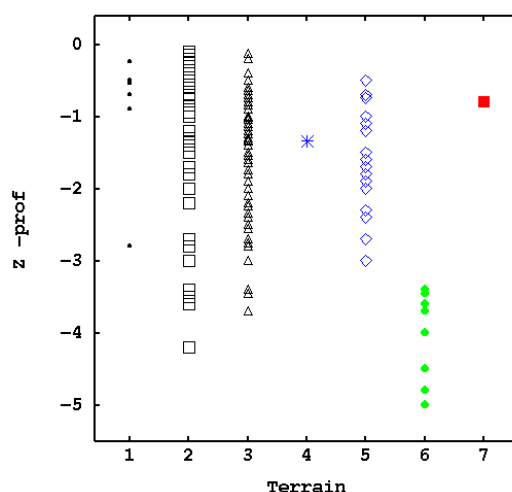


Figure 9. Nuage de corrélation terrains – profondeur

La figure 10 montre la fréquence de prélèvement dans les différents types de terrain, et la relation entre ces terrains et les échantillons dépassant le seuil de contamination (rouge). A part le goudron pour les HAP, la contamination est présente essentiellement dans les remblais, et surtout dans les types 2 et 3 plus largement échantillonnés. L'absence de contamination dans les terrains de type 4, 5 et 6 (remblai de laitier granulé ou induré, alluvions) est-elle significative ou simplement apparente à cause d'un plus faible échantillonnage dans ces terrains ?

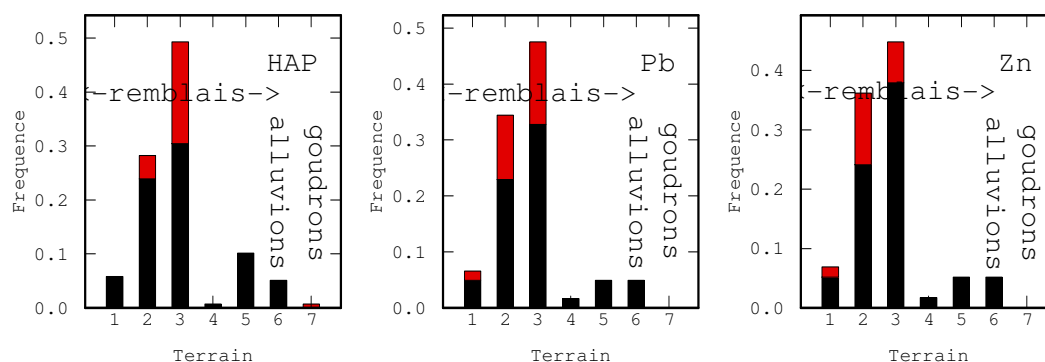


Figure 10. Fréquences de prélèvement dans les différents types de terrain. Rouge : concentration au-dessus du seuil de contamination pour chacun des types de polluant (gauche : HAP ; centre : plomb ; droite : zinc).

3.1.6 Etude du support des échantillons

Le support d'un échantillon est le volume qui a été prélevé et analysé. Pour caractériser complètement le support, il est nécessaire de décrire non seulement la mesure de ce volume, mais aussi sa forme et son orientation. Si on peut souvent s'abstenir d'étudier l'orientation (si celle-ci reste la même pour tous les échantillons par exemple), forme et volume influent directement sur les propriétés géostatistiques du jeu de données, en particulier sur le variogramme : des cubes de 10 cm et de 30 cm de côté n'auront pas le même variogramme, un cube de 30 cm de côté n'aura pas le même variogramme qu'un parallélépipède de base carrée d'1 dm² et de hauteur 90 cm, alors qu'ils ont le même volume. Des échantillons de support trop différent ne peuvent donc pas être regroupés pour le calcul du variogramme expérimental. Avant de commencer l'analyse géostatistique des données, on doit donc se poser la question du support des échantillons.

Sur le site T, deux codes d'échantillonnage coexistent : les données « ponctuelles », pour lesquelles nous n'avons aucune information sur le support, et des données « non ponctuelles », pour lesquelles nous disposons d'une profondeur minimum et maximum. Nous pouvons donc calculer la longueur théorique de l'échantillon.

Les données ponctuelles représentent 51% des échantillons. Elles n'existent que dans la campagne GEOSTOCK, et représentent 87,6% de celle-ci. Les données non ponctuelles sont présentes sur les deux campagnes.

On voit sur la figure 11 que la partie sud-est est plutôt échantillonnée par des données ponctuelles et la partie nord-ouest par des données non ponctuelles.

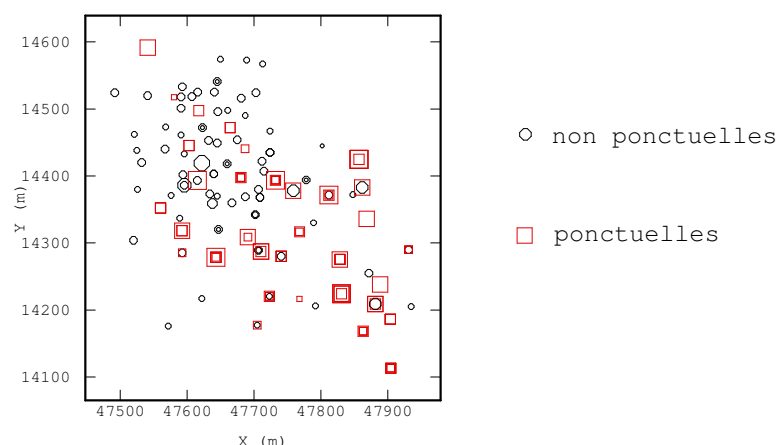


Figure 11. Carte des données de type « ponctuelles » (carrés rouges) et « non ponctuelles » (cercles noirs).

La figure 12 montre l'histogramme des longueurs théoriques des échantillons non ponctuels. On voit que les longueurs sont assez également distribuées entre 10 cm et 2,5 m. Mais que représente réellement cette longueur théorique sur un échantillon de 2,5 m ? Les profondeurs minimum et maximum représentent-elles réellement les limites de l'échantillon prélevé (qui devra ensuite être broyé dans sa totalité pour homogénéisation avant analyse) ou simplement les limites d'une zone dans laquelle un échantillon plus petit a été prélevé ?

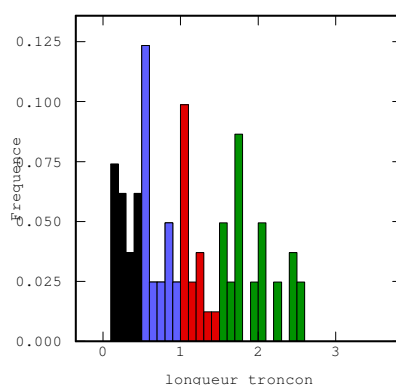


Figure 12. Histogramme de la longueur du tronçon pour les 81 échantillons de longueur connue. Noir : longueur entre 0,1 m et 0,5 m ; bleu : longueur entre 0,5 m et 1,0 m ; rouge : longueur entre 1,0 m et 1,5 m ; vert longueur supérieure à 1,5 m.

Le cas le plus favorable pour l'étude géostatistique serait celui d'échantillons de support inconnu mais identique partout, prélevé au milieu du minimum et maximum fourni pour l'échantillon. Cela nous permettrait de regrouper tous les échantillons. Par contre, si ces longueurs théoriques correspondent aux longueurs réelles, il ne sera pas possible d'effectuer ce regroupement. Il faudrait auparavant effectuer une régularisation (création d'échantillons fictifs identiques, par regroupement et subdivision d'échantillons réels). Malheureusement le choix de la longueur de ces échantillons régularisés, en l'absence d'informations sur la raison de la variabilité des longueurs théoriques et sur l'homogénéité des échantillons, ne peut se faire que sur la plus grande longueur, de l'ordre de 2 à 3 m, et le nombre d'échantillons régularisés serait alors trop faible pour calculer des variogrammes expérimentaux. De plus, il ne nous serait alors pas possible de regrouper les données non ponctuelles et les données ponctuelles qui ne peuvent pas être régularisées.

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

L'extrême hétérogénéité vue dans les fosses rend peut probable le prélèvement d'un échantillon de hauteur totale de l'ordre de 2,50m, mais sans information sur les campagnes et méthode d'échantillonnage, nous ne pouvons pas décider entre l'un et l'autre cas.

L'examen statistique des données permet-il de trancher ? Et nous permet-il au moins de décider avec quelle classe de longueur nous pouvons mélanger les données ponctuelles ?

Quand la différence entre deux populations est simplement due au support des échantillons alors que le phénomène sous-jacent et les méthodes d'analyses sont les mêmes, la variance (et le palier du variogramme) diminue quand le support augmente. Retrouvons-nous ce phénomène ici ? Nous avons recalculé les statistiques pour 4 classes de longueur théorique, moins de 50cm, entre 50cm et 1m, entre 1m et 1,50m, plus de 1,5m. Ces classes correspondent aux classes de couleur de la figure 12. Les tableaux 3 à 6 montrent ces statistiques. Nous nous intéresserons essentiellement à l'écart-type.

Tableau 3. Statistiques de base sur les 19 échantillons de longueur comprise entre 0,1 m et 0,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	18	0,30	6 500,00	682,96	1 547,78	2,27
Pb	6	45,00	3 625,00	1 152,00	1 275,49	1,11
Zn	6	61,00	6312,00	2 094,83	2 189,93	1,05

Tableau 4. Statistiques de base sur les 20 échantillons de longueur comprise entre 0,5 m et 1,0 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	18	0,20	431,00	57,22	99,04	1,73
Pb	12	40,00	8 913,00	1 775,17	2 684,98	1,51
Zn	12	272,00	20 000,00	4 073,17	5 904,67	1,45

Tableau 5. Statistiques de base sur les 19 échantillons de longueur comprise entre 1,0 m et 1,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	18	0,30	465,00	93,46	143,78	1,54
Pb	11	79,00	3 495,00	1 079,09	901,89	0,84
Zn	10	156,00	1 6050,00	3 188,80	4 380,51	1,37

Tableau 6. Statistiques de base sur les 22 échantillons de longueur supérieure à 1,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	19	4,00	4 969,00	331,79	1 097,97	3,31
Pb	10	77,00	4 600,00	892,00	1 326,93	1,49
Zn	8	125,00	5 544,00	1 597,75	1 594,44	1,37

On voit donc que les écarts-type, qui devraient diminuer en fonction de la longueur de l'échantillon, ne suivent pas la règle attendue. En regardant plus attentivement ces tableaux, on constate que l'ordre des écarts-types suit exactement l'ordre des valeurs maximales. Sur des faibles populations comme celles-ci (entre 19 et 22 échantillons par classe) et quand les maxima sont aussi extrêmes que dans notre cas, le calcul de l'écart-type est essentiellement contraint par les maxima.

Nous avons tenté d'atténuer cet effet en calculant les statistiques du logarithme de la concentration en HAP, ce qui a pour effet de diminuer l'influence des hautes teneurs (tableaux 7 à 10). Malheureusement, si cet effet est atténué il se maintient pour l'essentiel, les valeurs d'écart type suivent encore les valeurs des maxima, même si le lien est un peu moins strict.

Ce calcul des logarithmes n'a pas été effectué sur les métaux lourds : en raison du très faible nombre de points dans chaque classe, l'effet de la valeur maximale sur l'écart type est encore plus fort, et il est peu probable que le passage au logarithme amène un changement.

Tableau 7. Logarithmes des HAP : statistiques de base sur les 18 échantillons de longueur comprise entre 0,1 m et 0,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
log10 HAP	18	-0,52	3,81	1,82	1,14	0,63
log10(1+HAP/moyenne HAP)	18	0,11	3,81	1,88	1,04	0,55

Tableau 8. Logarithmes des HAP : statistiques de base sur les 18 échantillons de longueur comprise entre 0,5 m et 1,0 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
log10 HAP	18	-0,70	2,63	1,16	0,87	0,75
log10(1+HAP/moyenne HAP)	18	0,08	2,64	1,28	0,70	0,55

Tableau 9. Logarithmes des HAP : statistiques de base sur les 18 échantillons de longueur comprise entre 1,0 m et 1,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
log10 HAP	18	-0,52	2,67	1,39	0,79	0,57
log10(1+HAP/moyenne HAP)	18	0,11	2,67	1,45	0,70	0,48

Tableau 10. Logarithmes des HAP : statistiques de base sur les 18 échantillons de longueur supérieure à 1,5 m.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
log10 HAP	19	0,60	3,70	1,52	0,81	0,53
log10(1+HAP/moyenne HAP)	19	0,70	3,70	1,56	0,78	0,50

Nous ne pouvons donc pas trancher, même s'il est inhabituel que des teneurs très fortes se trouvent dans des classes autres que celle des longueurs les plus faibles.

Pouvons-nous décider s'il est possible de mélanger les échantillons de longueur inconnue (« ponctuels ») avec les échantillons de faible longueur ?

Tableau 11. Statistiques de base sur les 85 échantillons de longueur inconnue.

	Nb données	Minimum	Maximum	Moyenne	Ecart-type	σ/m
HAP	64	0,10	2 000,00	187,75	365,74	1,95
Pb	21	19,00	6 000,00	1 165,19	1 668,23	1,43
Zn	21	49,00	14 690,00	2 533,81	3 997,26	1,58

En comparant le tableau 11 (statistiques des échantillons de longueur inconnue) et le tableau 3 (statistiques des échantillons de longueur inférieure à 50 cm) on voit que là encore, c'est la concentration la plus élevée qui détermine l'écart-type le plus élevé. Il ne nous est pas possible de décider si les concentrations moins élevées sont dues à un effet de lissage quand on augmente la taille du support ou simplement au fait que l'échantillonnage a trouvé ou non les fortes concentrations.

Avec nos échantillons, seules deux configurations d'échantillonnage permettent un travail géostatistique fiable :

- Si les longueurs théoriques sont bien les longueurs vraies des échantillons, nous ne pouvons pas les mélanger et nous ne pouvons pas effectuer de régularisation à cause du trop faible nombre d'échantillons. Il aurait été envisageable de contourner cette difficulté en travaillant sur la variable accumulation (concentration $\times$ longueur) sur toute la hauteur de la couche étudiée, par exemple entre la surface et le toit des alluvions. Malheureusement, comme 7 fosses seulement atteignent le toit des argiles, les limites de la couche sont mal définies. Cette méthode ne nous permettrait pas non plus d'utiliser les analyses ponctuelles.
- Si tous les échantillons ont le même support ou des supports proches, on peut les utiliser tous ensemble pour le calcul des variogrammes expérimentaux. Malheureusement, nous n'avons pas les informations nous permettant de décider si ce mélange est autorisé ou non.

À ce stade, étant donnée l'association entre la non connaissance du support des échantillons et la très grande hétérogénéité du terrain (donc des mesures), il serait raisonnable de décider de reprendre en détail la description de l'échantillonnage avec les personnes en charge de ce dernier. Si comme ici les mesures sont trop anciennes pour pouvoir rassembler des informations complémentaires à leur sujet, il est sans doute plus réaliste de renoncer à poursuivre l'étude géostatistique : il est bien sûr possible de tenter de modéliser la variabilité spatiale et de 'utiliser ce modèle pour une cartographie, mais l'énorme incertitude sur le modèle rendra hasardeuse toute interprétation des propriétés des modèles, et a fortiori des résultats.

Nous avons néanmoins poursuivi l'étude, en tentant de montrer à chaque stade en quoi la mauvaise qualité des données obère la possibilité d'obtenir des résultats satisfaisants.

4. ANALYSE VARIOGRAPHIQUE ET ESTIMATION

Pour l'analyse variographique, nous avons choisi de considérer comme ponctuelles les données pour lesquelles nous n'avons pas d'information de longueur. Les deux types de données, ponctuelles et non ponctuelles sont d'abord étudiés séparément puis mélangées. Dans ce dernier cas, toutes les données ont été traitées comme ponctuelles : il aurait fallu avoir une information de longueur sur toutes les données pour pouvoir faire un autre choix.

En raison de leur corrélation négligeable, les HAP et les métaux lourds sont étudiés séparément. De plus, le nombre d'échantillons plus faible ne permet pas d'analyser séparément les données ponctuelles et non ponctuelles pour les métaux lourds, chacune de ces populations ne contenant pas assez d'échantillons.

4.1 Etude des données ponctuelles (HAP)

4.1.1 Variogrammes expérimentaux.

Pour calculer les variogrammes expérimentaux, il nous faut effectuer le choix des paramètres de calcul : directions des variogrammes, pas de calcul et tolérances.

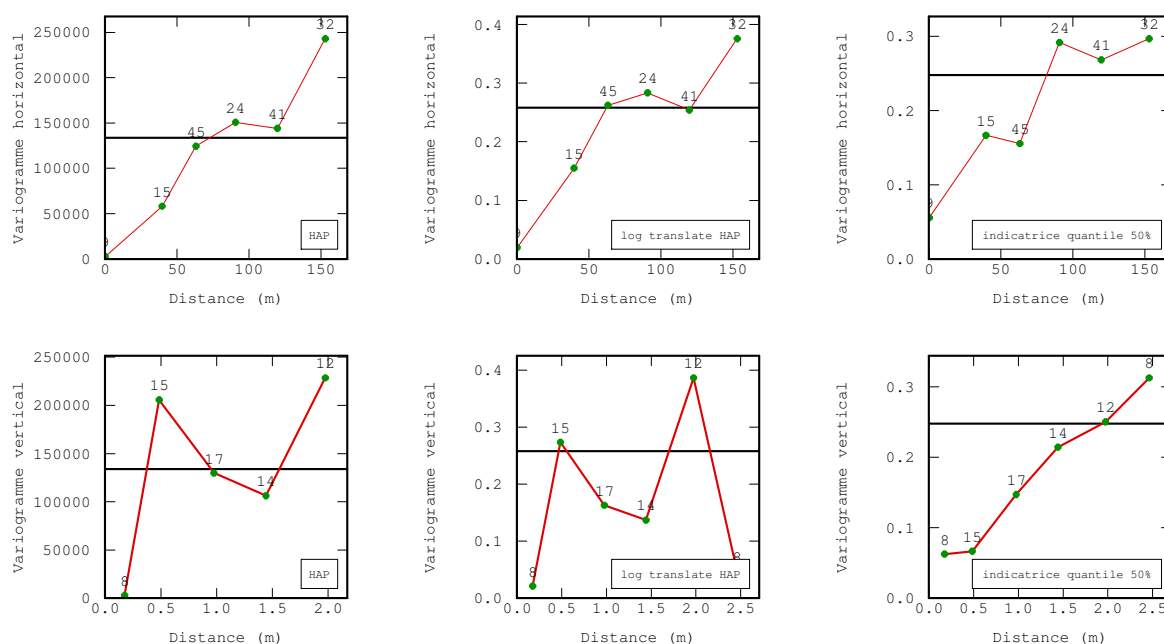


Figure 13. Variogrammes expérimentaux de la concentration (gauche), du logarithme traduit (centre), de l'indicateur du quantile 50%(droite). Haut : variogrammes verticaux ; bas : variogrammes horizontaux.

Etant donné le nombre de données assez faible, nous ne pouvons pas calculer les variogrammes expérimentaux dans un grand nombre de directions : il nous faudrait prendre des tolérances d'angle trop grandes (la tolérance d'angle permet de regrouper les calculs dans des directions proches), et la précision sur les directions serait illusoire. Nous avons donc choisi de calculer simplement un variogramme horizontal et un variogramme vertical. Les pas de calcul sont de 30 m dans la direction horizontale et 50 cm dans la direction verticale. Ces pas de calculs correspondent à un compromis entre la précision du variogramme expérimental et la possibilité de modélisation : si le pas est trop grand, on mélange la variabilité à des distances trop différentes et cela n'a aucun sens. Si le pas est trop petit, le variogramme est trop erratique et on ne peut pas le modéliser. La tolérance sur les distances, qui permet de mélanger des distances légèrement différentes, a été choisie à 50 % du pas de calcul. La figure 13 (colonne de

gauche) montre les variogrammes expérimentaux ainsi obtenus (haut : variogramme horizontal, bas variogramme vertical).

On voit en haut à gauche que le variogramme horizontal est relativement régulier et assez facile à modéliser. Par contre le variogramme vertical (en bas à gauche) est beaucoup plus erratique, nous ne disposons pas de modèle permettant d'ajuster facilement un tel variogramme. Lorsque les données sont peu nombreuses ou mal réparties, il est assez fréquent d'obtenir des variogrammes expérimentaux difficilement modélisables, surtout quand les distributions statistiques des mesures sont très fortement dissymétriques comme dans notre cas (beaucoup de valeurs très faibles et quelques valeurs très fortes). Dans le cas des distributions dissymétriques, nous disposons d'outils pour tenter d'obtenir des variogrammes modélisables en diminuant l'influence des valeurs extrêmes. Nous pouvons par exemple calculer le logarithme translaté de la variable d'étude (variogrammes du centre sur la figure 13) ou l'indicatrice du quantile 50 % (variogrammes de droite). Le logarithme translaté est $\log_{10}(1 + \text{HAP} / \text{moyenne HAP})$, où HAP est la concentration et m la moyenne des concentrations. L'indicatrice du quantile 50 % vaut 1 lorsque la concentration est inférieure à la valeur médiane, 0 sinon. Sur les variogrammes horizontaux de la figure 13 (ligne du haut) on voit que les trois variables, si elles ont bien des variogrammes différents, présentent des variogrammes avec des portées comparables (portée : distance à laquelle le variogramme se stabilise). Sur les variogrammes verticaux (ligne du bas) on voit que le logarithme translaté présente un comportement similaire à la variable brute, mais que l'indicatrice du quantile 50% est beaucoup plus facilement modélisable. Nous choisirons donc la portée issue du variogramme de cette indicatrice pour modéliser le variogramme vertical de la variable brute, puisque les portées sont en général comparables pour ces trois variables (cette portée s'obtient en étudiant conjointement les variogrammes horizontaux et verticaux : pour les variogrammes du quantile 50%, le palier se voit sur le variogramme horizontal, à condition d'exclure le dernier point qui est calculé avec très peu de paires ; on voit ensuite sur le variogramme vertical la valeur où la courbe atteint ce palier). On doit remarquer que le nombre de paires de points utilisées pour calculer chaque point du variogramme expérimental est faible (nombres sur les courbes de la figure 13) : ces variogrammes expérimentaux, même quand ils paraissent réguliers, sont peu fiables.

4.1.2 Ajustement des variogrammes

La figure 14 montre l'ajustement des variogrammes de la concentration en HAP. L'ajustement consiste à faire correspondre un modèle de variogramme (courbes lisses sur la figure 14) au variogramme expérimental (courbe segmentée reliant les points verts). Cet ajustement s'effectue interactivement en comparant les courbes. Il existe des méthodes automatiques ou semi-automatiques, mais un contrôle de la qualité de l'ajustement par l'utilisateur est indispensable.

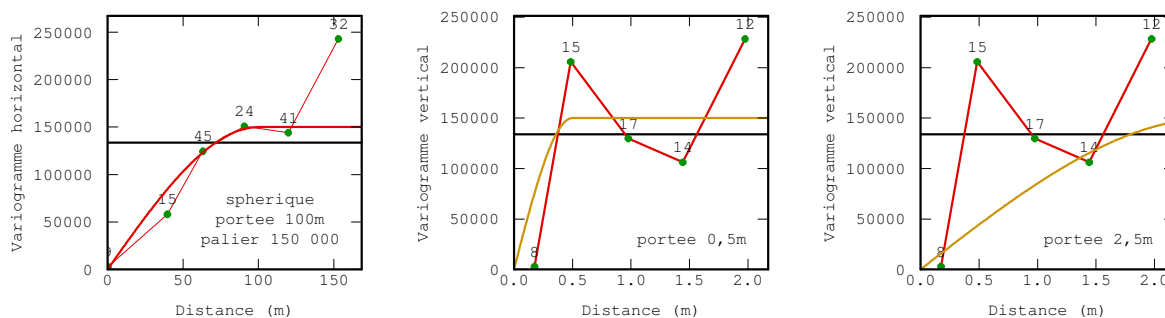


Figure 14. Ajustement des variogrammes horizontal (gauche) et vertical (centre, droit) de la concentration en HAP.

On voit sur la figure 14 que le variogramme horizontal (à gauche) se modélise assez facilement. Nous avons ajusté un modèle de variogramme sphérique, de portée 100 m et de palier (valeur du plateau de stabilisation) 150 000. Le variogramme vertical est plus difficile à ajuster. Les composantes horizontale et verticale du modèle sont obligatoirement liés, nous avons donc ajusté à nouveau un modèle sphérique de

palier 150 000 (le palier correspond à la variance théorique de la concentration, qui n'est pas obligatoirement égale à la variance expérimentale) : par contre les portées sont différentes. L'image du centre montre l'ajustement qu'on effectuerait à partir de la concentration brute seulement (portée 50 cm). L'image de droite montre l'ajustement effectué en tenant compte de l'information donnée par l'indicatrice du quantile 50 % (portée 2,5 m). Les deux ajustements sont visuellement mauvais. La portée correspondant à la distance de non corrélation, nous pouvons choisir de garder la portée de 2,5 m, faisant l'hypothèse que depuis 1974 les HAP ont pu diffuser dans les divers terrains à une distance supérieure à 50 cm et présentent donc probablement une continuité verticale marquée. Une comparaison entre les valeurs des paramètres et des caractéristiques connues des hétérogénéités du sous-sol (distance entre les murs enterrés par exemple) pourrait permettre de choisir la portée la plus réaliste, mais en l'absence de connaissance de ces caractéristiques des hétérogénéités la validation croisée peut aider à choisir entre ces deux modèles.

4.1.3 Validation croisée.

La qualité de l'ajustement des variogrammes peut être testée par validation croisée. On ôte successivement chacun des échantillons de la base de donnée, et on le ré-estime par krigeage à partir de tous les autres. Lorsque la validation croisée est bonne, cela signifie que les erreurs d'estimations (inévitables) sont cohérentes avec le modèle ajusté, qu'elles s'expliquent facilement par la variabilité inhérente au modèle. Lorsque la validation croisée est mauvaise, cela signifie que le modèle n'explique pas complètement la variabilité du phénomène.

Divers indicateurs sont utilisés pour évaluer la qualité d'une validation croisée. La figure 15 montre, pour l'estimation faite avec les deux modèles, de portée verticale 0,5 m et de portée verticale 2,5 m, le nuage de corrélation valeur vraie – valeur estimée (une estimation parfaite –irréalisable- aboutirait à un nuage sur la première bissectrice).

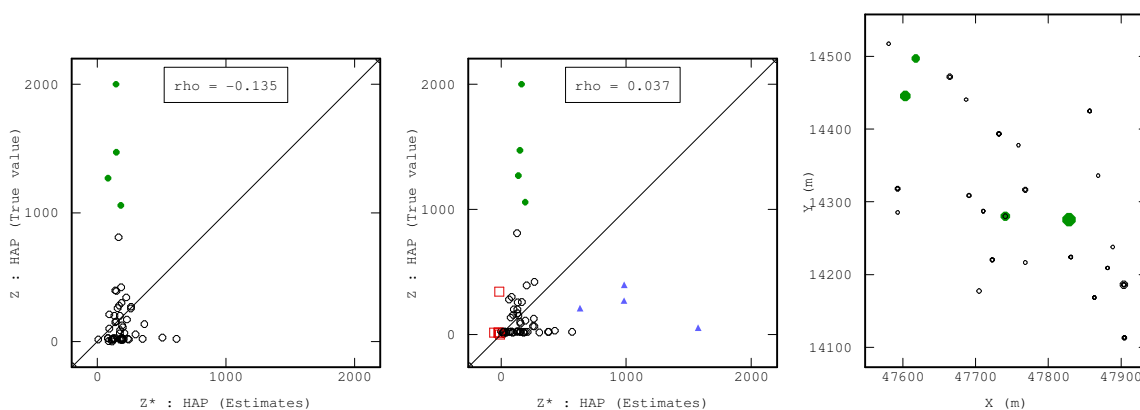


Figure 15. Evaluation de la qualité de l'ajustement par validation croisée. Gauche : portée verticale 0,5 m ; centre : portée verticale 2,5 m ; droite : carte des données ponctuelles.

Les résultats de la validation croisée sont ici très mauvais pour nos deux modèles : si les nuages de corrélation sont effectivement centrés sur la première bissectrice, on voit une forte sous-estimation des valeurs fortes (contamination non détectée, points verts) et dans le cas du modèle à forte portée verticale une surestimation de certaines valeurs faibles (forte contamination détectée à tort, points bleus). Les coefficients de corrélation sont très faibles, et même faiblement négatif (-0,135) pour le modèle à portée verticale 0,5 m. Par contre, le modèle à grande portée verticale, s'il est légèrement meilleur pour le coefficient de corrélation positif, est nettement plus mauvais en ce qui concerne la surestimation des valeurs faibles. Les fosses contenant des échantillons très fortement sur- ou sous- estimés sont représentées par des points verts sur la carte des données.

Le modèle à portée verticale 2,5 m est également plus mauvais que le modèle à portée verticale 0,5m pour plusieurs autres critères :

- il présente des valeurs estimées négatives (en rouge sur le nuage de corrélation),
- la variance de l'erreur standardisée est beaucoup trop forte, 2,95, alors qu'elle est égale à 1,06 pour le modèle à courte portée verticale, c'est-à-dire proche de 1 comme il est attendu. L'erreur standardisée est égale au quotient erreur de krigeage / écart type de krigeage. Quand le modèle de variogramme est correct, sa variance est proche de 1. Cela signifie que même si les valeurs estimées sont très éloignées des valeurs vraies, cette différence est bien expliquée par le modèle. Au regard de ce critère, le modèle à courte portée verticale est correct et c'est celui qui sera conservé.

Quand les variogrammes expérimentaux sont très mauvais ou incalculables, la validation croisée peut parfois permettre de proposer un modèle : on garde le modèle qui aboutit à la meilleure validation croisée (cette méthode est à éviter quand on dispose de variogrammes de bonne qualité). Dans notre cas, il n'a pas été possible de trouver un modèle meilleur que celui ajusté directement sur la variable brute (modèle à courte portée verticale).

Nous avons donc finalement choisi de conserver le modèle ajusté sur les données brutes, modèle sphérique de portée horizontale 100 m et de portée verticale 0,5 m. La validation croisée montre que l'estimation sera de mauvaise qualité, mais que le modèle ajusté représente assez bien la variabilité des données. Il ne nous est malheureusement pas possible de suggérer des pistes pour améliorer le modèle de variogramme tant que le problème du support des échantillons n'est pas résolu. Cette modélisation difficile peut par exemple s'expliquer par des données « ponctuelles » qui sont en réalité à support variable.

4.1.4 Cartographie : estimation par krigeage

Une fois le modèle ajusté et validé, on peut procéder à l'estimation, dans le but de dessiner une carte de la contamination par exemple. Le problème de support des données n'étant pas résolu, nous avons procédé à une simple interpolation numérique, c'est-à-dire à un krigeage ponctuel. La comparaison avec les seuils de contamination, évalués pour des moyennes sur des volumes plus grands, ne sera donc peut-être pas pertinente.

La figure 16 montre la carte de la moyenne des estimations de tous les plans horizontaux (à gauche) et de la moyenne des écarts-types de ces mêmes plans (à droite). L'écart type de krigeage est l'écart type de l'erreur d'estimation. C'est un indicateur de la qualité locale de l'estimation, dans la mesure où le modèle ajusté est correct. Quand il est élevé, l'erreur peut être élevée, quand il est faible l'erreur est très probablement faible. Pour interpréter un écart-type de krigeage, il est nécessaire d'avoir confiance en la qualité du modèle : un écart-type de krigeage calculé à partir d'un modèle faux n'a aucun sens. Il a été choisi de dessiner des cartes de moyennes pour résumer les résultats d'estimation ponctuelle. Une analyse détaillée des résultats demanderait bien sûr des cartes niveaux par niveau. Il serait cependant plus judicieux d'effectuer ces cartes niveau par niveau à partir d'estimations de blocs qui correspondraient mieux aux unités de sélection pour la dépollution éventuelle (voir paragraphe 4.1.6)

Dans notre cas, le modèle est probablement mauvais et nous n'essaierons pas d'interpréter l'écart-type de krigeage (qui de toute façon s'interpréterait sur des cartes niveau par niveau et pas sur la moyenne). Par contre, on peut voir sur cette carte un comportement assez universel : l'écart-type de krigeage est faible lorsque l'on est proche des données et augmente lorsque l'on s'en éloigne. Ce phénomène s'explique par le fait que la corrélation entre données et point à estimer diminue quand on s'éloigne des données. On a donc plus de chances d'avoir une erreur importante en s'éloignant.

La carte de l'estimation, qui montre des valeurs fortes en périphérie et faibles au centre, est irréaliste, puisque la contamination en HAP provient de l'usine, et donc du centre : il est extrêmement peu probable que la contamination ait ainsi « fuit » la zone d'origine.

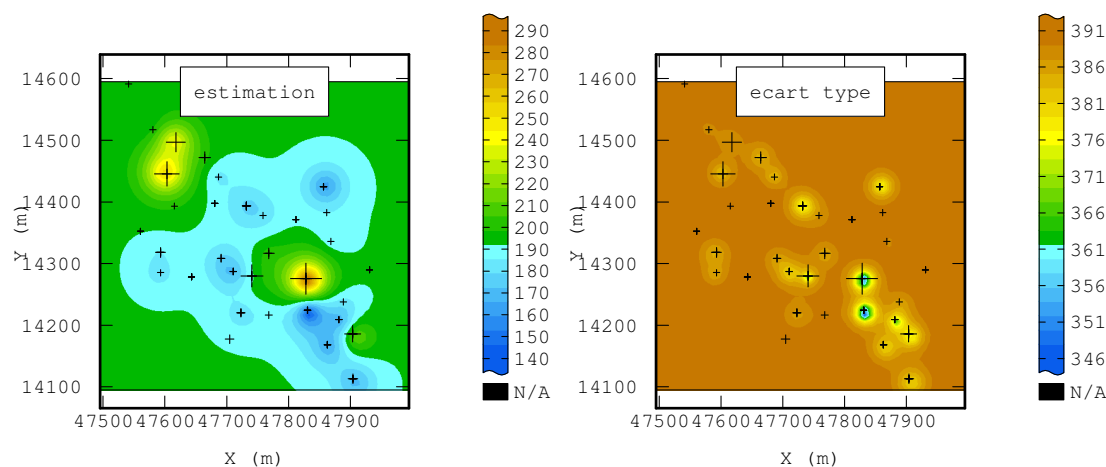


Figure 16. Moyennes des estimations par krigeage (gauche) et moyenne des écarts-type de krigeage (droite) : moyennes des plans horizontaux

4.1.5 Changement de coordonnée verticale : utilisation de la profondeur

Les résultats précédents ont été obtenus en utilisant l'altitude comme coordonnée verticale. L'avantage est que les distances horizontales ou obliques par rapport à la verticale sont les distances vraies. Mais la pollution provenant a priori de la surface, il peut être plus judicieux d'utiliser la profondeur comme coordonnée verticale (en lui affectant un signe négatif pour que l'orientation positive de l'axe vertical soit vers le haut). De plus, les cartes de concentration obtenues suivront la topographie du sol, ce qui correspond davantage à une dépollution par décapage.

Le jeu de données présentait cependant des incohérences : la variable profondeur (issue de calculs lors d'une étude passée ?) pouvait être la même pour deux échantillons différents, présentant les mêmes coordonnées horizontales mais des altitudes différentes. Les concentrations en HAP sont également différentes. En comparant les profondeurs et altitudes de l'ensemble des échantillons des sondages correspondants, il a été possible de retrouver les erreurs sur les profondeurs et de les corriger.

Si on trace les nuages de corrélation entre concentration en HAP et profondeur, les graphiques se ressemblent (figure 17). On peut cependant remarquer que les échantillons sont répartis de façon plus régulière lorsqu'on utilise la profondeur comme coordonnée verticale. Dans les deux cas, on voit que la partie la plus profonde de la zone d'étude ne présente que des concentrations faibles.

Avec cette nouvelle coordonnée verticale, les variogrammes de la concentration en HAP, de son logarithme translaté et de l'indicateur du quantile 50 % ont à nouveau été calculés et ajustés. La figure 17 montre qu'il est possible d'ajuster la concentration en HAP en utilisant les paramètres du modèle du logarithme translaté (les paliers de chaque structure de base sont réajustés). Par contre, on voit sur la figure 18 qu'il n'est pas possible d'ajuster la concentration en HAP en se contentant de modifier les paliers du modèle du quantile 50 %. Nous avons cependant conservé ce dernier modèle, qui propose un ajustement de mauvaise qualité, pour le tester par validation croisée.

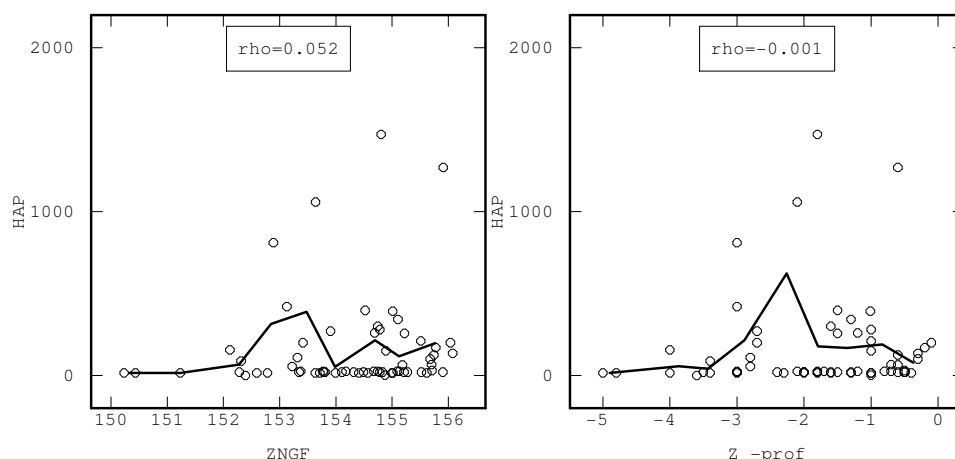


Figure 17. Nuage de corrélation HAP fonction de la coordonnée verticale. Gauche : coordonnée verticale altitude ; droite coordonnée verticale profondeur. Ligne noire : espérance conditionnelle (moyenne par classes)

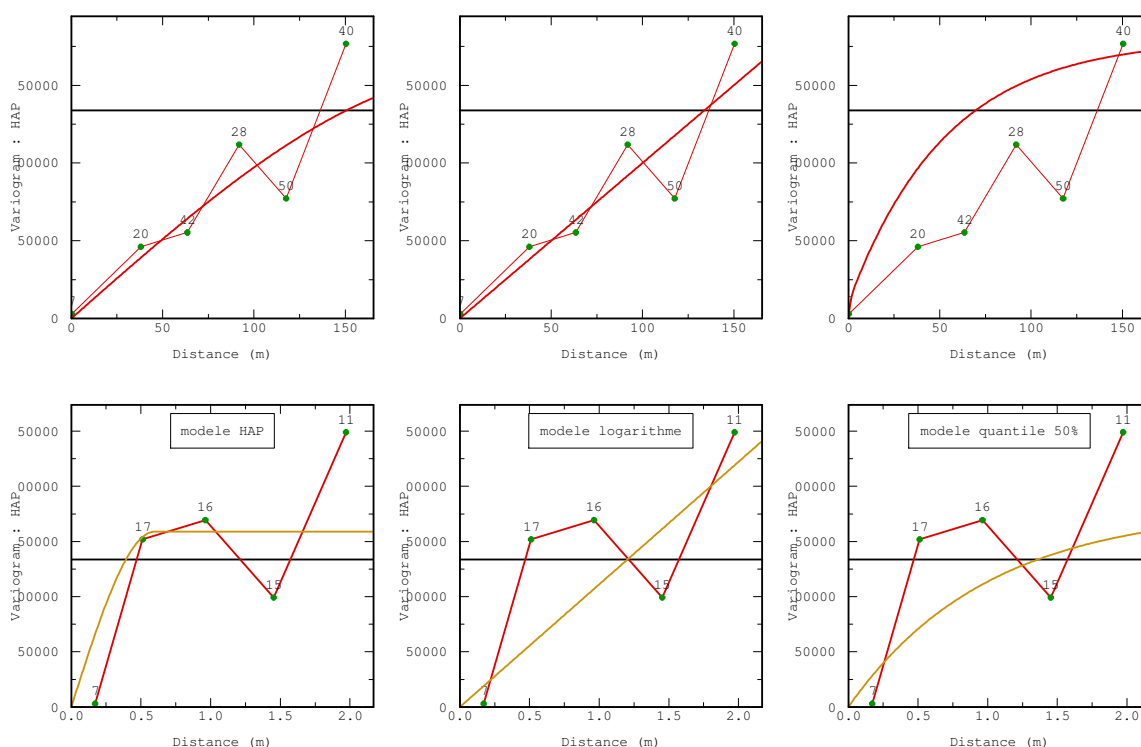


Figure 18. Ajustement des variogrammes horizontal (haut) et vertical (bas) de la concentration en HAP. Gauche : modèle de la concentration ; centre : modèle du logarithme translaté (recalé) ; droite : modèle de l'indicateur du quantile 50% (recalé).

Les modèles ajustés sont résumés dans le tableau 12 (pour le modèle 2, linéaire, les valeurs notées dans le tableau sont un coefficient multiplicatif et non un palier, et des facteurs d'échelle et non des portées).

On voit sur la figure 18 que les ajustements calculés à partir du logarithme translaté et à partir de l'indicateur du quantile 50 % sont de mauvaise qualité : pour la direction verticale seulement pour le premier, pour les deux directions pour le second. Tous les modèles sont toutefois testés par la validation croisée.

Tableau 12. Modèles ajustés pour la concentration en HAP. Modèle 1 : modèle calculé uniquement à partir de la concentration des HAP ; modèle 2 : modèle basé sur le logarithme translaté de la concentration en HAP ; modèle 3 : modèle calculé à partir de l'indicatrice du quantile 50% de la concentration en HAP.

	Modèle 1	Modèle 2	Modèle 3	
			Structure 1	Structure 2
Type de modèle	Sphérique	Linéaire	Exponentiel	Exponentiel
Palier	159 000	500 000	10 000	170 000
Portée horizontale	230 m	500 m	3 m	160 m
Portée verticale	0,75 m	4,5 m	3 m	3 m

Remarquons que le variogramme horizontal expérimental de la concentration en HAP ne montre pas clairement la portée horizontale du modèle (distance à laquelle le variogramme atteint son palier). Pour mieux la mettre en évidence, le variogramme horizontal a été recalculé avec des pas de calculs de 50 m puis 60 m. Comme le calcul a à nouveau été mené pour 6 pas de calcul, la distance maximale atteinte est plus élevée est plus grande dans ces deux cas. La figure 19 montre comment le modèle ajuste les trois variogrammes expérimentaux, pas de 30 m, de 50 m et de 60 m. On voit que le modèle ajuste assez bien les deux premiers variogrammes expérimentaux, et moins bien le variogramme expérimental calculé avec un pas de 60 m, mais il est possible que celui-ci moyenne des distances trop différentes à chaque pas de calcul, et qu'il ne soit pas vraiment significatif.

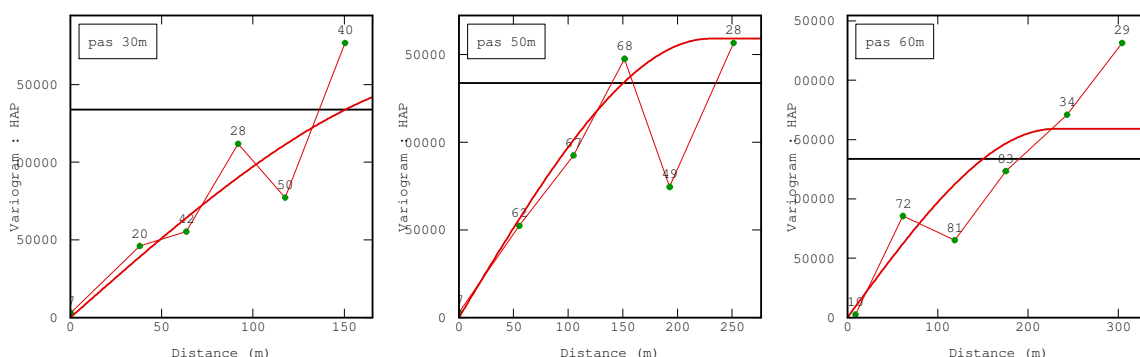


Figure 19. Ajustement des variogrammes expérimentaux des HAP calculés avec les pas de 30 m (gauche), 50 m (centre) et 60 m (droite). Modèle numéro 1 (sphérique avec une portée de 230 m).

La validation croisée donne de mauvais résultats pour ces 3 modèles. Le tableau 13 résume les résultats de ces trois validations croisées. Dans chaque cas, nous avons souligné le meilleur résultat et fait figurer en rouge la valeur « idéale » d'une estimation parfaite (irréaliste).

Si la distribution de l'erreur standardisée (erreur divisée par l'écart-type d'estimation) est gaussienne et si le palier du modèle a été correctement estimé, le nombre de données non robustes doit idéalement représenter 1% du nombre total de données (soit 0,64 arrondi à 1). Une estimation parfaite valeur vraie = valeur estimée donnerait un coefficient de corrélation égal à 1. Il est irréaliste d'attendre un tel résultat, mais on souhaite s'en approcher le plus possible. L'erreur standardisée et la valeur estimée ne devraient normalement pas être corrélées. On cherche à avoir une erreur la plus faible possible, sa moyenne doit donc être 0 (estimateur sans biais) et sa variance la plus faible possible.

Tableau 13. Résultats des validations croisées des trois modèles de variogrammes. Voisinage sphérique de rayon 200m

	Modèle 1	Modèle 2	Modèle 3
Nombre de données non robustes 1	6	6	<u>4</u>
Coefficient de corrélation valeur vraie / valeur estimée 1	0,011	<u>0,133</u>	0,090
Coefficient de corrélation erreur standardisée / valeur estimée 0	0,505	0,442	<u>0,439</u>
Moyenne de l'erreur 0	13,36	<u>7,52</u>	9,75
Variance de l'erreur	180 500	165 750	<u>163 500</u>
Variance de l'erreur standardisée 1	2,01	3,25	<u>1,65</u>

Enfin, la variance de l'erreur standardisée est égale à 1 quand le modèle explique correctement l'erreur : nettement plus faible que 1 l'estimation est inexplicablement bonne, nettement plus élevée le modèle n'explique pas bien l'erreur constatée. Mais un simple changement de palier (coefficient multiplicatif) peut modifier considérablement cette erreur standardisée, elle n'est donc utile que si le palier est correctement ajusté, au moins au vu des variogrammes expérimentaux. Dans notre cas, le voisinage de krigeage est de grandes dimensions (diamètre de 400 m). Le krigeage utilise alors les valeurs du variogramme aux grandes distances. Comme les comportements des différents variogrammes aux longues distances sont très différents de l'un à l'autre, cela affecte la variance de krigeage de façon importante, et cela peut être responsable de cette différence sur les résultats affectant l'erreur standardisée. On s'attachera donc plutôt aux statistiques de l'erreur non standardisée et au coefficient de corrélation valeur vraie / valeur estimée (lignes grisées).

Le coefficient de corrélation valeur vraie / valeur estimée est un peu plus élevé pour le modèle 2, mais reste tellement faible qu'on ne basera pas un choix de modèle sur ce dernier. Ce modèle est également légèrement meilleur que les autres pour la moyenne de l'erreur, alors que c'est le modèle 3 qui est le meilleur en ce qui concerne la variance de l'erreur.

On peut noter que le modèle 1, ajusté directement à partir de la concentration en HAP, est le plus mauvais pour tous les critères de choix. Ce peut être dû à la portée verticale trop faible, proche du pas de calcul du variogramme expérimental. La validation croisée peut également être une aide à l'ajustement quand le variogramme expérimental ne peut pas être calculé convenablement ou ne permet pas de faire un choix (par exemple quand le nombre de données est trop faible). Dans notre cas, les mauvais résultats sont probablement dus à un voisinage trop grand : on sait que la zone d'étude est extrêmement hétérogène, alors que les variogrammes horizontaux ne présentent pas d'effet de pépité. Cela semble incohérent, et il est sans doute également incohérent d'aller chercher des points de donnée trop éloignés pour l'estimation.

Il aurait été possible de tester également un variogramme présentant un effet de pépité. Cela aurait fourni une estimation plus « pessimiste » (plus lisse, avec un écart type de krigeage plus élevé). Nous souhaitons voir tout d'abord si le modèle optimiste fournissait un intervalle de confiance conventionnel permettant de discriminer des zones probablement contaminées et probablement non contaminées, quitte à corriger ensuite cet excès d'optimisme en réintroduisant un effet de pépité. Nous verrons que même avec le modèle optimiste, l'intervalle de confiance conventionnel ne permet pas cette discrimination, ou trop rarement. Il n'a donc pas été considéré nécessaire de réintroduire un effet de pépité.

Pour le modèle 1, si on fait varier la taille du voisinage, le meilleur résultat de validation croisée en ce qui concerne la variance de l'erreur est obtenu pour un voisinage ellipsoïde de rayon 75 m dans le plan horizontal et de demi-axe 15 m dans la direction verticale (mais la moyenne de l'erreur est un peu plus forte que pour le grand voisinage). Cette dimension de 15 m, bien supérieure à l'épaisseur de la zone d'étude, signifie que l'on prend toute la hauteur de la colonne pour les sondages proches du point estimé, mais que pour les sondages les plus éloignés seuls les points situés au même niveau sont pris en considération. Diminuer le demi-axe vertical de l'ellipsoïde à 2 m entraîne des résultats plus mauvais.

Pour ce voisinage plus restreint, les validations croisées des trois modèles donnent les résultats fournis dans le tableau 14.

Cette fois-ci, le modèle ajusté directement sur les HAP (qui ajuste mieux les variogrammes expérimentaux) est le moins mauvais pour tous les critères sauf le nombre de données robustes et la moyenne de l'erreur. C'est finalement ce modèle que nous utiliserons, avec ce petit voisinage, pour l'estimation, même si nous anticipons une estimation de mauvaise qualité.

Il faut remarquer que ce « meilleur » voisinage présente des demi-axes de dimensions bien inférieures à la portée dans le plan horizontal, et bien supérieures dans la direction verticale. La règle parfois appliquée d'utiliser les valeurs des portées comme dimensions des demi-axes est peu judicieuse dans ce cas. Elle n'offre qu'un ordre de grandeur pour commencer les tests de voisinages.

L'utilisation de la profondeur comme coordonnée verticale a l'avantage de permettre l'estimation dans une couche d'épaisseur constante suivant la topographie. Dans le chapitre 4.1.4 il avait été nécessaire d'estimer sur une grille d'épaisseur d'au moins 6 m pour pouvoir couvrir toute la zone en raison d'une dénivellation de la topographie d'environ 6 m. Il est maintenant possible de se limiter à la couche de 3 m qui intéresse réellement cette étude.

Tableau 14. Résultats des validations croisées des trois modèles de variogrammes. Voisinage ellipsoïdal de rayon 75 m dans la direction horizontale, et de demi-axe 15 m dans la direction verticale.

	Modèle 1	Modèle 2	Modèle 3
Nombre de données non robustes 1	6	6	<u>4</u>
Coefficient de corrélation valeur vraie / valeur estimée 1	<u>0,107</u>	0,103	0,091
Coefficient de corrélation erreur standardisée / valeur estimée 0	<u>0,444</u>	0,461	0,489
Moyenne de l'erreur 0	20,36	<u>15,16</u>	25,64
Variance de l'erreur	<u>171 000</u>	182 500	184 600
Variance de l'erreur standardisée 1	<u>1,49</u>	3,06	1,71

4.1.6 Coordonnée verticale : profondeur. Estimation de bloc

L'utilisation de la profondeur comme coordonnée verticale nous permet d'estimer directement la concentration moyenne sur une maille d'épaisseur 3 m (pour l'estimation précédente, il aurait fallu extraire la couche superficielle de 3 m de l'ensemble des mailles minces estimées. Cela nécessiterait une mesure ou une estimation de la topographie à chaque nœud de maille horizontale.

Le modèle de variogramme utilisé pour effectuer un krigeage de bloc est celui choisi à partir de la validation croisée du chapitre 4.1.5 (modèle 1, ajusté directement sur les données brutes). Le voisinage est également identique (petit voisinage de rayon 75 m dans le plan horizontal et de demi axe 15 m dans la direction horizontale), à ceci près qu'il est en plus indispensable d'y ajouter une discrétisation du bloc. Cette discrétisation revient à remplacer, pour le calcul de variogramme moyen ou de covariance moyenne utilisé dans le système de krigeage, une intégrale sur le bloc par une moyenne de valeurs ponctuelles calculées dans le bloc. Plus la discrétisation est fine (plus le nombre de points est élevé), meilleure sera l'approximation mais cette amélioration se fait aux dépens de la vitesse de calcul. Il faut trouver un compromis efficace.

Pour plusieurs choix de discrétisation, les covariances moyennes ont été calculées. Dans chaque cas, la position des points de discrétisation est légèrement décalée de manière aléatoire, pour 11 évaluations séparées. Quand les résultats de ces 11 évaluations sont suffisamment stables, on considère que la discrétisation convient. A noter que la discrétisation verticale est normalement limitée par le rapport entre la dimension verticale de la maille et la dimension verticale de l'échantillon. Dans notre cas, nous ne connaissons pas la dimension verticale des échantillons. Nous supposons qu'elle est d'environ 50 cm et n'utiliserons pas de discrétisation supérieure à 6 points dans la direction verticale.

Le tableau 15 montre quelques-uns des tests de discrétisation effectués. On voit que la discrétisation doit être suffisante dans les trois directions, il est inutile de discrétiser finement dans les directions horizontales si la direction verticale n'est pas suffisamment discrétisée. La discrétisation finalement choisie est celle dont l'écart-type des 11 valeurs est le plus faible, ce qui est un signe de stabilité. Il est bien entendu possible de discrétiser davantage, mais cela se fait au détriment du temps de calcul pour le krigeage, pour un avantage minime. Nous utiliserons donc la discrétisation 8 x 8 x 5, respectivement dans les directions est-ouest, nord-sud et verticale, ce qui signifie qu'on remplace le bloc par un total de $8 \times 8 \times 5 = 320$ points pour le calcul du système de krigeage.

La figure 20 montre les résultats de l'estimation. Sur les 2 500 mailles, seules 1 634 ont été estimées (un voisinage de krigeage petit laisse des mailles pour lesquelles aucun échantillon n'est présent dans le voisinage).

Tableau 15. Tests de discrétisation pour le krigeage de blocs

Discrétisation	moyenne	écart type
10 x 10 x 3	108 792,899	439,623514
9 x 9 x 4	108 320,981	421,125578
<u>8 x 8 x 5</u>	<u>108 014,831</u>	<u>248,192704</u>
7 x 7 x 6	107 978,69	293,041129

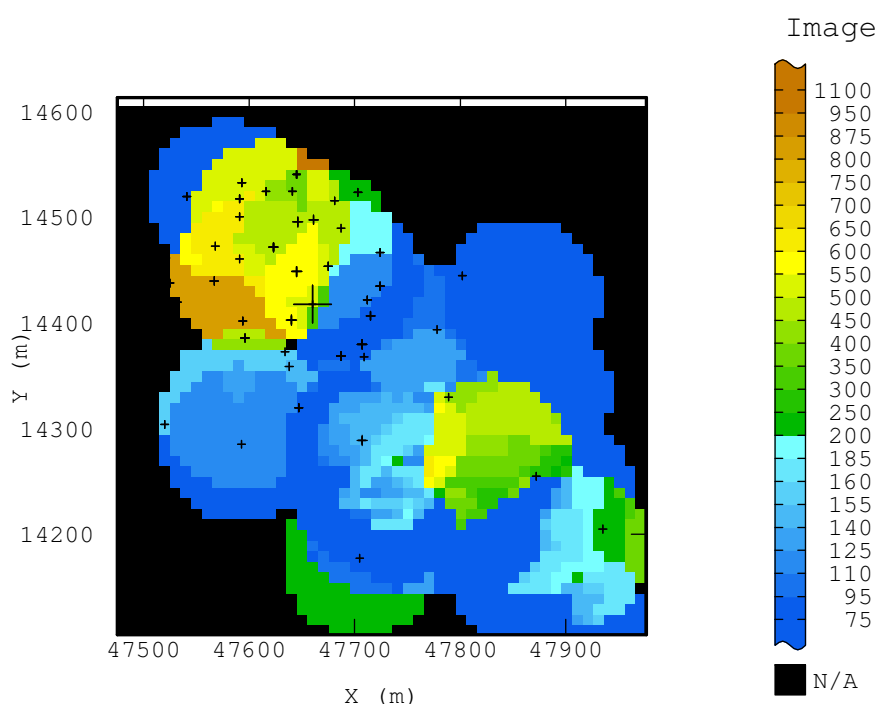


Figure 20. Estimation de la concentration en HAP sur les mailles de 10x10x3 m³. Utilisation du voisinage de 75 m.

A partir de ces résultats, on peut construire l'intervalle de confiance conventionnel $[z_k - 2\sigma_k, z_k + 2\sigma_k]$. Lorsque le maximum de cet intervalle de confiance est inférieur au seuil de contamination, on peut considérer que la maille n'est pas polluée avec une probabilité élevée (dans la mesure où on a confiance dans la qualité de l'ajustement de variogramme, ce qui n'est pas notre cas ici). Aucune maille n'est dans ce cas. Lorsque le minimum de cet intervalle de confiance est supérieur au seuil de contamination, on peut dire que la maille est polluée avec une probabilité élevée, toujours dans la mesure où on fait confiance au

modèle. C'est le cas ici pour 216 mailles, soit environ 13% des mailles estimées. Cet intervalle de confiance conventionnel est calculé dans le cas où la distribution de la variable étudiée n'a pas été modélisée. Par analogie avec une distribution normale pour lequel il correspond à une probabilité de 95%, on suppose que la probabilité d'être dans cet intervalle est très élevée, sans toutefois pouvoir la quantifier exactement. Il existe bien entendu des distributions pour lesquelles la probabilité d'être dans cet intervalle est plus faible. L'intervalle de confiance conventionnel donne une indication très utile, mais qui doit être confirmée par des calculs de probabilités de dépassement de seuil plus précis, provenant d'autres méthodes comme les simulations.

On voit que l'estimation présente de nombreux artéfacts, des cercles montrant approximativement la forme du voisinage de krigeage. Ces artéfacts sont fréquents quand la petite taille du voisinage ne permet pas de sélectionner un nombre suffisant de données pour le krigeage. Il nous faudra donc choisir entre un voisinage donnant une validation croisée un peu moins mauvaise et un voisinage permettant d'éviter les artéfacts. Pour comparaison, la figure 21 montre l'estimation faite à partir du voisinage de 200 m de rayon.

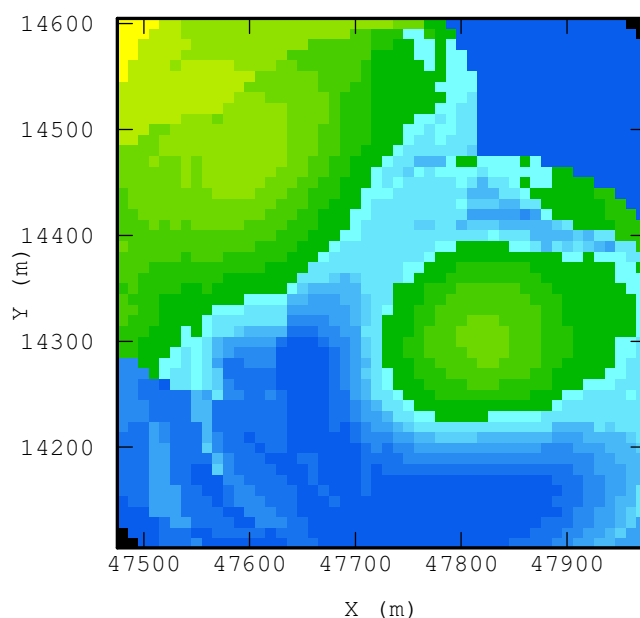


Figure 21. Estimation de la concentration en HAP sur les mailles de $10 \times 10 \times 3 \text{ m}^3$. Utilisation du voisinage de 200 m.

Les artéfacts liés à la forme du voisinage ne sont plus présents qu'en bordure de zone, en extrapolation. Ils disparaîtraient si on se limitait à la zone de mesure. Mais la carte par contre est beaucoup plus lissée, probablement trop. Cette fois-ci, en se limitant aux mailles où l'estimation a été effectuée précédemment (pour éviter une extrapolation excessive) il n'y a plus que 26 mailles où le minimum de l'intervalle de confiance conventionnel est supérieur au seuil de contamination (pollution quasi certaine) et toujours aucune maille où le maximum de l'intervalle de confiance conventionnel est inférieur au seuil de contamination (non pollution quasi certaine). Il faut remarquer que cette carte est nettement différente de la carte obtenue en utilisant l'altitude comme verticale (chapitre 4.1.4). En particulier, la partie extrapolée est très différente. Nous retrouvons par contre sur les deux cartes les deux zones de concentration élevées au nord-ouest et au centre-est.

4.2 Etude des données non ponctuelles (HAP)

Si on suppose que les longueurs théoriques des données non ponctuelles correspondent bien aux longueurs réelles des échantillons, il est nécessaire d'effectuer une régularisation pour les utiliser. Là encore, pour pouvoir comparer aux résultats du chapitre 4.1, nous avons effectué une estimation ponctuelle.

Si on suppose que les échantillons longs correspondent à des échantillons homogènes, on peut construire des échantillons fictifs (composites) de longueur assez faible, en regroupant les échantillons courts et divisant les échantillons longs. La régularisation effectuée sur des tronçons de 50°cm abouti aux résultats de la figure 22. Les variogrammes expérimentaux (ligne du haut) sont d'assez mauvaise qualité et l'ajustement n'est pas très bon. La carte d'estimation (en bas à gauche) est très différente de celle obtenue à partir des données ponctuelles. Cette différence est bien entendu due à l'implantation très différente des échantillons, avant la différence entre les modèles.

Si on suppose que les échantillons longs ne sont pas homogènes, il faudra effectuer la régularisation sur des tronçons longs et on n'obtiendra pas assez de composites pour calculer les variogrammes expérimentaux.

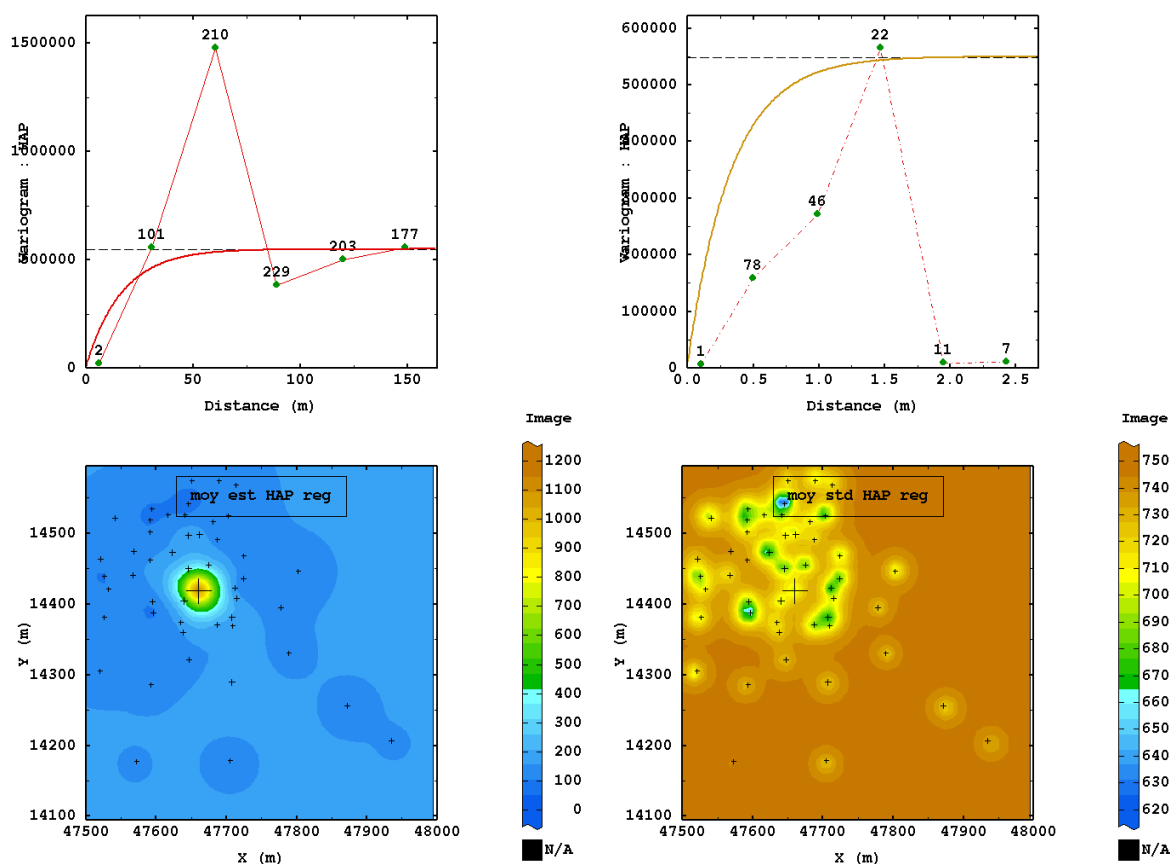


Figure 22. Echantillons « non ponctuels ». Variogrammes expérimentaux et ajustement des HAP régularisés par tronçons de 50cm (haut). Carte des moyennes verticales de l'estimation et des moyennes cumulées de l'écart type de krigeage des HAP (bas)

4.3 Utilisation de toutes les données (HAP).

4.3.1 Estimation ponctuelle

Si tous les échantillons sont de support identique, ou tout au moins très proche, il est possible de les regrouper pour le calcul des variogrammes et l'estimation. Dans un premier temps, nous avons effectué une estimation ponctuelle, le problème de support des échantillons étant toujours présent. Cela nous permet de plus de comparer avec les estimations précédentes.

La figure 23 montre le résultat de ce regroupement. Le variogramme expérimental est encore d'assez mauvaise qualité (difficile à ajuster).

La carte d'estimation ne présente plus la bordure de valeurs fortes qu'on obtenait pour l'estimation à partir des valeurs ponctuelles seules, et montre une petite zone contaminée au sud-est qui n'était pas détectable à partir des valeurs non ponctuelles. Elle semble donc de meilleure qualité, mais sans la confirmation de l'hypothèse de travail (toutes les données ont le même support) on ne peut pas aller plus loin dans son analyse. A nouveau, la carte d'écart-type n'est fiable que si le modèle est correct.

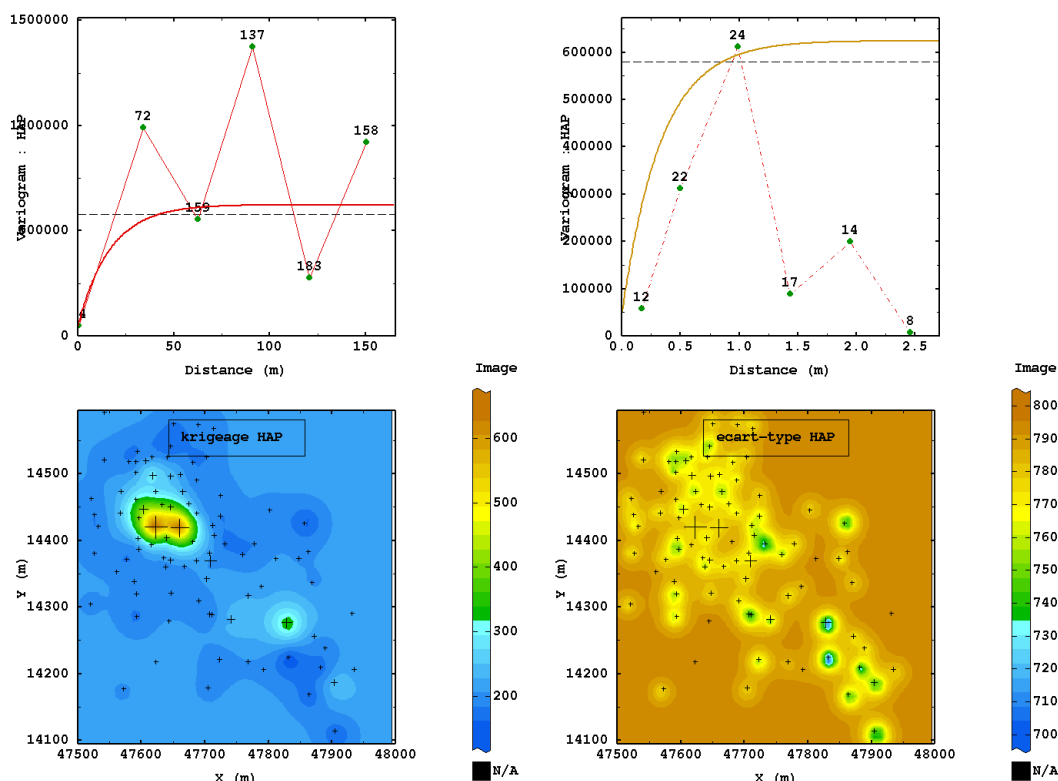


Figure 23. Totalité des données. Variogrammes expérimentaux des HAP et ajustement (haut). Carte des moyennes verticales de l'estimation et des moyennes cumulées de l'écart type de krigeage des HAP (bas).

Remarquons que l'écart-type de krigeage est partout plus grand que le maximum de l'estimation, et plus grand que le seuil de contamination. Si on construit à partir de ces résultats l'intervalle de confiance conventionnel $[z_k - 2\sigma_k, z_k + 2\sigma_k]$ (z_k représente la valeur estimée pour un point donné, σ_k l'écart-type de krigeage au même point), la borne minimale sera partout négative et la borne maximale partout plus grande que le seuil de contamination : il est impossible de dire à partir de ces résultats si le sol a une probabilité élevée d'être ou non contaminé.

4.3.2 Utilisation de la profondeur comme coordonnée verticale

A nouveau, nous avons utilisé la profondeur (affectée d'un signe négatif) comme coordonnée verticale.

Les variogrammes expérimentaux de la concentration en HAP et de son logarithme translaté sont de mauvaise qualité, et seront très difficile à modéliser (Figure 24). Par contre l'indicateur du quantile 50% présente des variogrammes expérimentaux satisfaisants (Figure 25, haut).

Tableau 16. Description du modèle de variogramme de la concentration en HAP (toutes les données, coordonnée verticale : profondeur)

	Type de modèle	Palier	Portée horizontale	Portée verticale
Structure 1	Effet de pépite	34 500		
Structure 2	Sphérique	72 7000	95m	3m

A partir de ce modèle de l'indicateur du quantile 50 %, nous avons simplement modifié les paliers pour proposer un ajustement de la concentration en HAP (Figure 25, bas). Les variogrammes expérimentaux de cette concentration étant de mauvaise qualité, les ajustements sont également plutôt mauvais. Nous avons cependant conservé ce modèle pour l'estimation, étant donné qu'il est à peu près compatible avec les variogramme expérimentaux. Les paramètres du modèle sont donnés dans le tableau 16.

A nouveau, la validation croisée donne un résultat médiocre (voir tableau 17). On peut noter que cette fois-ci, le voisinage de 200 m de rayon donne des résultats légèrement meilleurs, dans un ensemble qui reste mauvais. Comme ce grand voisinage permet d'éviter les artefacts graphiques nous le conserverons, mais le résultat sera probablement trop lissé (moins bonne prise en compte des moyennes locales)

Il reste à discrétiser le bloc. Le calcul de la covariance moyenne sur le bloc est moins stable, et nous a conduits à une discrétisation de 12 x 12 x 6, ce qui est une discrétisation très fine (864 points).

Tableau 17. Résultats de la validation croisée

	Voisinage 75 m	Voisinage 200 m
Nombre de données non robustes 1	<u>10</u>	11
Coefficient de corrélation valeur vraie / valeur estimée 1	0,036	<u>0,069</u>
Coefficient de corrélation erreur standardisée / valeur estimée 0	0,524	<u>0,514</u>
Moyenne de l'erreur 0	19,64	<u>15.16</u>
Variance de l'erreur	802 500	<u>794 300</u>
Variance de l'erreur standardisée 1	<u>2,29</u>	2,43

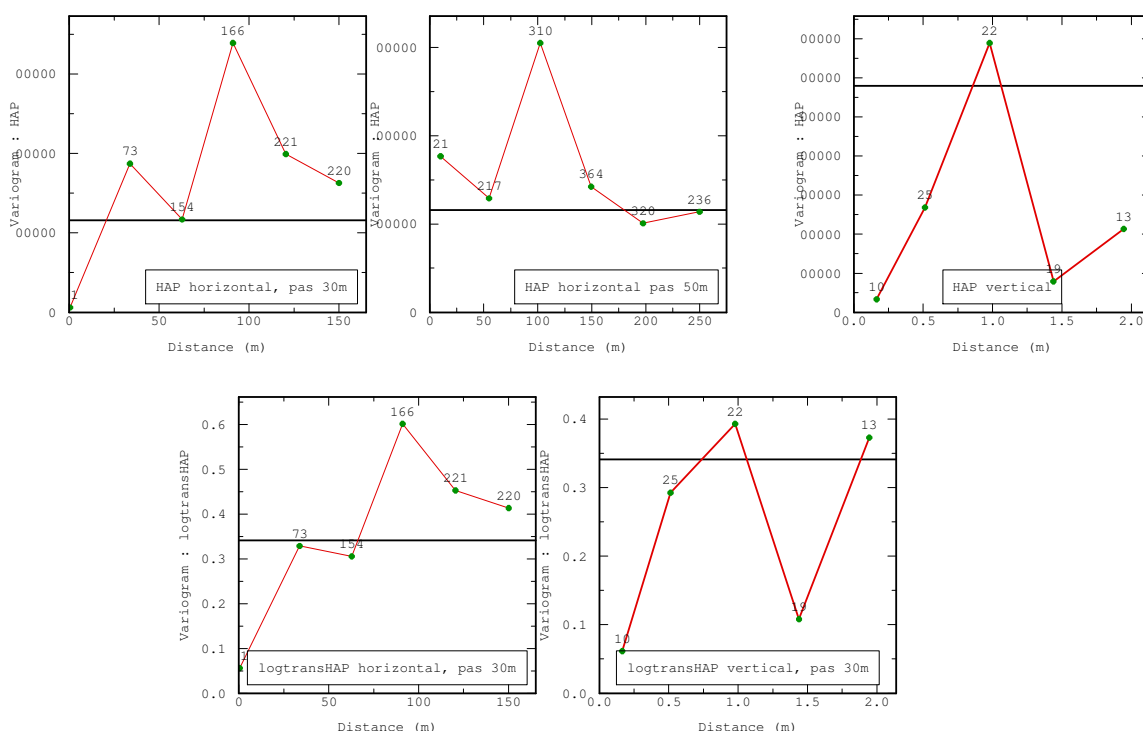


Figure 24. Variogrammes expérimentaux de la concentration en HAP (haut) et de son logarithme translaté (bas).

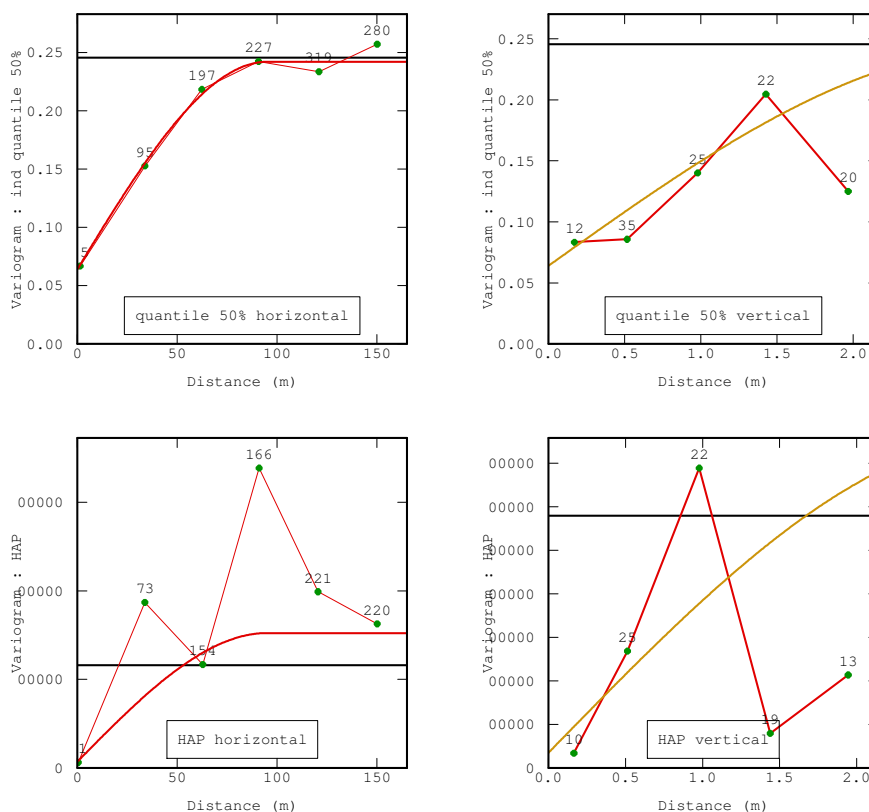


Figure 25. Variogrammes ajustés de l'indicatrice du quantile 50% des concentrations en HAP (haut) et des concentrations en HAP (bas).

La figure 26 montre la carte de l'estimation des blocs de $10 \times 10 \times 3 \text{ m}^3$. Le seuil de contamination correspond au contact bleu-vert. On retrouve les deux taches contaminées qu'on voyait sur l'estimation précédente, mais elles sont plus étendues. L'écart type d'estimation est presque partout supérieur au seuil sauf en quelques mailles situées dans des endroits où les données sont nombreuses.

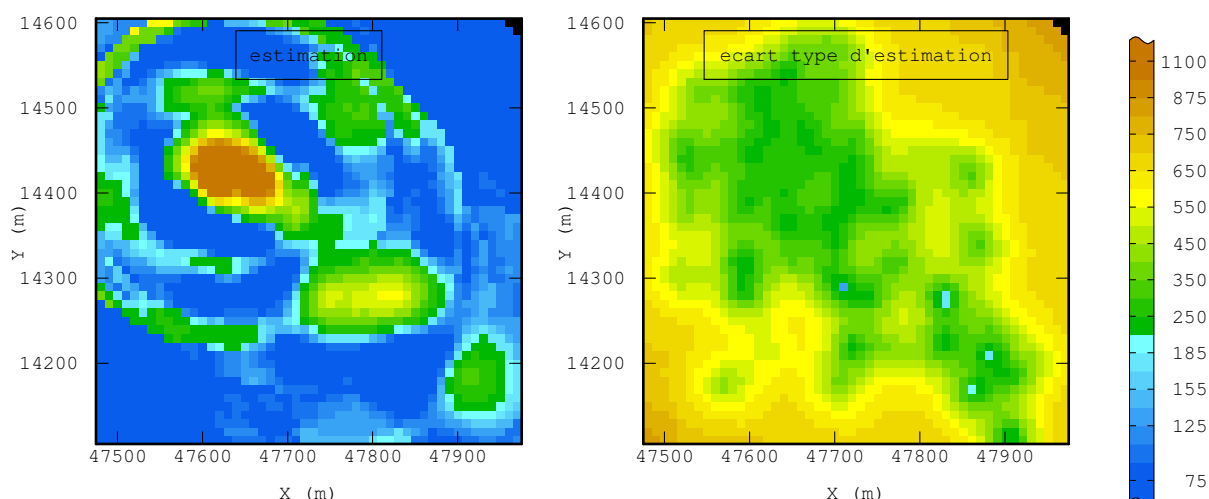


Figure 26. Carte d'estimation (à gauche) et d'écart type d'estimation (à droite)

La carte d'estimation montre là encore des artéfacts montrant la forme du voisinage dans les zones où l'estimation correspond à une extrapolation.

Si on se limite à la zone estimée précédemment avec le voisinage de 75 m, il n'y a à nouveau pas de maille où le maximum de l'intervalle de confiance conventionnel soit inférieur au seuil de contamination : pas de maille très probablement non polluée. Par contre, il y a 1 207 mailles sur 1 634, soit environ 74 %, où le minimum de l'intervalle de confiance conventionnel est supérieur au seuil de contamination (mailles très probablement polluées).

Pour aller plus loin, il faudrait utiliser d'autres méthodes de géostatistique, comme les simulations. Ces dernières sont très sensibles à la qualité de la modélisation, du variogramme mais aussi de l'histogramme des données. Nous avons vu que nous n'avons pas confiance dans la modélisation du variogramme. Des résultats de simulation seraient encore plus sujets à caution que les résultats du krigeage, et nous avons estimé inapproprié d'utiliser cette méthode pour ce jeu de données particulier.

4.4 Etude du plomb et du zinc

Pour les variables plomb et zinc, nous devrions également nous poser la question du support des données. Malheureusement, le jeu de données est trop petit pour nous permettre de le séparer en deux parties. Nous ferons donc l'hypothèse que tous les échantillons ont le même support, pour pouvoir utiliser toutes les données en même temps.

4.4.1 Variogrammes et ajustement

Les variables plomb et zinc présentant une corrélation élevée ($\rho = 0,84$) nous souhaitons en tenir compte pour la modélisation. Quand la corrélation est très élevée, il n'est pas forcément utile d'effectuer une modélisation multivariable lorsque chaque échantillon d'une variable contient également les informations de l'autre variable. Ici, la modélisation multivariable reste utile car il y a plus de mesures de concentrations de plomb que de concentrations de zinc. Les mesures de plomb pourront être utilisées comme données auxiliaires pour l'estimation du zinc. Le coefficient de corrélation est élevé, mais $R^2 = 0,84^2 = 0,70$. Chaque variable explique 70% de la variabilité de la variable associée, et il reste 30% à modéliser individuellement. Cela sera directement pris en compte par la modélisation multivariable, fournissant une meilleure cohérence de l'estimation des variables que dans le cas de deux krigeages séparés, même lorsque tous les échantillons du voisinage sont identiques. D'autre part, il y a moins d'échantillons de zinc que de plomb, il est donc utile d'utiliser les échantillons de plomb là où les échantillons de zinc manquent (en particulier si l'échantillonnage est peu dense aux environs des données manquantes).

En plus du variogramme, la modélisation multivariable nécessite l'emploi du variogramme croisé, qui traduit le lien spatial entre les deux variables. La figure 27 montre les variogrammes et variogramme croisé du plomb et du zinc et leur ajustement dans le plan horizontal. Là encore, le modèle choisi est « optimiste », avec un effet de pépité réduit. Il aurait été possible de choisir un modèle composé uniquement d'un effet de pépité : vu la grande irrégularité des variogrammes expérimentaux, aucun choix ne s'impose.

La figure 28 montre les variogrammes et le variogramme croisé dans la direction verticale, ainsi que leur ajustement. Il faut noter que, dans cette direction verticale, le nombre de paires de points est si faible qu'on ne peut avoir aucune confiance dans cet ajustement. Par contre, il n'est pas possible de mélanger les deux directions, les variogrammes sont si différents qu'un variogramme omnidirectionnel qui mélange tout serait principalement modélisé par un effet de pépité : les cartes qui en découlent sont presque plates, et ne reflèteraient essentiellement qu'une moyenne locale dans le voisinage de krigeage.

Les variogrammes étant assez irréguliers, il n'a pas été possible d'effectuer un ajustement parfait. Dans un modèle multivariable, tous les variogrammes et variogrammes croisés sont liés. Il est nécessaire d'utiliser une procédure semi-automatique pour l'ajustement, afin de préserver ce lien.

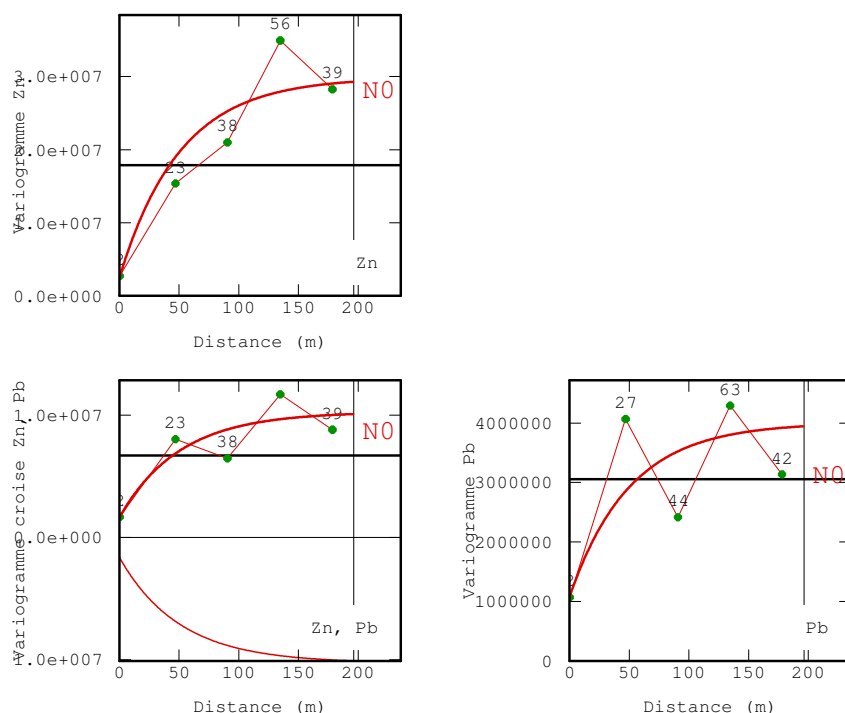


Figure 27. Variogrammes expérimentaux et ajustement du plomb (en bas à droite), du zinc (en haut) et variogramme croisé (en bas à gauche) dans le plan horizontal.

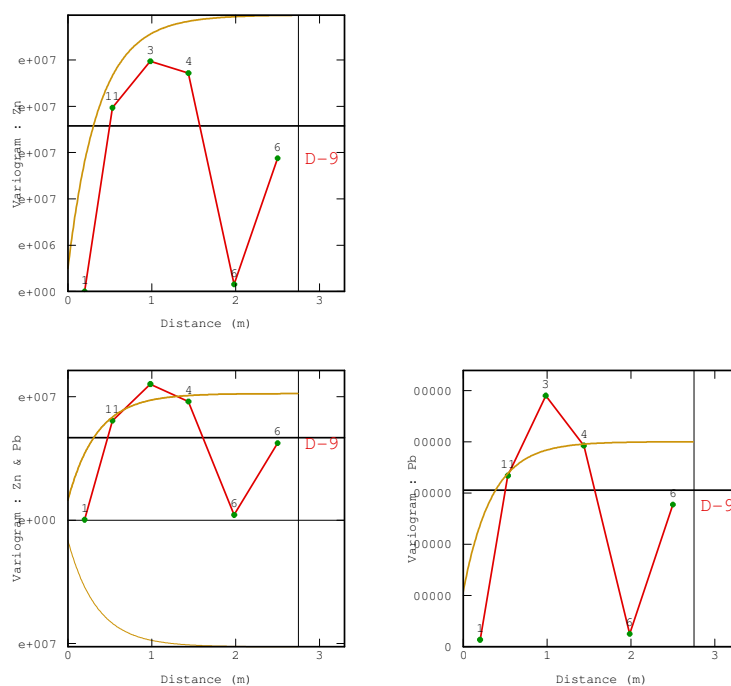


Figure 28. Variogrammes expérimentaux et ajustement du plomb (en bas à droite), du zinc (en haut) et variogramme croisé (en bas à gauche) dans la direction verticale.

4.4.2 Validation croisée multivariable

La validation croisée obéit aux mêmes principes que la validation croisée monovariée, mais un choix supplémentaire doit être fait : lorsqu'on enlève la donnée à ré-estimer du jeu de données, enlève-t-on seulement la variable cible ou enlève-t-on aussi la variable auxiliaire ? Le premier choix donne de meilleurs résultats de validation croisée, mais ceux-ci ne sont représentatifs de la qualité de l'estimation fournie par le modèle qu'aux environs immédiats des points où la variable auxiliaire est présente alors que la variable cible est absente. Dans notre cas, cela ne concerne que l'estimation du zinc au voisinage des points où seul le plomb a été mesuré. Lorsque l'on est loin de ces points ou lorsque l'on considère l'estimation du plomb, c'est la validation croisée où l'on enlève les deux variables qui doit être prise en compte.

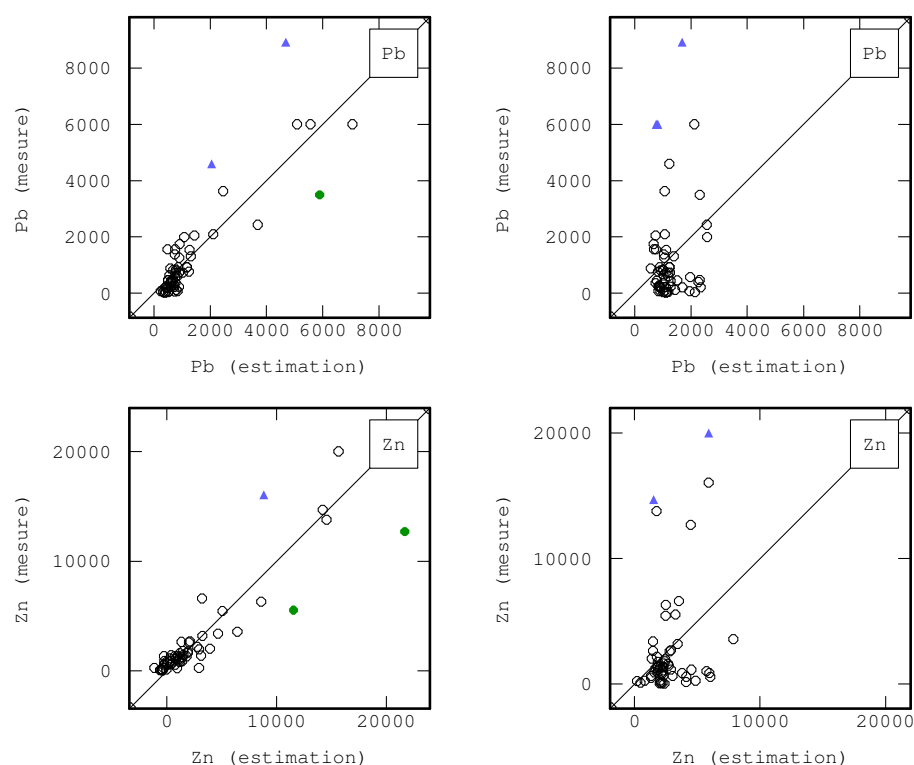


Figure 29. Validation croisée multivariable. Haut : plomb ; bas zinc. Gauche : avec utilisation de la variable auxiliaire au point à estimer ; droite : sans utilisation de la variable auxiliaire au point à estimer. Triangles bleu : valeurs fortement sous-estimées ; ronds vert : valeurs fortement surestimées.

Sur la figure 29, on voit que si on enlève seulement la variable cible (partie de gauche) les résultats de la validation croisée sont très bons, en particulier les points des nuages de corrélation valeur vraie – valeur estimées sont distribués autour de la première bissectrice (biais conditionnel absent ou faible). Seuls trois points présentent une erreur trop forte pour être bien expliquée par la variabilité du modèle (points verts et bleus). Si on enlève les deux variables, les points ne sont plus distribués autour de la première bissectrice (biais conditionnel plus élevé), mais il n'y a que deux points qui présentent une erreur trop forte pour le modèle. Cela pourrait signifier que l'erreur d'estimation s'explique bien par la variabilité du modèle, mais le critère de la variance de l'erreur standardisée nous dit le contraire : celle-ci n'est jamais proche de 1, soit nettement trop forte quand on conserve la variable auxiliaire au point cible, soit nettement trop faible quand on l'enlève. Comme on pouvait s'en douter au vu de l'ajustement approximatif des variogrammes, (en particulier pour l'ajustement du zinc), le modèle représente mal les données.

4.4.3 Cartographie : estimation par cokrigage

La méthode d'estimation multivariable proposée en géostatistique est le cokrigage (extension du krigeage au cas multivariable). La figure 30 montre à gauche les cartes de concentration estimées pour le plomb (en haut) et le zinc (en bas), et à droite les cartes d'écart-type correspondant.

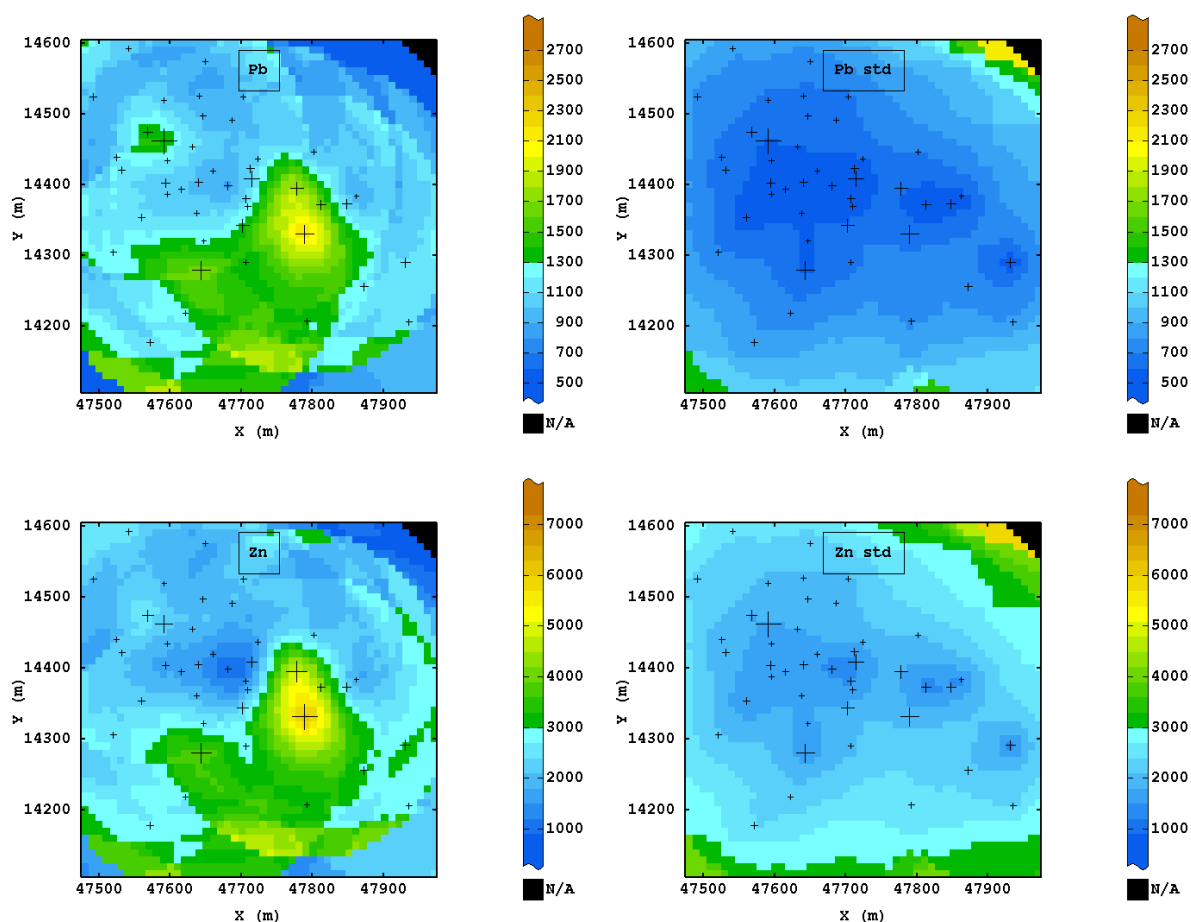


Figure 30. Cartes des estimations (gauche) et écarts types de krigeage (droite) pour le plomb (haut) et le zinc (bas).

Les deux variables étant fortement corrélées, les deux cartes sont très semblables. Pour toutes les cartes, le seuil de contamination correspond à la limite cyan/vert. Bien que les valeurs des écarts types de krigeage soient inférieures aux seuils de contamination (1300 ppm pour le plomb, 3000 ppm pour le zinc) sauf dans les zones extrapolées, l'intervalle de confiance conventionnel englobe pour tous les points à la fois des concentrations correspondant à une contamination et une concentration correspondant à une non-contamination. On ne peut donc détecter ni zone très probablement polluée, ni zone très probablement non polluée à partir de ces cartes. Là encore, il faudrait effectuer des simulations pour calculer des probabilités de dépassement de seuil. Mais avec une aussi grande incertitude sur l'ajustement du variogramme les résultats obtenus seraient de qualité douteuse, ces simulations n'ont donc pas été effectuées.

5. ETUDE DU LOGARITHME TRANSLATE DE LA CONCENTRATION EN HAP

Lorsque la variable d'intérêt présente un histogramme extrêmement dissymétrique comme c'est le cas ici, il peut être judicieux d'étudier le logarithme de cette variable, et de calculer les intervalles de confiance à partir du logarithme : on peut calculer les bornes de l'intervalle de confiance conventionnel sur le logarithme et en calculer ensuite les exponentielles pour revenir à la variable initiale. Cette opération peut se faire soit sur le logarithme de la variable étudiée, soit sur son logarithme translaté (utile lorsqu'il y a beaucoup des valeurs nulles, le logarithme de 0 n'étant pas défini).

Il faut remarquer que si la variable étudiée est une variable additive (la valeur moyenne de la réunion de deux blocs de dimensions identiques est la moyenne des valeurs des deux blocs) leur logarithme (ou logarithme translaté) ne l'est plus. Il ne sera alors plus possible d'effectuer une estimation de bloc (le logarithme moyen sur le bloc n'est pas le logarithme de la moyenne de la variable initiale sur le bloc) mais seulement une estimation ponctuelle (sauf hypothèses particulières).

Si le logarithme présente une distribution bigaussienne (les valeurs au point x et au point y présentent une loi de probabilité bivariable gaussienne, et ce quels que soient les points x et y) il sera également possible d'effectuer un krigeage log normal : non seulement on effectue le krigeage du logarithme, mais on transforme ensuite les résultats, estimation et variance de krigeage, pour obtenir une estimation par krigeage lognormal et une variance de krigeage lognormal pour la valeur brute. Le krigeage lognormal s'effectue à partir du logarithme naturel, et non du logarithme translaté.

5.1 Estimation du logarithme translaté et du logarithme naturel des HAP.

Cette estimation a été effectuée en regroupant toutes les données. Le problème de support des données se pose toujours, mais dans un but de démonstration il a été choisi de ne pas en tenir compte : toutes les mesures sont supposées effectuées sur le même support ponctuel.

5.1.1 Calcul du logarithme translaté

Pour éviter de donner trop de poids aux valeurs très faibles, leur logarithme ayant une très forte valeur négative, il a été choisi de travailler avec le logarithme translaté. Le logarithme translaté est $\log_{10}(1 + z/m)$, où z est la concentration et m la moyenne des concentrations.

5.1.2 Analyse variographique du logarithme translaté

Les emplacements des données étant évidemment les mêmes que pour la variable non transformée, les paramètres de calcul des variogrammes expérimentaux sont les mêmes : 30 m dans la direction horizontale, 50 cm dans la direction verticale, tolérance sur les distances 0,5 fois le pas dans la direction considérée. La figure 31 présente les variogrammes expérimentaux dans les deux directions, horizontale et verticale, et le modèle ajusté sur ces variogrammes.

Le modèle ajusté est un modèle sphérique, de palier 0,34 et de portées 30 m pour la direction horizontale et 0,7 m pour la direction verticale. Il n'a pas été modélisé d'effet de pépité, nous avons choisi de modéliser le premier point du variogramme expérimental alors que celui-ci ne représente que peu de paires de points. Ce modèle, où les portées sont de l'ordre de grandeur du pas de calcul du variogramme, n'est pas fiable. La mauvaise qualité du jeu de donnée qu'on avait notée sur les variables brutes ne s'améliore pas lors du passage au logarithme. La modélisation aurait pu être faite également avec un effet de pépité pur, peut-être avec des résultats de validation croisée plus cohérents qu'au chapitre 5.1.3. Par contre, l'écart-type de krigeage aurait été plus important. Nous savons que notre modèle est trop optimiste en ce qui concerne l'écart type de krigeage, mais nous avons souhaité le conserver pour calculer l'intervalle de confiance conventionnel. Un modèle optimiste donne-t-il des résultats sélectifs ou non ?

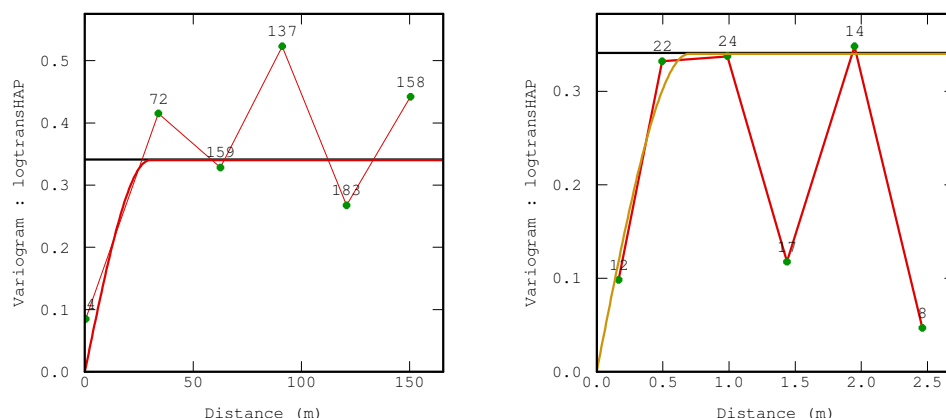


Figure 31. Variogrammes expérimentaux et modèle du logarithme translaté des HAP. Gauche : variogramme horizontal ; droite : variogramme vertical.

5.1.3 Validation croisée

Le nuage de corrélation (Figure 32) montre là encore que l'estimation est mauvaise : le coefficient de corrélation, 0,087, est très faible. Il y a des valeurs fortes très sous-estimées (ronds verts) et quelques valeurs nettement surestimées carrés rouges). A noter qu'un échantillon fortement sous-estimé peut se trouver dans le même sondage qu'un échantillon fortement sur estimé (flèche sur la carte de droite de la figure 32). C'est en général dû à des échantillons proches de mesure très contrastée, chacun influant sur la modélisation de l'autre et perturbant donc son estimation. Le coefficient de corrélation est un peu plus élevé que pour le krigeage de la valeur brute, mais avec des valeurs aussi faibles ce n'est pas significatif.

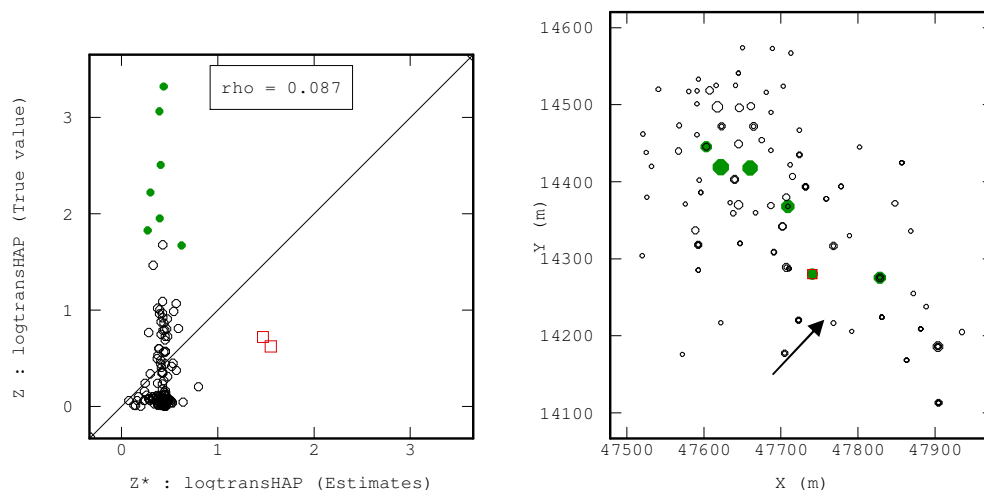


Figure 32. Validation croisée du logarithme translaté des HAP. Gauche : nuage de corrélation de la valeur vraie en fonction de la valeur estimée. Droite : carte des données, montrant les valeurs fortement sous estimées (ronds verts) et sur estimées (carrés rouges).

5.1.4 Estimation et calcul de l'intervalle de confiance conventionnel. Dépassement de seuil.

Une fois le krigeage effectué, l'intervalle de confiance conventionnel $[z_k - 2\sigma_k, z_k + 2\sigma_k]$ est calculé (z_k représente la valeur estimée pour un point donné, σ_k l'écart-type de krigeage au même point).

Le seuil de contamination de 200 ppm correspond à une valeur de $\log_{10}(1+200/243,37) = 0,2604$ soit 0,26 en première approximation. Nous comparerons les bornes de l'intervalle de confiance conventionnel calculé à partir de l'estimation du logarithme translaté à cette valeur de 0,26.

A nouveau, le maximum de l'intervalle de confiance conventionnel n'est jamais inférieur au seul de contamination. Tous les points peuvent être contaminés. Ils le seront avec quasi-certitude (probabilité > 97,5% dans le cas d'une erreur de distribution gaussienne) si le minimum de l'intervalle de confiance conventionnel est supérieur à 0,26. C'est le cas pour 15 points dans notre grille. A noter que 7 de ces 15 points se trouvent dans le niveau le plus superficiel. Alors que le modèle est trop optimiste (il propose une précision illusoire), il ne permet malgré tout pas de mettre en évidence des zones très probablement soit polluée soit non polluée.

Si on fait le bilan, 48 points dans la grille ne sont pas estimés (avec notre voisinage de krigeage de 200 m de diamètre, ils sont trop loin des points de données), 15 points ont une estimation qui montre une forte probabilité de contamination (cette probabilité est d'au moins 97,5 % si l'erreur d'estimation présente une distribution gaussienne) et 29 937 points ont une estimation pour laquelle l'intervalle de confiance conventionnel ne permet pas de trancher, bien qu'on ait ajusté un modèle optimiste qui fournit un intervalle de confiance conventionnel moins large. Si on avait la certitude qu'en tout point, l'erreur d'estimation présente une distribution gaussienne, il nous serait possible de montrer des points contaminés ou non avec une probabilité connue mais différente. Malheureusement, nous ne connaissons pas la distribution de cette erreur.

A nouveau nous n'avons pas souhaité poursuivre avec des simulations étant donné l'incertitude sur le modèle. D'autant que le modèle le plus adapté aux données est peut-être l'effet de pépite pur, pour lequel les simulations présentent un aspect de bruit aléatoire.

5.2 Krigeage lognormal ? Etude de la distribution du logarithme naturel des HAP

Si la distribution du logarithme naturel de la variable étudiée est multigaussienne il est possible d'effectuer un krigeage lognormal.

Ce krigeage lognormal consiste à effectuer le krigeage du logarithme naturel de la variable, puis effectuer une transformation du résultat et un calcul de l'écart-type relatif de krigeage, en corrigeant le biais apporté par la transformation inverse « exponentielle ». Il est important de bien utiliser le logarithme naturel et non pas le logarithme de base 10, ou la transformation du résultat serait fausse.

Il est à peu près impossible de prouver qu'une variable présente une distribution multigaussienne (distribution gaussienne multivariable quand on prend en compte la variable en plusieurs points du domaine d'étude). Mais pour être multigaussienne, la variable doit tout d'abord présenter une distribution monovariée gaussienne. Cette hypothèse peut se vérifier (ou s'invalider) à partir de plusieurs méthodes. Par exemple en vérifiant que l'histogramme présente apparemment la forme « en cloche » bien connue (ou au moins qu'il est approximativement symétrique. Il est également possible de comparer les quantiles de la distribution expérimentale du logarithme naturel aux quantiles théoriques d'une distribution gaussienne de même moyenne et même variance (« Q-Q plot »).

Pour le logarithme naturel des HAP, la figure 33 montre l'histogramme du logarithme naturel des données et la courbe de comparaison Quantile-Quantile.

Nous voyons que l'histogramme n'est pas vraiment symétrique, et que les quantiles de la distribution du logarithme naturel des HAP ne sont pas ceux d'une distribution gaussienne (la courbe n'est pas une ligne droite). Cela pourrait être dû à un nombre de données trop faible qui entraîne des fluctuations statistiques importantes. Il est possible d'effectuer des tests statistiques disant si la distribution est compatible ou non avec une hypothèse de distribution gaussienne en tenant compte du nombre d'échantillons (comme le test du χ^2). Mais même si la réponse est positive, pour faire un krigeage lognormal il faut avoir plus de certitude sur la distribution gaussienne du logarithme. Etant donné les mauvais résultats des tests de la figure 33, ce krigeage n'est pas recommandé dans notre cas.

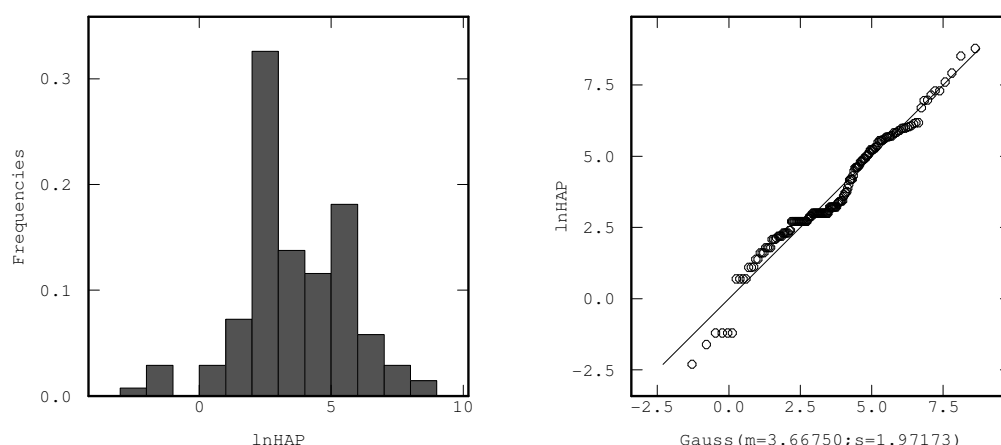


Figure 33. Droite : histogramme du logarithme naturel des HAP. Gauche : comparaison Quantile-Quantile avec une distribution gaussienne de même moyenne et de même variance.

6. ETUDE DU LOGARITHME TRANSLATE DU PLOMB ET DU ZINC

Les logarithmes traduits des concentrations en métaux lourds ont également été étudiés. La figure 34 montre que le coefficient de corrélation, 0,92, est encore plus élevé que pour les concentrations brutes. En effet, l'effet des valeurs extrêmes est atténué et elles déforment moins le nuage de corrélation.

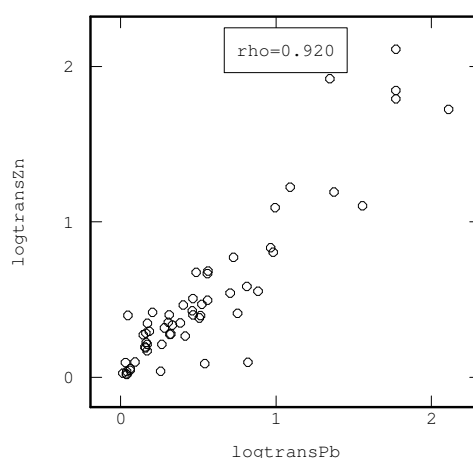


Figure 34. Nuage de corrélation entre les logarithmes traduits du Plomb et du Zinc.

Par contre, les variogrammes expérimentaux sont de mauvaise qualité (Figures 35 et 36). Très erratiques, ils imposent d'ajuster un modèle avec un fort effet de pépité. Cela signifie que l'estimation sera très peu discriminante, les zones de concentration faible ou de concentration élevée seront mal distinguées, et les variances d'estimation seront partout élevées. L'estimation du logarithme traduit apportera donc probablement peu d'amélioration, voire aucune.

Effectivement, pour le plomb, lorsqu'on effectue le krigeage du logarithme traduit et que l'on calcule les intervalles de confiance conventionnels, le seuil transformé égal à 0,31 est toujours dans l'intervalle de confiance. Le krigeage du logarithme traduit n'a apporté aucune amélioration par rapport au krigeage de la concentration brute.

Pour le zinc, un seul point donne un minimum de l'intervalle de confiance conventionnel supérieur au seuil transformé égal à 0,32. Peut-on réellement considérer ce résultat comme une amélioration ?

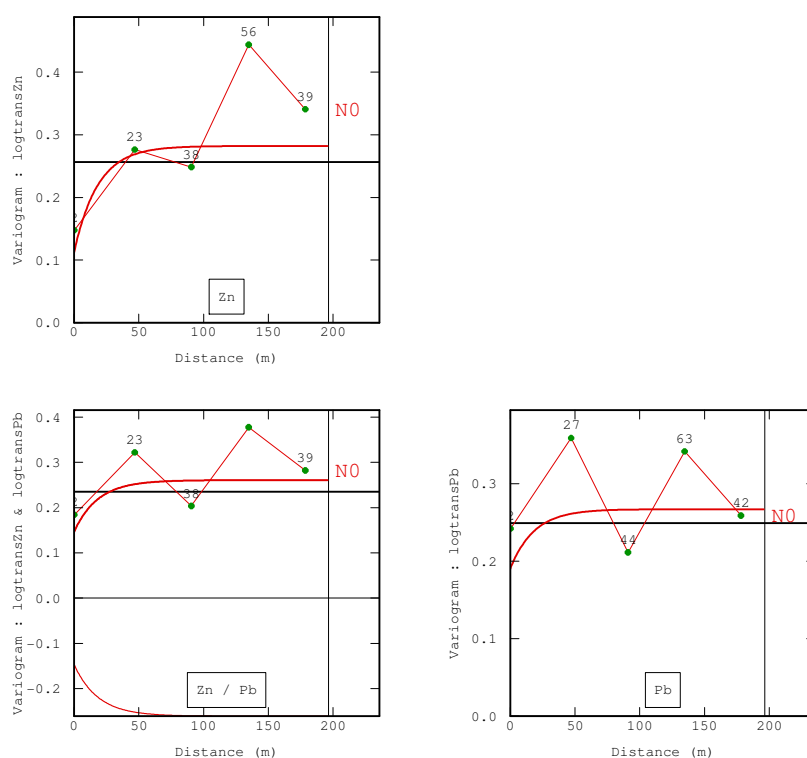


Figure 35. Variogrammes et variogramme croisé horizontal des logarithmes translats du Plomb et du Zinc. Variogrammes expérimentaux et modèle ajusté.

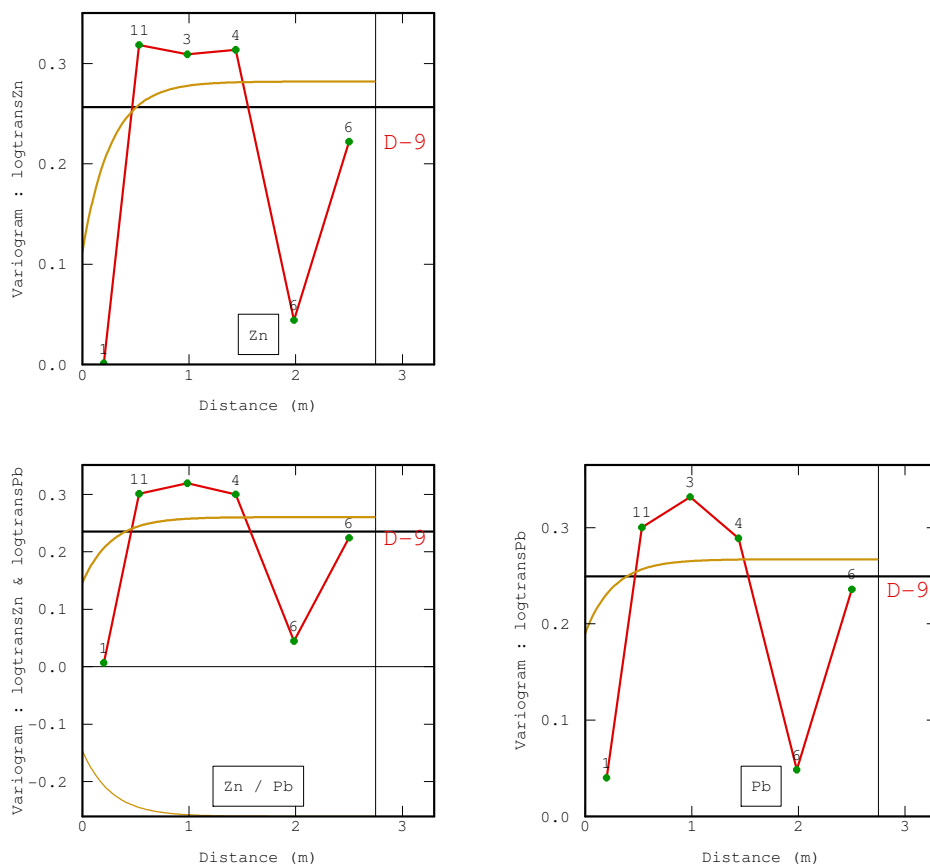


Figure 36. Variogrammes et variogramme croisé vertical des logarithmes translats du Plomb et du Zinc. Variogrammes expérimentaux et modèle ajusté.

7. VALIDATION SUR LES DONNEES NOUVELLEMENT FOURNIES

Une fois l'étude décrite dans les chapitres précédents terminée, Arcelor-Mittal a transmis de nouveaux résultats d'analyse de la couche superficielle du sol pouvant servir de données de validation. Nous n'avons pas tenté cette validation. Outre la date très tardive d'arrivée de ces données, elles ne pourraient servir qu'à valider ou invalider un modèle dont nous savons déjà qu'il est probablement très mauvais, en raison de la qualité non adaptée de l'échantillonnage et de l'extrême hétérogénéité du terrain. Si la validation nous donne un résultat favorable ce sera très probablement par chance, si elle nous donne un mauvais résultat il est déjà prévisible au vu de l'analyse présentée dans ce rapport.

La validation n'a donc pas été tentée.

8. CONCLUSION

L'étude géostatistique du site T n'a pas pu être menée à bien de façon correcte en raison d'un manque d'information sur les méthodes d'échantillonnage et le support des échantillons. Des décisions fondamentales pour les calculs à mettre en œuvre en dépendent. Il n'est pas possible de mélanger sans précaution des échantillons de support différent. L'hypothèse de support constant pour les données sans indication de profondeur minimum et maximum est inévitable, mais invérifiable en l'absence de ces informations. Lorsque ces indications de profondeur minimum et maximum sont présentes, elles ne sont utilisables que si on considère que l'échantillon a bien été prélevé sur l'intégralité de la hauteur, ce qui est assez peu vraisemblable au vu de l'hétérogénéité des fosses. Et même dans ce cas, la prise en compte de l'hétérogénéité interne des échantillons entraînerait une telle diminution du nombre d'échantillons lors de la construction des échantillons composites que le nombre de données utilisables deviendrait insuffisant pour une étude géostatistique.

Si le but de cette étude avait été réellement de planifier et guider une dépollution, la mauvaise qualité des données aurait probablement conduit à l'abandon ou à la suspension temporaire de l'utilisation de la géostatistique, éventuellement en l'attente d'une reprise de l'échantillonnage ou d'une ré-étude détaillée des échantillons déjà existants et de l'historique des prélèvements. L'âge déjà ancien des campagnes d'échantillonnage n'a pas permis de les réétudier, en particulier d'interroger les personnes qui étaient en charge de ces campagnes. L'étude géostatistique a été malgré tout poursuivie, pour une démonstration des conséquences néfastes d'un mauvais échantillonnage.

La seule hypothèse de travail utilisable en l'état actuel des informations est de supposer que tous les échantillons ont été prélevés avec le même support, hypothèse invérifiable donc. Cette hypothèse nous permet de regrouper toutes les données pour le calcul des variogrammes expérimentaux et d'obtenir un modèle. Les cartes alors obtenues auraient pu conduire à une première délimitation des zones décontaminées à dépolluer. Malheureusement, l'écart-type de krigeage toujours supérieur au seuil de contamination oblige à considérer la totalité de la zone comme contaminée, même quand l'estimation donne une valeur nettement inférieure à ce seuil. Pour pallier ceci, l'estimation a été refaite à partir du logarithme translaté des HAP. Cette variable plus symétrique et présentant des valeurs moins extrêmes permet habituellement une modélisation géostatistique plus aisée. Effectivement, l'écart-type d'estimation n'est plus partout supérieur au seuil transformé, mais la construction de l'intervalle de confiance conventionnel en chaque point montre que, si on détecte bien quelques-unes des zones très probablement contaminées, on ne peut toujours pas mettre en évidence de zones très probablement non polluées. Le plus souvent, on reste dans l'incertitude. Des méthodes de géostatistique plus avancées (simulations par exemple) pourraient peut-être permettre de lever cette incertitude, et en tout cas de mieux déterminer les intervalles de confiance locaux.

Il existe pour ce site T des données supplémentaires que nous souhaitons utiliser pour la validation des modèles et de la méthode. Nous avons reçu ces données (couche superficielle du sol) peu de temps avant la remise du présent rapport. Il nous a paru plus intéressant de procéder à des démonstrations méthodologiques complémentaires plutôt que de procéder à une validation dont les résultats, qu'ils soient bons ou mauvais, ne peuvent pas être significatifs.

Bien que l'étude géostatistique du site T soit décevante, elle attire l'attention sur un point préalable extrêmement important : la qualité de l'échantillonnage. Les méthodes d'échantillonnages doivent dépendre des buts à atteindre. Il n'est sans doute pas fondamental de prélever des échantillons de support identiques quand on veut simplement détecter la présence ou l'absence de contaminants dans le sol, alors que c'est indispensable si on veut ensuite utiliser les mesures à l'aide de la géostatistique dans un but de cartographie ou d'évaluation d'un volume à dépolluer. Il faut au moins être capable d'homogénéiser le support des échantillons par le calcul (régularisation) s'ils ne sont pas identiques, et cela impose des contraintes particulières lors du prélèvement : échantillons contigus le long des sondages, volumes des échantillons connus.

D'un point de vue méthodologique, le fait que les modèles calculés directement sur les données brutes soient de mauvaise qualité nous a conduit à proposer des méthodes pouvant aider à la modélisation : modélisation du logarithme, de l'indicatrice du quantile 50%, utilisation d'une variable auxiliaire dans le cas du plomb et du zinc. La validation croisée permet en outre de savoir si des résultats apparemment décevants peuvent s'expliquer par la variabilité du modèle ou si c'est le modèle lui-même qui n'est pas adapté. Enfin, si les résultats du krigeage, toujours trop lisses, ne permettent pas de calculer les probabilités de dépassement de seuil, nous avons montré comment ils pouvaient éventuellement permettre, grâce au calcul de l'intervalle de confiance conventionnel, d'avoir une première idée des zones très probablement contaminées ou très probablement non contaminées. Première idée qui doit de toute façon être confirmée et affinée par des calculs de probabilités de dépassements de seuil en faisant des simulations, mais cette méthode très sensible à la qualité du modèle n'a pas été utilisée ici.

Tableau 18. Principales étapes réalisées pour l'étude de démonstration et pertinence de ces étapes dans le cadre d'une étude géostatistique classique

Étape	Commentaire
Analyse de l'échantillonnage et des informations fournies	Indispensable, elle a conduit ici à de nombreuses remarques sur les spécificités de l'échantillonnage et a eu des conséquences sur tout le processus de traitement des données. Elle conduit ici sur de gros doutes concernant l'utilisabilité de ces données pour une délimitation correcte de zones contaminées.
Analyse statistique	Elle devrait permettre un contrôle qualité des données et les premiers commentaires sur les niveaux de contamination rencontrés. Dans notre cas, elle conforterait la décision de ne pas poursuivre l'étude sauf à demander une reprise de l'échantillonnage (reprise de l'historique détaillé s'il est accessible, ou nouvelles campagnes de mesure). Les résultats ultérieurs seront trop soumis à caution et inutilisables en pratique.
Modélisation des teneurs en HAP et métaux lourds au sein de cette couche	Modélisation monovariante pour les HAP qui n'est pas corrélée aux autres variables, bivariable pour le Plomb et le Zinc qui sont bien corrélés. A partir de cette étape, la poursuite de l'étude ne vise qu'un projet de démonstration sur ce qui pourrait être fait avec un échantillonnage acceptable. Aucun résultat n'est utilisable dans un cadre opérationnel.
Estimation des concentrations et cartographie de la contamination	Etant donnée la très mauvaise qualité des données et donc des modèles, les résultats de cette étape ne sont pas utilisables dans un cadre opérationnel.
Estimation des volumes contaminés et volumes à excaver, calculs de probabilités de dépassement de seuil.	Cette étape devrait être indispensable, mais les méthodes de mise en pratique reposent sur des hypothèses fortes concernant le modèle et sa qualité qui ne sont pas vérifiées ici. Il a donc été décidé ici de ne pas poursuivre et de ne pas effectuer ces calculs.