Document public

Établissement de fonds pédogéochimiques urbains en parallèle à l'Opération ETS du Ministère de l'Écologie

Rapport Intermédiaire

BRGM/RP-66306-FR
octobre 2016

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

Géosciences pour une Terre durable
brgm

Document public

Établissement de fonds pédogéochimiques urbains en parallèle à l'Opération ETS du Ministère de l'Écologie

Rapport intermédiaire

BRGM/RP-66306-FR
octobre 2016

Étude réalisée dans le cadre de la convention n° 1372C0016 ADEME-BRGM

JF. Brunet
Avec la collaboration de
F. Guiet, E. Taffoureau, B. Bourgine, C. Blanc et L. Sancho

Vérificateur :	**Approbateur :**
Nom : L. Callier	Nom : H. Léprond
Fonction : Responsable scientifique de Programme	Fonction : Responsable de l'unité Sites, Sols et Sédiments Pollués
Date : 18/11/2016	Date : 08/12/2016
Signature :	Signature :

Le système de management de la qualité et de l'environnement est certifié par AFNOR selon les normes ISO 9001 et ISO 14001.

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

Géosciences pour une Terre durable
brgm

Mots-clés : Fond pédogéochimique urbain, FGU, ETM, BDSolU, Gestion des sites et sols pollués, Terres excavées, France.

En bibliographie, ce rapport sera cité de la façon suivante :

Brunet J.-F. avec la collaboration de **Guiet F., Taffoureau E., Bourgine B., Blanc C. et Sancho L.** (2016) - Établissement de fonds pédogéochimiques urbains en parallèle à l'Opération ETS du Ministère de l'Écologie. Rapport intermédiaire. BRGM/RP-66306-FR, 91 p., 4 fig., 2 ann.

Synthèse

La convention n° 1372C0016 entre l'ADEME et le BRGM encadre le projet « Établissement d'un fond géochimique urbain et industriel en parallèle à l'opération ETS » dit FGU, du 12 septembre 2014 au 12 septembre 2017. Les travaux consistent à poursuivre et compléter les tâches entamées au cours de la première convention ADEME-BRGM (2010-2014) :

- étude bibliographique ;
- bancarisation d'analyses de sols urbains « exempts » de pollution directe et répartis sur l'ensemble du territoire national, obtenus dans le cadre du projet « Diagnostic des sols dans les établissements accueillant des enfants et des adolescents » dit ETS ;
- proposition de pistes d'amélioration de la norme NF EN ISO 19258 « Qualité du sol - Guide pour la détermination des valeurs de bruit de fond » ;
- refonte du système de collecte des données et de la base de données pour étendre son spectre d'application ;
- recherche de réponses aux questions méthodologiques, notamment statistiques ;
- détermination de fonds pédogéochimiques pour les différents paramètres analysés dans les principales agglomérations françaises.

L'objectif est d'apporter un appui aux différents acteurs impliqués dans la gestion des sites (potentiellement) pollués et d'améliorer la connaissance de la qualité des sols urbains. Il s'inscrit dans le contexte de la méthodologie nationale de cette thématique précisée par les textes publiés par le Ministère en charge de l'environnement en février 2007 et construits sur un principe de réhabilitation en fonction de l'usage. En l'absence de valeurs réglementaires pour le milieu « sol », cette méthodologie préconise la comparaison de la qualité des sols investigués à celle des sols voisins exempts de tout impact direct et représentatifs du fond pédogéochimique. En milieu rural, les fonds pédogéochimiques anthropisés sont relativement bien connus grâce aux bases de données de l'INRA (Base de données des éléments traces métalliques - BDETM - et Réseau de mesures de la qualité des sols - RMQS). Mais en milieu urbain, il faut tenir compte de contaminations diffuses liées aux activités humaines des villes et des industries. Le projet FGU consiste à déterminer les fonds pédogéochimiques anthropisés spécifiques aux différentes agglomérations urbaines en France. Les connaissances ainsi acquises sur la qualité des sols urbains pourront servir, entre autres, à la mise en œuvre de la démarche de gestion des terres excavées (après adaptation et compléments). Le projet FGU s'appuie sur le projet lancé par le ministère en charge de l'environnement intitulé « Diagnostics des sols dans les lieux accueillant des enfants ou des adolescents » (projet « Établissements sensibles - ETS »).

Le présent rapport rappelle rapidement le fonctionnement du projet et décrit les différentes tâches en cours ou réalisées. En octobre 2016, la base de données FGU compte 73 078 résultats d'analyses correspondant à 632 échantillons de sols « témoins » obtenus dans le cadre du projet ETS. Les premiers essais de traitement géostatistique des données recueillies se heurtent à des difficultés dues aux faibles effectifs des analyses et à la répartition hétérogène des points de prélèvement. Cependant un protocole de traitement statistique a été élaboré dans le cadre d'un stage de fin d'étude d'ingénieur. L'exploitation des données tiendra ainsi compte des faibles effectifs des populations et du taux de valeurs inférieures aux limites de quantification analytiques, d'autant plus quand ces dernières sont élevées. Cette adaptation des méthodes statistiques mises en œuvre aux spécificités et au contexte des données FGU, va permettre une valorisation plus affinée des résultats.

L'alimentation de la base de données avec des analyses complémentaires reste cependant nécessaire pour atteindre les objectifs fixés. La deuxième convention ADEME-BRGM (2014-2017) prévoit donc la refonte de la base de données et du protocole permettant son alimentation. Il s'agit de bancariser des analyses de sols et leurs métadonnées dans le cadre de projets hors ETS. Ces données pourront être obtenues selon des protocoles de prélèvement et d'analyse variés et notamment, concerner des sols profonds. Cette nouvelle base est appelée Base de données des analyses de Sols Urbains (BDSolU). Elle doit être alimentée par les données ETS déjà acquises et celles de plusieurs projets du BRGM en cours ou achevés. L'alimentation de la base au moyen des données recueillies localement par certaines collectivités urbaines est également une solution pour laquelle le BRGM recherche des partenariats.

Enfin, le BRGM a contribué, à travers le projet FGU, à la consultation sur la norme ISO 19258 et aux travaux du Groupe de travail « Bruit de fond » conduit par l'ADEME. Il contribue ainsi à la réflexion en cours sur l'ensemble des questions méthodologiques posées par la détermination des fonds pédogéochimiques anthropisés.

Sommaire

Liste des figures

Liste des annexes

1. Introduction

1.1. CONTEXTE

La France a mis en place une méthodologie de gestion des sites et sols (potentiellement) pollués basée sur la prévention et sur la gestion des risques suivant l'usage pour les pollutions anciennes. Toutefois, notre pays ne s'est pas doté de valeurs guides réglementaires concernant la qualité géochimique de ces sols. En cas de suspicion de pollution, la démarche française privilégie la comparaison de l'état du sol considéré à celui des sols « sains » voisins de la zone d'investigation. Il s'agit de distinguer le **fond géochimique « naturel »** et notamment les anomalies géochimiques locales des contaminations ou des pollutions attribuables aux activités du site (1).

De son côté, la démarche de gestion des terres excavées considère qu'une terre est exempte de pollution dès lors que ses caractéristiques sont cohérentes avec **le fond géochimique naturel local**[1]. Un sol peut être considéré sans danger pour les populations lorsqu'il est conforme à son état naturel initial, et lorsqu'il est conforme à l'état d'un sol dont il est admis qu'il ne pose pas de problème particulier pour l'usage envisagé (2).

Au cours d'un diagnostic de sol ou en cas d'excavation de terres, les résultats d'analyse obtenus sur le terrain sont confortés par comparaison aux référentiels disponibles sur la qualité de sols (3) :

- du BRGM disponible sur le site InfoTerre http://infoterre.brgm.fr :
  - IMN - Inventaire Minier National du BRGM,
- de l'INRA disponibles sur le site du GIS SOL www.gissol.fr :
  - BD ETM - Base de Données des Éléments Traces Métalliques,
  - RMQS - Réseau de Mesure de la Qualité des Sols,
  - ASPITET - Apport d'une Stratification Pédologique pour l'Interprétation des Teneurs en Éléments Traces.

Mais les échantillons destinés à ces bases de données, ne couvrent pas tout le territoire, et ont été prélevés et analysés selon des protocoles différents de ceux employés dans le domaine des sites et sols (potentiellement) pollués. Par exemple, la minéralisation[2] des échantillons de sol a le plus souvent été réalisée au moyen d'acide fluorhydrique pour tendre vers une dissolution complète et ainsi atteindre les concentrations dites « totales ». Cependant, l'analyse des échantillons de sol dans le domaine des sites et sols

[1] Pour les textes officiels, il s'agit de distinguer d'une part les éventuelles pollutions attribuables au site auquel on s'intéresse, d'autre part, les anomalies naturelles et les contributions anthropiques n'impliquant pas le site (1). La notion de « fond géochimique naturel », que les textes associent à un état initial de l'environnement exempt de toute pollution anthropique, semble cohérente avec le « fond pédogéochimique naturel » défini par D. Baize (9) : Concentration naturelle d'un élément majeur ou trace dans un horizon de sol, résultant uniquement de l'évolution géologique et pédologique, à l'exclusion de tout apport d'origine anthropique. Mais il est aujourd'hui illusoire de rechercher le « fond pédogéochimique naturel » pour bon nombre de substances. Il convient donc de tenir compte de la superposition des contributions diffuses dues aux activités anthropiques (en dehors de celles du site considéré) au fond pédogéochimique naturel : le Fond PédoGéochimique Anthropisé (FPGA).

[2] La « minéralisation », dans ce cadre, est une mise en solution par attaque acide des éléments contenus dans un échantillon de sol, en vue de son analyse.

(potentiellement) pollués se contente d'une attaque dite « pseudo-totale » à l'eau régale (mélange d'acides chlorhydrique et nitrique)[3].

On retiendra surtout que les échantillons des bases de données usuelles sont prélevés en milieu rural. Or, dans les agglomérations urbaines des contributions anthropiques se superposent au fond pédogéochimique naturel local car les sols y sont le réceptacle des retombées atmosphériques locales dues à l'artisanat, à l'industrie (y compris minière), aux chauffages urbain et individuel, au trafic routier, etc…

De plus, ces « sols » sont souvent constitués de remblais d'origine naturelle (ex. : sables) ou anthropique (ex. : gravats) qui peuvent contenir des quantités importantes de substances indésirables (ex. : bitumes, scories, mâchefers). Ces dépôts et ces remblais peuvent modifier les concentrations de certaines substances par rapport aux valeurs rurales[4].

Dans ces conditions, l'usage d'un référentiel rural, pourrait biaiser les études sur la qualité des sols urbains et il convient de déterminer un **fond pédogéochimique anthropisé urbain**.

1.2. LE PROJET FOND GÉOCHIMIQUE URBAIN - FGU

1.2.1. Objectifs

Le projet intitulé « Établissement d'un fond géochimique urbain et industriel » est réalisé dans le cadre de la convention de financement FGU n° 1372C0016 signée entre l'ADEME et le BRGM qui se déroule entre le 12 septembre 2014 et le 12 septembre 2017. Cette convention fait suite à une première convention FGU n°1072C0046 (2010-2014) (4). Ces deux conventions ont pour objectif principal l'établissement de fonds pédogéochimiques anthropisés dans les principales agglomérations françaises. La constitution de ces fonds pédogéochimiques anthropisés urbains s'appuie, dans un premier temps, sur le recueil des analyses de sols réalisé dans 411 villes françaises par le BRGM pour le compte du ministère de l'écologie dans le cadre du projet « Diagnostic des sols dans les établissements accueillant des enfants et des adolescents ». Au cours de ce projet lancé depuis 2008, aussi appelé « Établissements sensibles » ou ETS, plus de 2 400 établissements devraient faire l'objet en France, à terme, de visites, de prélèvements et d'analyses pour évaluer la qualité des milieux de vie des populations dites « sensibles ». Les diagnostics ETS font appel à plusieurs prélèvements dits « témoins » réalisés sur des sites voisins, pour comparer les résultats des analyses de sols obtenues au droit des établissements.

Initialement on espérait ainsi obtenir environ autant d'analyses d'échantillons « témoins » que d'établissements retenus par la démarche ETS, soit 2 400 à la fin de l'opération[5]. Dans ces conditions, seulement cinq agglomérations de France métropolitaine présenteront, à terme, une population de plus de 30 échantillons. Par conséquent le projet ETS ne pourra pas, à lui seul, fournir le volume de données statistiquement nécessaire pour déterminer le ou les référentiels recherchés à l'échelle de chaque agglomération française.

[3] De plus, l'usage de l'acide fluorhydrique est de plus en plus contraint pour des raisons de sécurité.

[4] Dans certains lieux résultant d'importants remaniements le fond pédogéochimique naturel local n'a plus d'influence.

[5] Le projet ETS est actuellement engagé sur 2 premières phases qui ne couvrent que 1 400 établissements. Une partie de ces établissements ne fera pas l'objet d'un diagnostic (établissements finalement hors démarche avant le début du diagnostic, refus du diagnostic par le maître d'ouvrage). Enfin une partie des prélèvements réalisés en dehors des consignes prescrites s'avèrent inexploitables. Certains de ces prélèvements sont découverts avant analyse (phase 1 du diagnostic), d'autres après analyse lors des phases 2 des diagnostics. Dans ce dernier cas, les analyses sont bancarisées dans la base de données mais les échantillons SLU sont rebaptisés SLE ou refusés. Au final on estime à moins de 1 000 le nombre d'échantillons « témoins » dont les analyses seront bancarisées à l'issue des 2 conventions FGU entre l'ADEME et le BRGM (4).

Les informations recueillies doivent donc être complétées et l'ensemble bancarisé de façon appropriée. Les objectifs de la convention en cours sont donc de :

- poursuivre la bancarisation d'analyses de sols urbains « exempts » de pollution directe et répartis sur l'ensemble du territoire national, obtenus dans le cadre du projet « Diagnostic des sols dans les établissements accueillant des enfants et des adolescents » ;
- refondre le système de collecte et la base pour permettre la bancarisation de données obtenues dans le cadre de projets divers et selon des protocoles de prélèvement et d'analyse différents. La base contiendra donc des analyses représentatives :
  - d'une population de points « témoins » comparables à ceux initialement recherchés par la première convention du projet FGU,
  - de populations de points renseignant sur la qualité de l'ensemble des sols urbains ;
- déterminer les fonds pédogéochimiques pour l'ensemble des paramètres analysés dans les principales agglomérations françaises.

La première convention ayant mis l'accent sur plusieurs questions méthodologiques, il s'agit aussi au cours de cette deuxième étape de :

- compléter l'étude bibliographique, notamment pour le traitement statistique des données ;
- participer au Groupe de Travail « Bruit de Fond » mis en place par l'ADEME pour la rédaction d'un « Guide de bonnes pratiques pour la détermination de fonds pédogéochimiques anthropisés pour la gestion d'un site pollué en milieu urbain, rural ou industriel » ;
- et enfin, de participer à la révision de la norme NF EN ISO 19258 « Qualité du sol - Guide pour la détermination des valeurs de bruit de fond ».

1.2.2. Méthode d'obtention et de bancarisation des analyses ETS

Les échantillons « témoins » de l'opération ETS sont codés SLU (Sols Urbains). Les espaces verts, et préférentiellement les jardins publics, ont été retenus pour la réalisation de ces prélèvements car ce sont les plus accessibles pour les équipes de préleveurs. En outre, ils sont jugés *a priori,* exempts d'impact polluant ponctuel, mais cependant représentatifs du cumul des dépôts atmosphériques diffus urbains.

En phase 1 des diagnostics ETS, des échantillons SLU sont systématiquement prélevés. Conformément à l'organigramme de la figure 1, ils ne sont analysés qu'en cas d'absence de prélèvement et d'analyse de sols au cours de la phase 2. Ces échantillons SLU sont prélevés entre 0 et 5 cm de profondeur et sont représentatifs des sols de surface accessibles aux populations sensibles par un porté main-bouche.

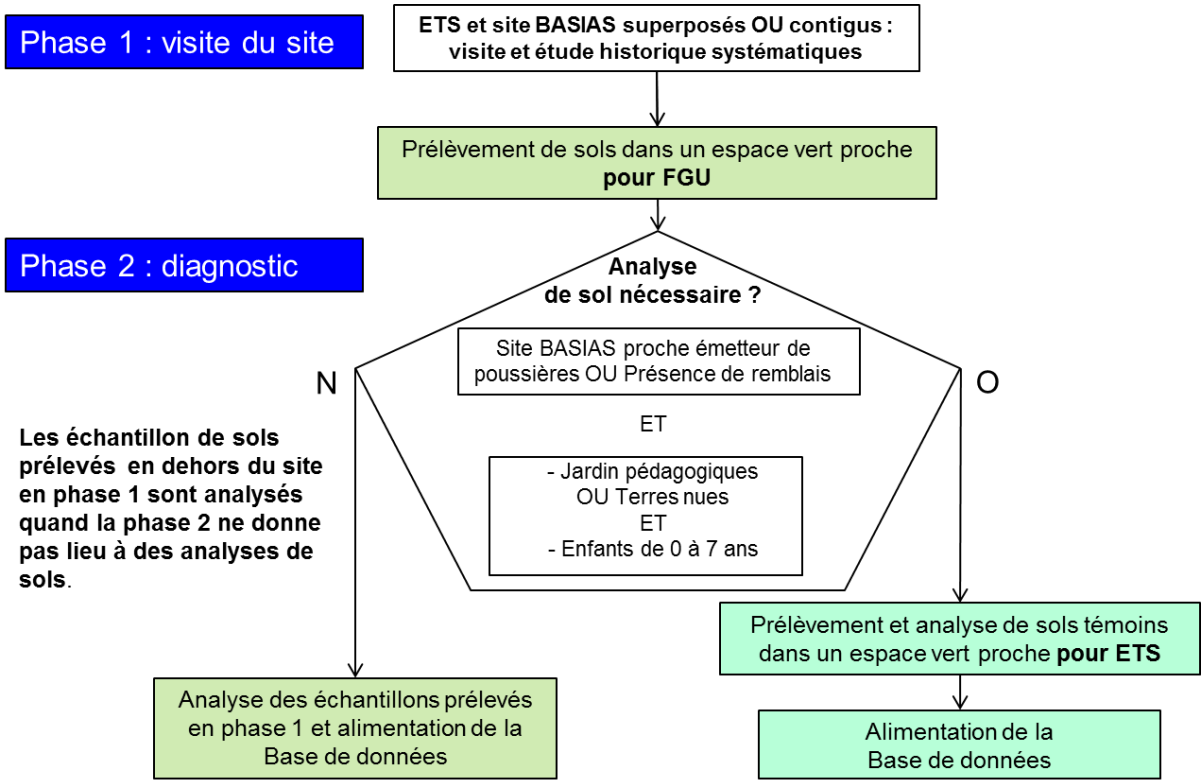


Figure 1 : Organigramme de sélection des échantillons SLU dans le cadre des diagnostics ETS.

En phase 2, les diagnostics ETS conduisent dans certains cas à des prélèvements d'échantillons « témoins » SLU et à des prélèvements dans les établissements. Ces derniers sont codés SLE (Sol des Etablissements). Pour compléter la connaissance de la qualité des sols en milieu urbain, les analyses de ces échantillons sont également bancarisées dans la base de données. Les échantillons de phase 2 sont prélevés entre 0 et 5 cm ou entre 0 et 30 cm de profondeur pour prendre en compte l'ingestion de légumes racines/tubercules en contact avec le sol quand un jardin potager pédagogique est présent dans l'établissement.

L'ensemble des échantillons SLU retenus dans la base de données provient de villes de plus de 5 000 habitants. Les effets dits « pépites » sont minimisés par des échantillonnages composites réalisés par 5 prélèvements aux coins et au centre de carrés de trois mètres de côté. Les éléments grossiers et les éventuels systèmes racinaires sont éliminés.

En raison de la configuration de certains établissements, les échantillons SLE peuvent provenir de villes de moins de 5 000 habitants et être obtenus dans des conditions différentes : prélèvements ponctuels, à l'emplanture des arbres, dans des bacs de fleurs accessibles aux populations sensibles, …

Les échantillons sont tamisés à 2 mm et la phase inférieure broyée à 80 µm. Les analyses des substances minérales sont réalisées après solubilisation des échantillons dans l'eau régale. Les analyses visent les principaux éléments traces métalliques (cuivre, chrome, plomb, zinc, nickel, cadmium, mercure), un métalloïde (arsenic) et des substances persistantes organiques (cyanures totaux, hydrocarbures aromatiques polycycliques (HAP), polychlorobiphényles (PCB), dioxines (PCDD), furanes (PCDF).

2. Bilan de la collecte de données

Au 13 octobre 2016, la base de données FGU compte 1 554 échantillons de sols (SLU, SLE et « refusés »). Ces échantillons ont été prélevés à proximité ou au droit de 839 établissements implantés dans 293 villes métropolitaines réparties dans les 20 régions concernées par la première et la seconde tranche de l'opération ETS. Le nombre de résultats d'analyse bancarisés s'élève à 73 078. Les 1 554 échantillons se répartissent comme suit :

- 632 échantillons prélevés dans des espaces verts, dits SLU (sols urbains) et représentatifs du fond pédogéochimique anthropisé urbain au sens de la convention FGU ADEME-BRGM. Ils correspondent à 30 635 résultats d'analyse ;
- 877 échantillons prélevés au droit des établissements scolaires, dits SLE (sols des établissements) et donc potentiellement contaminés par les activités liées à la présence d'un ancien site inventorié dans BASIAS ;
- 45 échantillons dits « refusés », impropres à une valorisation pour absence de respect des consignes de prélèvements (mais dont les analyses sont malgré tout bancarisées).

Le graphique de la figure 2 décrit l'évolution de la bancarisation des données dans la base de données créée lors de la première et de la deuxième convention.

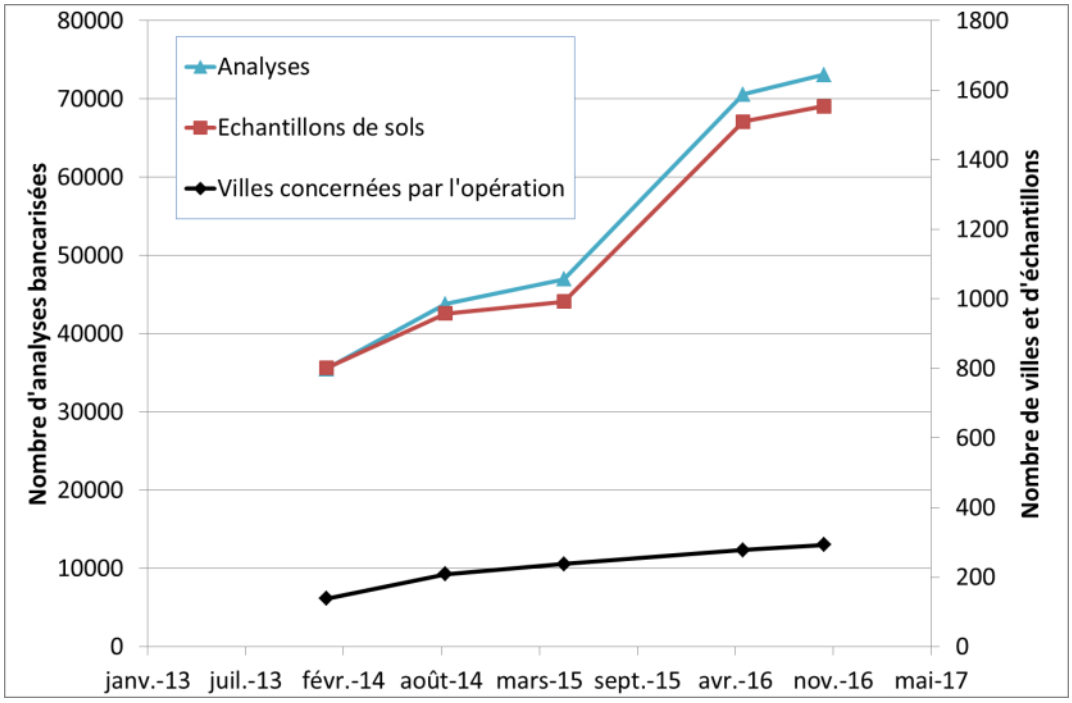


Figure 2 : Évolution du nombre d'analyses bancarisées, du nombre d'échantillons et de villes associés.

3. Traitement des données

3.1. ESSAIS DE TRAITEMENT GÉOSTATISTIQUE

Une partie des données bancarisées jusqu'en 2015 (4) a été sélectionnée et soumise à des essais de traitement géostatistique. Les critères choisis pour la sélection des données sont :

- le nombre d'analyses disponibles : en 2015 seules 3 agglomérations, notées A, B et C, présentaient un nombre de points de prélèvements de l'ordre de 30 ou plus ;
- la présence d'une substance minérale et d'une substance organique : pour ces essais l'arsenic et la dioxine OCDD ont été retenues ;
- un faible nombre de valeurs inférieures à la limite de quantification analytique pour les populations des substances choisies.

Le tableau de la figure 3 récapitule les paramètres de cette sélection.

	Agglomération B		**Agglomération C**	
Substance	Arsenic	OCDD	Arsenic	OCDD
Analyses disponibles	47	20	26	14
Nombre de valeurs < LQ	3 (6 %)	0 (0 %)	0 (0 %)	1 (7 %)

Figure 3 : Tableau récapitulatif des critères retenus pour la sélection des données traitées en géostatistique.

Les essais sont réalisés sur l'ensemble des jeux de données disponibles. Les résultats obtenus suite au traitement géostatistique de ces données sont récapitulés en annexe 1. Malgré la sélection des analyses les plus appropriées parmi les données disponibles, ils font globalement apparaître :

- des paramètres dont il est difficile de garantir la fiabilité dans le cas de trop faibles effectifs des populations ;
- des difficultés de traitement en cas de répartition hétérogène des points de données (points alignés ou grandes zones sans données) ;
- des corrélations spatiales parfois inexistantes, faibles, ou difficiles à caractériser de façon suffisamment claire.

Ces éléments engendrent des difficultés pour obtenir une interpolation cartographique fiable. Dans les conditions les plus contraignantes (trop peu de données, pas de corrélation spatiale), une interpolation n'aurait même aucun sens et est tout simplement à éviter.

3.2. TRAITEMENT STATISTIQUE

Au cours de la première convention ADEME-BRGM (2010-2014), plusieurs questions méthodologiques relatives au traitement statistiques des données recueillies par le projet FGU ont été identifiées :

- comment gérer les valeurs extrêmes (outliers) ?
- comment déterminer un seuil au-delà duquel les valeurs diffèrent significativement du fond pédogéochimique anthropisé ?

Les données FGU présentent, d'une part des effectifs faibles par agglomération, et d'autre part des taux élevés de valeurs inférieures aux limites de quantification. Les réponses à ces questions doivent donc tenir compte de ces spécificités à chaque étape du traitement :

- préparation des données au moyen de tests et de transformations adaptés ;
- détermination de statistiques descriptives ;
- interprétation des données ;
- analyse des résultats par croisement des données (par exemple, confrontation aux données existantes de l'INRA dans le milieu rural avoisinant les agglomérations étudiées).

Dans cet objectif, et suite aux résultats des essais géostatistiques, une étude approfondie a été conduite au cours d'un stage de 5^{e} année en école d'ingénieur entre avril et septembre 2016. L'objectif du stage était de constituer, sur la base d'une étude bibliographique, un protocole de traitement adapté aux données FGU à chacune de ces étapes.

Les principaux résultats de cette étude se trouvent en annexe 2, dans le rapport de stage de L. Sancho (Ingénieur Université Pierre et Marie Curie, Polytech'Paris, spécialité Sciences de la Terre). L'étude bibliographique repose essentiellement sur les travaux de Reimann (5) et Helsel (6) et conduit à l'élaboration d'un arbre de décision pour un traitement des données ajusté aux situations rencontrées dans le cadre du projet FGU.

En outre, cette étude confirme la nécessité de faire appel à des méthodes adaptées aux spécificités des données de départ, au contexte de l'étude et aux besoins exprimés. Certaines méthodes, appliquées avec raison dans des conditions habituelles, s'avèrent parfois inadaptées aux données du projet FGU, notamment pour :

- la détermination des quantiles ;
- la transformation logarithmique ;
- la substitution des données inférieures à la limite de quantification par 50 % de cette limite.

Cette dernière méthode, par exemple, est très souvent utilisée pour le traitement des analyses d'explorations minières. En effet, si ces analyses comportent des résultats inférieurs aux limites de quantification analytiques (LQ), les ignorer, les remplacer par 0 ou par 100 % de la LQ peut entrainer un biais préjudiciable à l'interprétation des résultats. La substitution par 50 % de la LQ est un compromis rapide et facile qui permet de tenir compte de ces valeurs. Elle est bien adaptée à l'exploration minière qui cherche à mettre en évidence des teneurs élevées de matériaux précieux. Elle semble aussi la solution la plus fréquemment employée pour gérer les valeurs inférieures aux limites de quantification analytiques dans le cadre de recherche de valeurs seuil ou de fonds géochimiques (7). Pourtant, dans ces contextes cette méthode doit être affinée, notamment dans le cas de substances peu présentes dans l'environnement, ou de LQ particulièrement élevées.

4. Refonte de la base de données

Une base de données a été créée pour les besoins de la première convention ADEME-BRGM (2010-2014) afin de bancariser les données obtenues au cours du projet ETS. Conçue sous Access® / Oracle®, elle est alimentée via un fichier constitué de trois tableaux réalisés sous le tableur Excel®. Le premier tableau permet la saisie des informations concernant les échantillons, c'est-à-dire essentiellement : nom, localisation, profondeur. Le deuxième tableau permet d'associer l'échantillon à l'établissement diagnostiqué par le projet ETS. Ces 2 tableaux constituent les métadonnées des données analytiques contenues dans le troisième tableau. Pour éviter la saisie des noms des substances et des unités sous des formes multiples (ex : mg/kg et milligrammes par kilogrammes), des lexiques ont été créés :

- pour les substances (ex : COMP001 correspond à l'arsenic) ;
- pour les unités (ex : UNI001 correspond à mg/kg de matière sèche).

Cette base de données répond aux besoins simples de la bancarisation des données du projet ETS toujours acquises selon les mêmes protocoles.

Il est toutefois apparu très tôt que :

- le nombre d'échantillons recueillis dans le cadre du projet ETS serait, par construction, insuffisant pour élaborer des fonds pédogéochimiques anthropisés dans l'ensemble des agglomérations françaises (4). À terme, seulement quatre à cinq agglomérations devraient disposer de plus de 30 échantillons dans la base. Cette valeur est considérée comme une limite inférieure pour *« dresser un histogramme représentatif ou pour calculer un percentile représentatif »* (8) ;
- les attentes des parties impliquées dans le domaine des sols urbains s'étendent au-delà de l'objectif de la première convention ADEME-BRGM et la détermination du fond pédogéochimique anthropisé. Elles comprennent la connaissance complète de la qualité pédogéochimique des sols de leur territoire en surface comme en profondeur (notamment pour la gestion des terres excavées).

Dans un premier temps, et toujours dans le cadre des diagnostics ETS, la base de données a pu bancariser, en plus des analyses d'échantillons témoins (SLU), les analyses des échantillons prélevés au droit des établissements scolaires (SLE). Dans certaines conditions[6], et en fonction du classement des établissements à l'issue des diagnostics ETS, ces échantillons pourraient répondre aux deux besoins identifiés ci-dessus[7].

Pour aller plus loin dans cette démarche, la seconde convention ADEME-BRGM (2014-2017) a prévu la refonte totale de la base de données. Il s'agit de bancariser les analyses provenant de projets divers obtenues selon des protocoles différents de celui du projet ETS. La nouvelle base contiendra donc un nombre plus élevé de données descriptives (méta données) de l'échantillon, du point de prélèvement, des analyses et des intervenants (préleveurs, laboratoires, etc.).

[6] Il s'agit d'un travail qui reste à conduire, comprenant vérification et tri, et qui devra tenir compte des normes du domaine et des recommandations du guide que fera paraître le Groupe de Travail « BdF » de l'ADEME.

[7] Les établissements scolaires ont déjà été pris en compte dans d'autres projets comme G-BASE et GSUE au Royaume-Uni (7).

Cette refonte de la base, désormais appelée BDSolU (Base de données des analyses de Sols Urbains) est aussi une opportunité pour :

- développer la base sous le langage PostGreSQL et mettre en place son alimentation via un site internet dédié qui permettra une première vérification automatisée des fichiers postés par les fournisseurs de données (tâche assurée actuellement entièrement manuellement et très chronophage) ;
- adosser les informations bancarisées à des références reconnues (lexiques SANDRE[8], Corine Land Cover, BASIAS, INSEE,…) et rendre la base cohérente avec des outils existants (GDM®[9], BSS[10], BASIAS,…).

Le BRGM et l'ADEME voient dans BDSolU l'opportunité d'établir une large base de connaissance sur la qualité pédogéochimique des sols urbains sur l'ensemble du territoire national. La base de données sera à terme publique, et les différents producteurs de données pourront y déposer leurs analyses grâce à un outil en ligne. Dans un premier temps BDSolU sera alimentée par les données du projet ETS et celles de plusieurs projets achevés ou en cours, auxquels le BRGM participe dans différentes agglomérations en France.

De plus, le BRGM propose aux collectivités concernées par cette thématique un accompagnement basé sur l'échange de données collectées et de services. En fonction de modalités à convenir entre les parties (contrat ou convention), les fournisseurs de données pourront bénéficier :

- du modèle de gestion des données développé pour BDSolU par le BRGM ;
- de l'expérience du BRGM et de protocoles mis au point et validés au niveau national par les membres du Groupe de Travail « BdF » de l'ADEME ;
- d'un appui du BRGM pour la collecte, l'homogénéisation (dans les limites imposées par les différents protocoles de prélèvement, d'échantillonnage et d'analyse mis en œuvre), l'analyse critique et le traitement de leurs données.

[8] SANDRE : Service d'Administration Nationale des Données et Référentiels sur l'Eau.

[9] GDM est une suite logicielle développée par le BRGM pour le traitement géostatistique et la visualisation des informations sur le sous-sol en 3D.

[10] BSS : Banque de données du Sous-Sol gérée par le BRGM.

5. Norme NF EN ISO 19258

La norme NF EN ISO 19258 de septembre 2011 « Guides pour la détermination des valeurs de bruit de fond » (indice de classement X 31-606) reproduit intégralement la norme internationale ISO 19258:2005 « Guidance on the determination of background values ». Elle fait partie des normes internationales qui font l'objet d'un examen systématique tous les trois ans. Suite à une enquête publique conduite en 2009, un processus de révision a été entamé en 2014.

Dans notre pays, cette révision a été conduite par le BRGM (nommé chef de projet pour la France) avec l'IRD et l'INRA. En 2015 le Committee Draft a émis un vote positif pour engager les révisions en tenant compte des remarques des différents pays. Une enquête publique a été conduite de mai à juin 2016. Le dépouillement final globalisé pour l'ensemble des pays aura lieu fin octobre 2016 et la nouvelle version de la norme devrait paraître en 2017. La révision de cette norme était une opportunité pour :

- tenir compte des contraintes liées au contexte urbain ;
- clarifier certaines définitions spécifiques à la norme NF EN ISO 19258 (cependant la plupart des définitions se trouvent dans la norme NF EN ISO 11074) ;
- prendre en compte les contraintes liées à la détermination d'un fond pédogéochimique anthropisé pour les molécules organiques, dont certaines ne sont pas présentes naturellement dans l'environnement ;
- mettre à jour les références bibliographiques de la norme.

Les lignes directrices retenues dans la révision de la norme sont :

- la détermination de valeurs de « bruit de fond » selon les principales méthodes relatives aux substances minérales et organiques présentes dans les sols, aux échelles locale et régionale ;
- l'identification des méthodes d'échantillonnage et des stratégies d'échantillonnage ;
- la prise en compte des méthodes d'échantillonnage ;
- la prise en compte des méthodes de traitement des données.

En revanche, la norme ne concerne par la détermination de valeurs de bruit de fond :

- à l'échelle du site ;
- des eaux souterraines et des sédiments.

En France l'enquête publique a recueilli trois réponses en provenance de l'UPDS, du BRGM et d'un anonyme. Les principaux commentaires concernent :

- la traduction de la version anglaise en français. Il s'agit essentiellement de points relatifs à la norme NF EN ISO 11074 dite « Vocabulaire » dont la révision est prévue annuellement mais qui doit faire l'objet d'un consensus entre 10 pays ;
- des demandes pour détailler les méthodes d'exploitation des données. Une norme « ISO/DIS 18400-104 Soil quality — Sampling — Part 104: Strategies », est en cours de rédaction sur la stratégie et les statistiques mais il n'existe pas actuellement de version stabilisée ;
- des besoins de clarifications divers.

La réponse BRGM correspond à la relecture critique de la proposition de norme révisée dans le cadre du projet FGU.

6. Participation au Groupe de Travail « Bruit de Fond »

Suite aux différentes questions méthodologiques mises en évidence au cours de la première convention ADEME-BRGM (2010-2014) et mentionnées dans le rapport de 2015 (4), l'ADEME a mis en place un Groupe de travail « Bruit de fond » animé par l'YNCREA (ex ISA de Lille)[11].

Dans le cadre de la deuxième convention FGU (2014-2017), le BRGM participe activement au Groupe de travail et alimente les discussions en concertation avec l'ADEME et YNCREA. Le tableau de la Figure 4 récapitule les exposés présentés au cours des rencontres du Groupe de travail (GT) en session restreinte au groupe d'experts scientifiques ou élargie à l'ensemble des acteurs du domaine.

Date et type des rencontres	**Présentations du BRGM dans le cadre du projet FGU**
27 mai 2015 – GT restreint	• Présentation du projet FGU - Évolution et résultats de la première convention (2010-2014)
22 septembre 2015 – GT élargi	• Présentation du projet FGU - Évolution et résultats de la première convention (2010-2014)
29 juin 2016 – GT restreint	• Convention (2014-2017) - Bancarisation des données ETS - Construction de la base de données BDSolU • Révision de la norme « Guidance on the determination of background values ISO 19 258 » • Établissement d'un Protocole d'Analyse Statistique pour la Construction d'un Fond Géochimique des Sols Urbains
06 octobre 2016 – GT restreint	• Représentation schématique des contaminations diffuses sur une portion d'agglomération

Figure 4 : Tableau récapitulatif des rencontres du GT « BdF » et des contributions du projet FGU.

[11] Les trois associations « Groupe HEI ISA ISEN Lille », « ISEN Brest » et « ISEN Toulon » se sont fédérées le 20 juin 2016 au sein d'Yncréa (www.yncrea.fr).

7. Conclusions

La seconde convention FGU ADEME-BRGM (2014-2017) poursuit les tâches entamées lors de la convention précédente (2010-2014) :

- étude bibliographique ;
- bancarisation des données produites par le projet ETS (632 échantillons témoins pour plus de 73 000 analyses bancarisées en octobre 2016) ;
- contribution à la révision de la norme NF EN ISO 19258 ;
- recherche de réponses aux questions méthodologiques dans le cadre d'une participation aux travaux du Groupe de Travail « Bruit de fond » et d'un stage relatif au traitement statistique des données.

Une nouvelle tâche est aussi engagée pour refondre totalement la base de données et son mode de fonctionnement imaginés au cours de la première convention. La nouvelle base de données des analyses de sols urbains, BDSolU, permettra la bancarisation des analyses du projet « Établissements Sensibles » et celles d'autres projets. Certains d'entre eux conduits par le BRGM sont déjà identifiés et des partenariats avec des collectivités urbaines détentrices de données sont également à l'étude.

Ces travaux doivent permettre la détermination de fonds pédogéochimiques anthropisés dans plusieurs agglomérations françaises. Ils contribueront également à l'amélioration de la connaissance générale de la qualité géochimique des sols en milieu urbain.

8. Bibliographie

1. MEDDE. Note du 8 février 2007. *Site internet du MEDDE.* [En ligne] 2007. www.developpement-durable.gouv.fr/Note-du-8-fevrier-2007-Sites-et.html.

2. BLANC, Céline. Guide de réutilistion hors site des terres excavées en technique routière et dans des projets d'aménagement. *Site du Ministère du Développement durable.* [En ligne] 2012. www.developpement-durable.gouv.fr/spip.php?page=doc&id_article=27486.

3. MEDDE. Bases de données relatives à la qualité des sols : contenu et utilisation dans le cadre de la gestion des sols pollués. [En ligne] Avril 2008. [Citation : 23 Juin 2015.] www.developpement-durable.gouv.fr/spip.php?page=doc&id_article=19946.

4. BRUNET J.-F. avec la collaboration de GUIET F., BLANC C., LAPERCHE V., BALON P., AUBERT N. *Etablissement de fonds pédo-géochimiques urbains et industriels en parallèle à l'Opération ETS du Ministère du Développement durable.* BRGM. 2015. Rapport final. BRGM/RP-64845-FR.

5. REIMANN C., FILZMOSER P., DUTTER R. *Statistical Data Analysis Explained : Applied Environmental Statistics with R.* s.l. : John Wiley & Sons, Ltd, 2008.

6. HELSEL, D.R. *Statistics for censored environmental data using Minitab and R.* Denver, Colorado : John Wiley & Sons, Inc., 2012.

7. JARZABEK, Maxime. *Rapport bibliographique : retour d'expérience sur les fonds pédogéochimiques de sols urbains – Pratiques à l'étranger.* s.l. : BRGM , 2014.

8. Guides pour la déterminatin des valeurs de bruit de fond. *Norme Européenne - Norme Française - Qualité du sol.* 2011. NF EN ISO 19258. X 31-606.

9. BAIZE, Denis. *Petit Lexique de Pédologie.* s.l. : INRA, Paris, 2004. ISBN : 2-7380-1114-4 - ISSN : 1159-5663.

Annexe 1

Essai de traitement géostatistique des données

B. Bourgine (BRGM DGR/GSO) - Février 2016

1. Données de l'agglomération B

1.1. Traitement des données arsenic

Pour la variable As, 47 points de mesure sont disponibles. La concentration en As varie de 0,5 à 18,4 mg/kg, avec une moyenne de 11 et un écart-type de 4,45.

Noter que 3 points ont une concentration indiquée <0,5. En première approche ces points ont été affectés de la valeur 0,5.

1.1.1. Histogramme

L'histogramme (Fig. 1.) ne montre pas de dissymétrie très marquée dans la distribution des concentrations. La carte de répartition des valeurs ne fait pas apparaître de répartition privilégiée dans le domaine d'étude (Fig. 2). Les valeurs faibles et fortes peuvent s'observer en tout point.

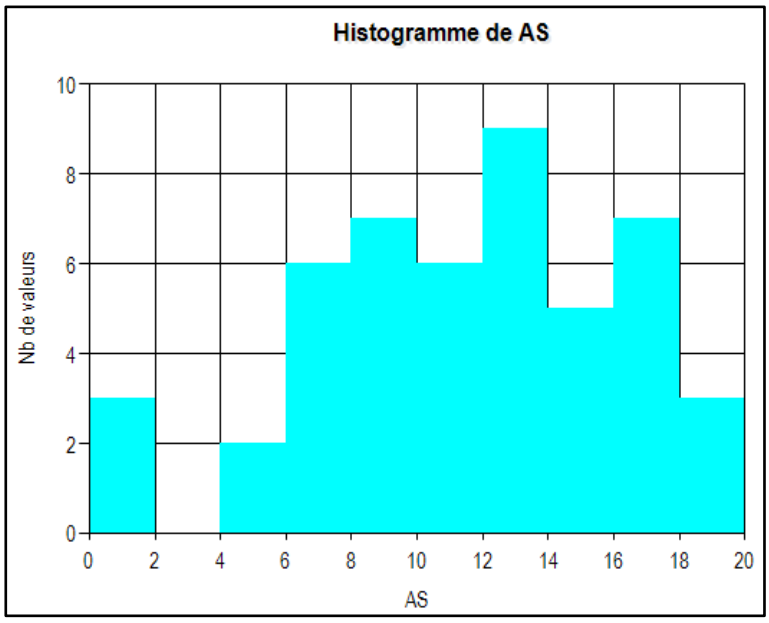


Figure 1 : Agglomération B - Histogramme de As (mg/kg).

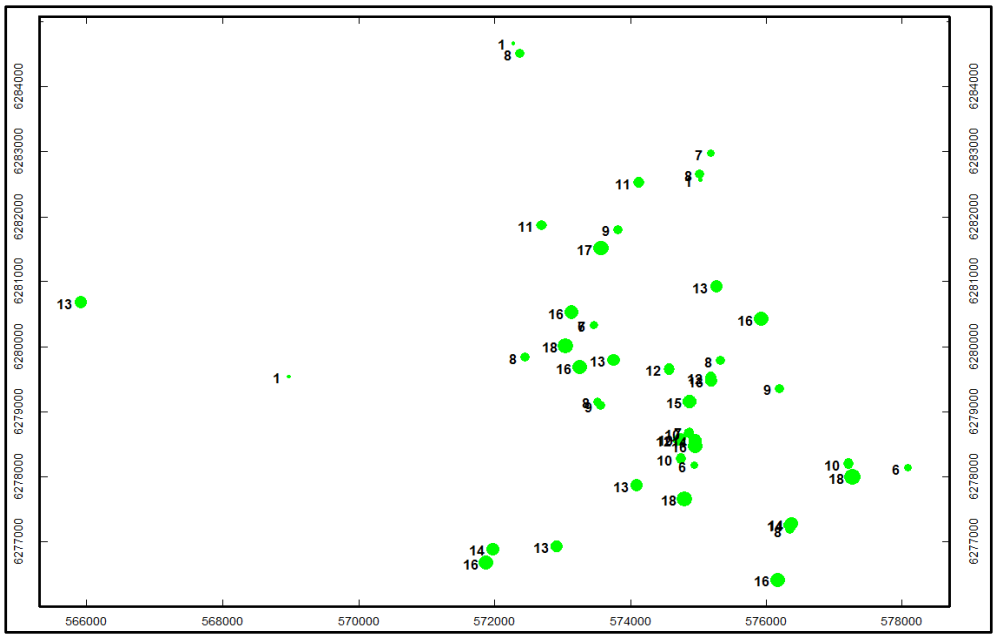


Figure 2 : Agglomération B - Carte des données de As (mg/kg). Coordonnées en Lambert 93. Taille du cercle proportionnel à As. NB : les valeurs affichées « 1 » sont en fait les concentrations à 0,5 mg/kg.

1.1.2. Variogramme

Un variogramme est calculé au pas de 250 m, jusqu'à 20 pas (5 km), selon plusieurs directions (Fig. 3).

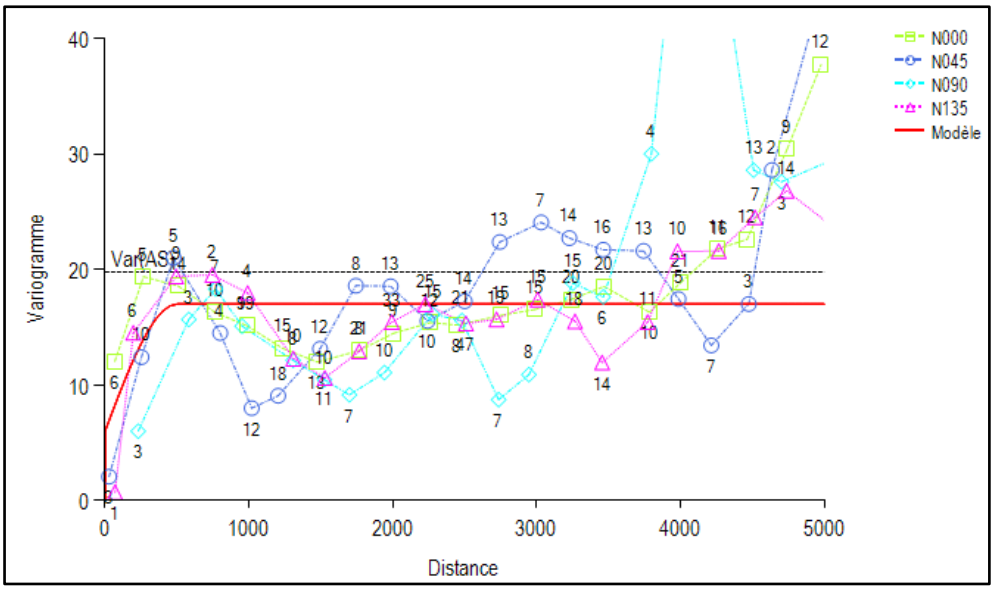


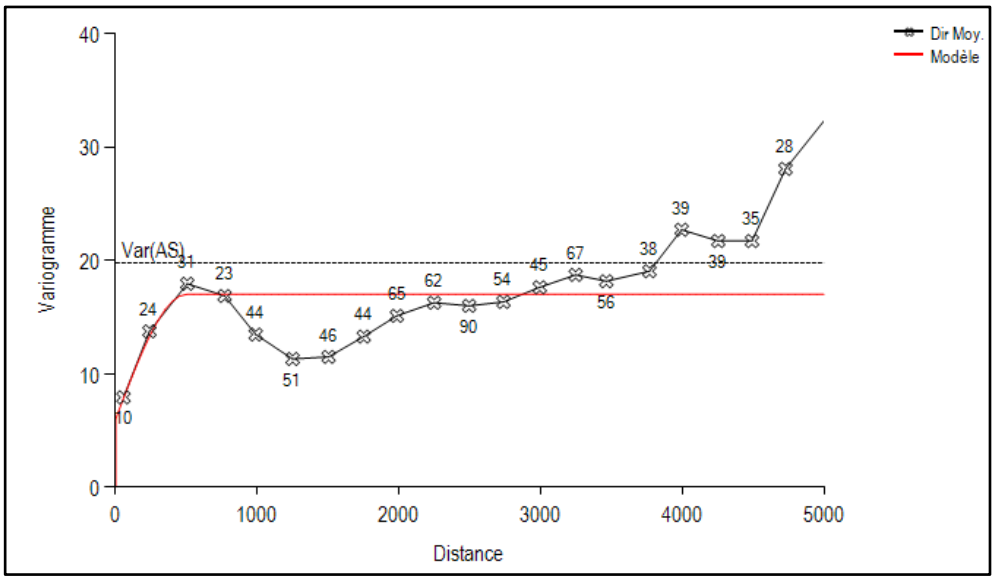
Figure 3 : Agglomération B - Variogramme directionnel de As.

Ce variogramme peut être considéré comme isotrope. Un variogramme « toutes directions confondues » est donc calculé (Fig. 4).

Ce variogramme montre une structure de courte portée (500 m) ainsi qu'un effet de pépite, voir un effet de trou suivi d'une légère dérive (Fig. 5). Toutefois le faible nombre de données rend le variogramme instable. En particulier la croissance du variogramme aux « grandes »

distances (au-delà de 2 à 3 km) est essentiellement liée aux 3 valeurs mesurées à 0,5 mg/kg et notamment à celle qui est isolée.

Figure 4 : Agglomération B - Variogramme des concentrations en As toutes directions confondues.



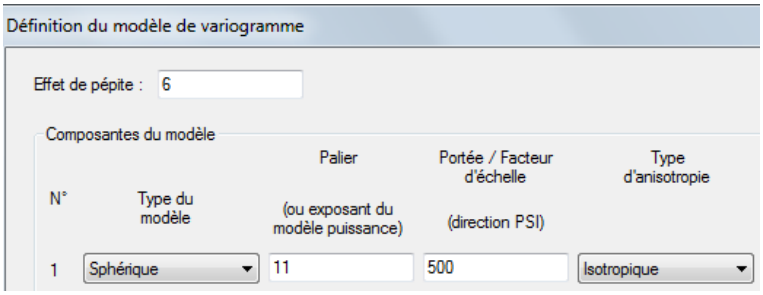


Figure 5 : Agglomération B - Paramètres du modèle de variogramme de As.

1.1.3. Validation croisée

Le modèle de variogramme ci-dessus est vérifié par validation croisée en choisissant un voisinage d'interpolation assez large (20 km).

Le tableau statistique des résultats de la validation croisée (Figure 6) montre que la moyenne des erreurs est nulle (non biais de l'estimation) et que l'écart-type des « erreurs réduites » assez proche de 1, ce qui signifie que le modèle est bien calé en terme de prévision des incertitudes.

On note toutefois 3 points « non robustes ». Ces trois points sont ceux dont la concentration en As est égale à 0,5 mg/kg. Deux de ces points sont situés tout près de mesures à 8 mg/kg et l'autre est un point isolé.

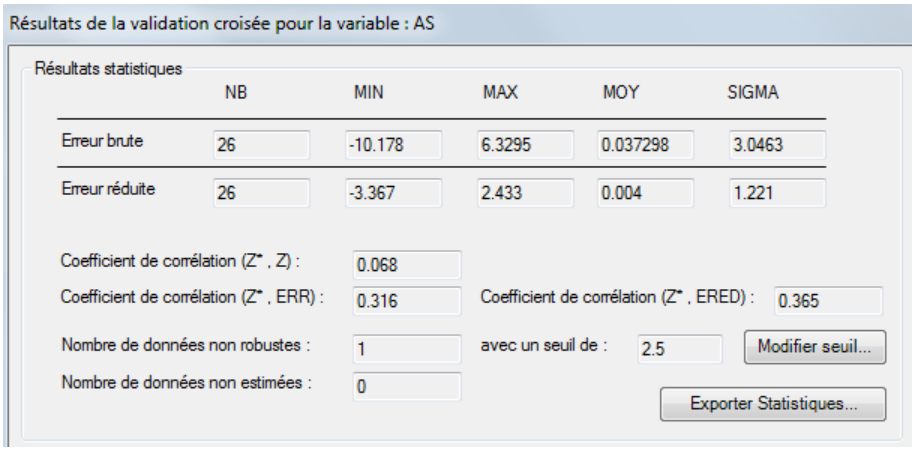


Figure 6 : Agglomération B - Validation croisée de As.

1.1.4. Carte des concentrations

À titre d'illustration, on a réalisé une interpolation des teneurs en As. L'interpolation est réalisée à l'intérieur d'un polygone arbitraire.

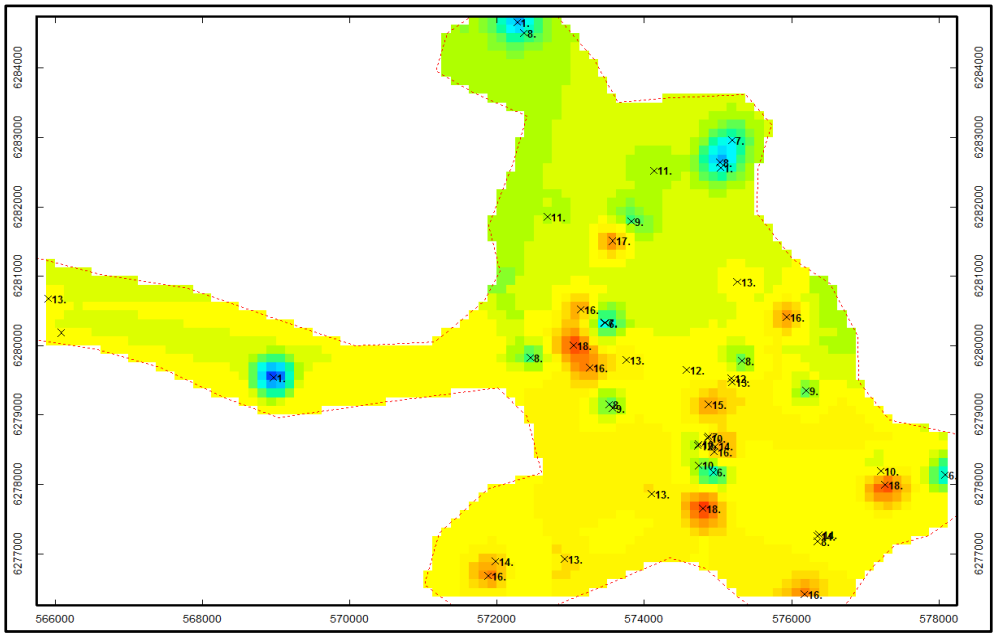


Figure 7 : Agglomération B - Interpolation de la concentration en As.

1.2. Traitement des données dioxine OCDD

1.2.1. Histogramme

Pour OCDD, on ne dispose que de 20 points de donnée (Fig. 8).

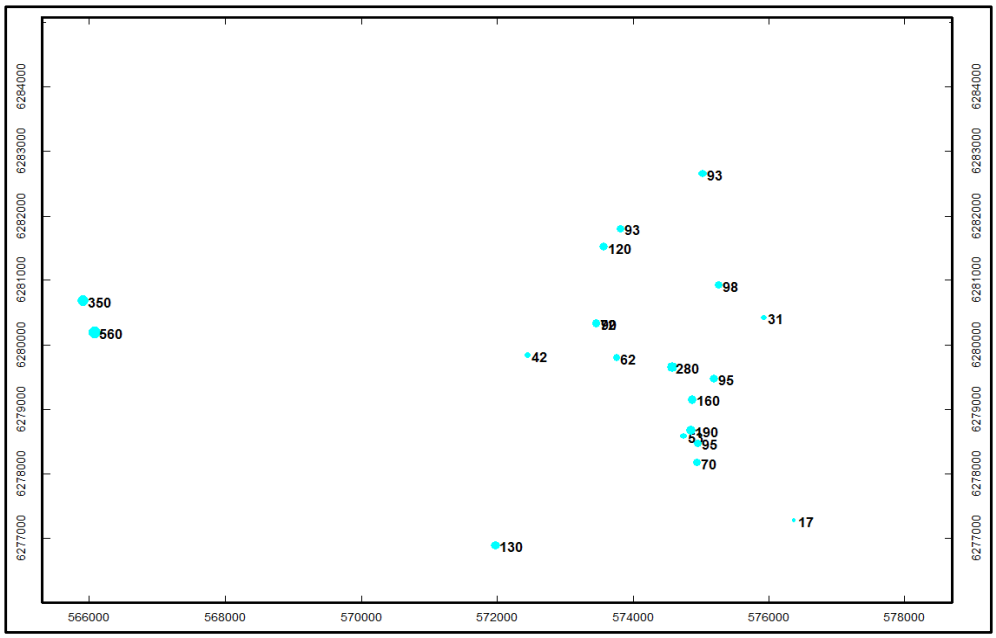


Figure 8 : Agglomération B - Concentrations en OCDD. Symbole proportionnel à Log(OCDD).

L'histogramme de ces données apparait comme assez fortement dissymétrique, avec quelques valeurs « élevées ».

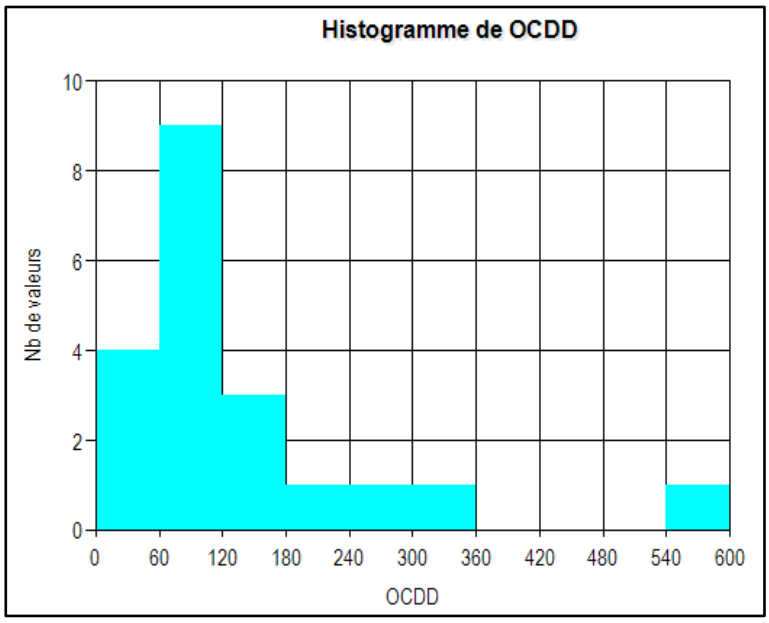


Figure 9 : Agglomération B - Histogramme de OCDD (valeurs en ng/kg).

En passant au logarithme on observe un histogramme beaucoup plus symétrique se rapprochant de celui d'une courbe gaussienne, mais avec beaucoup plus de valeurs autour de la médiane :

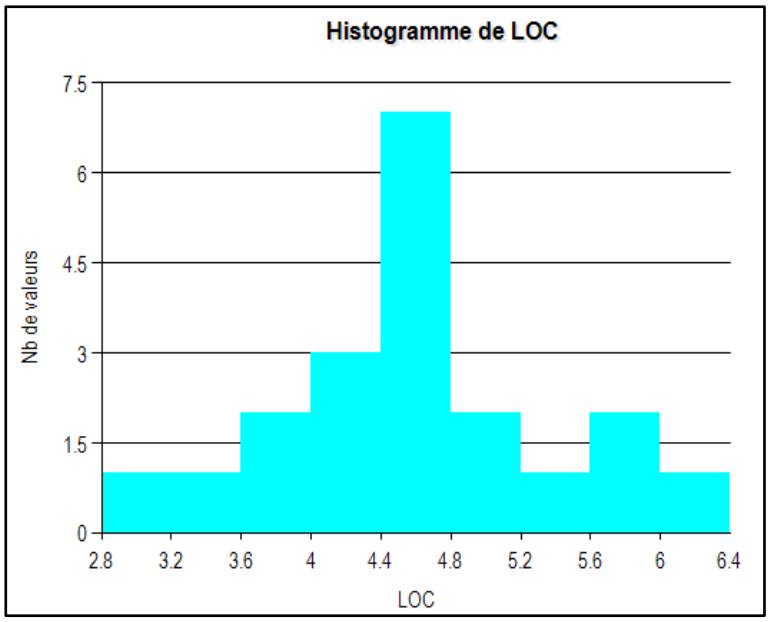


Figure 10 : Agglomération B - Histogramme de Ln(OCDD).

1.2.2. Variogrammes

Le variogramme de la concentration OCDD s'avère totalement inexploitable. Un passage au logarithme permet de limiter l'influence des valeurs extrêmes de la distribution et d'obtenir un variogramme exploitable (Fig. 11).

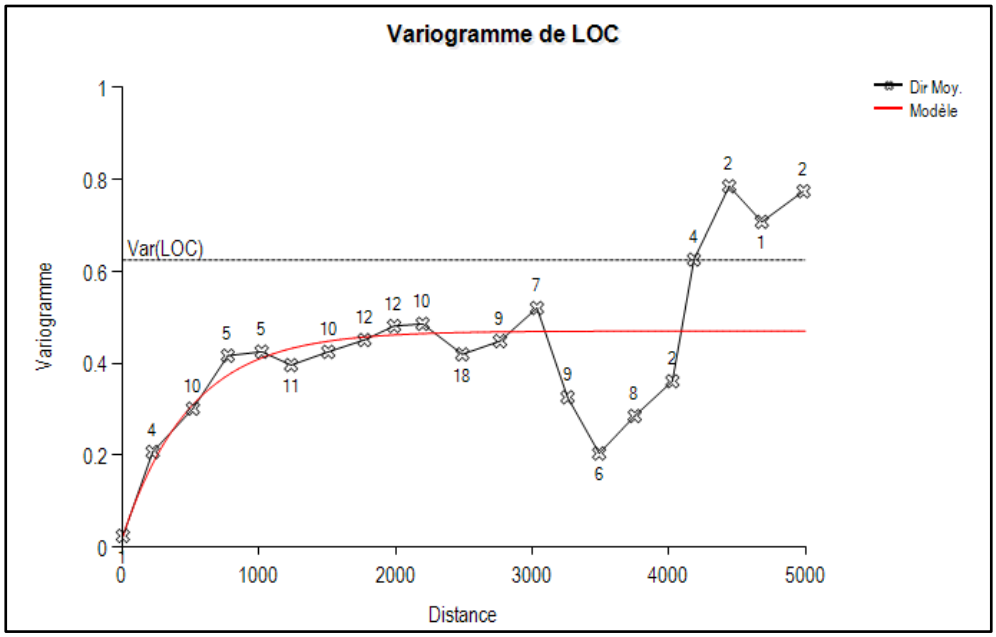


Figure 11 : Agglomération B - Variogramme de Ln(OCDD).

Ce variogramme repose toutefois sur très peu de couples et sa fiabilité est donc sujette à critique. On peut l'ajuster par un modèle exponentiel de portée pratique environ 1 500 m (3 fois le facteur d'échelle indiqué Figure 12).

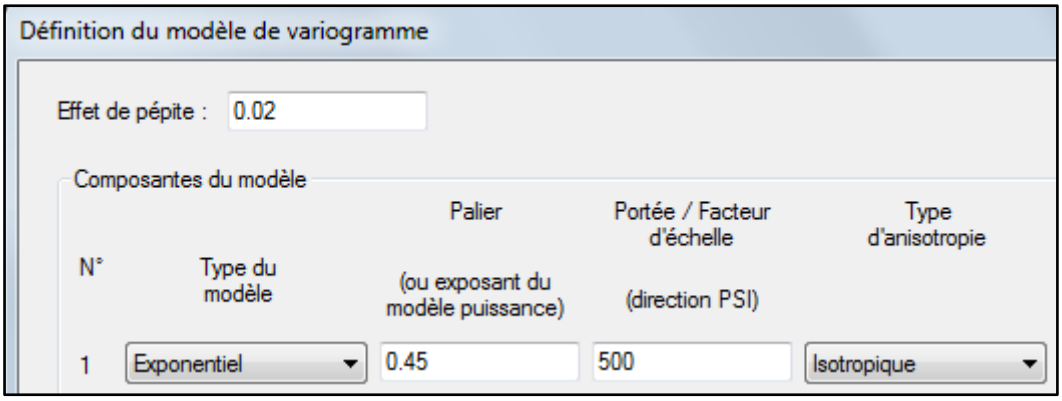


Figure 12 : Agglomération B - Modèle de variogramme de Log(OCDD).

1.2.3. Interpolation

Le logarithme des données pour l'OCDD est interpolé puis l'exponentielle du résultat est calculé. C'est-à-dire qu'on revient à la valeur brute par Z*=Exp(Y* + 0.5 Terme correctif) pour obtenir une estimation non biaisée de la moyenne.

L'interpolation est réalisée à la maille de 125 m pour avoir un bon rendu visuel. Le résultat est représenté Figure 13. La méthode d'interpolation est le krigeage, avec le modèle de variogramme ajusté précédemment.

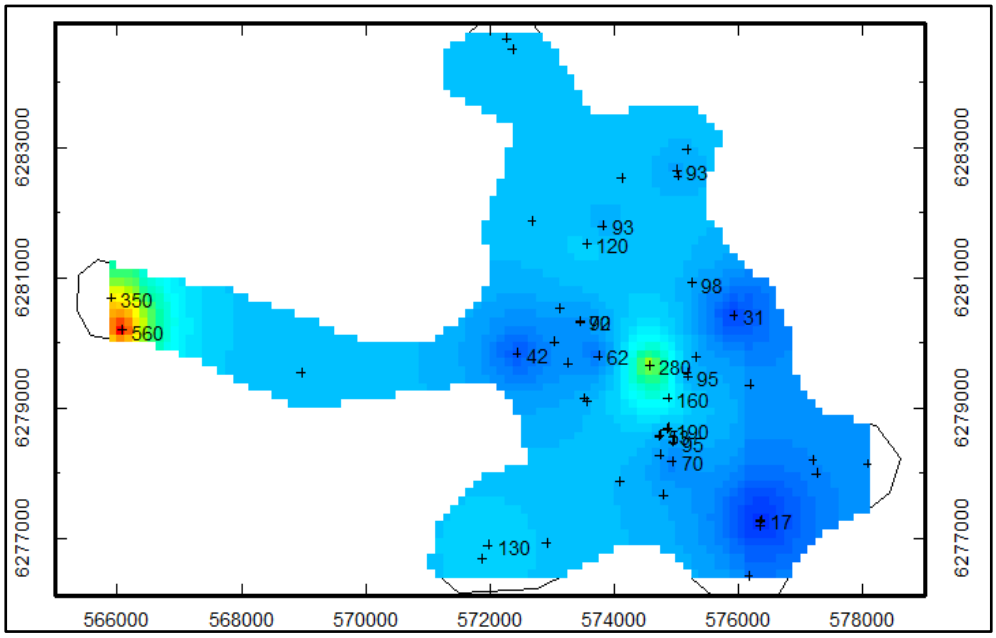


Figure 13 : Agglomération B - Interpolation de OCDD.

2. Données de l'agglomération C

2.1. Traitement des données arsenic

On dispose de 26 données (carte Figure 14). La concentration en As varie entre 1,5 et 17 mg/kg, pour une moyenne de 8,2 et un écart-type de 2,9. Il ne semble pas y avoir de répartition préférentielle des fortes/faibles valeurs.

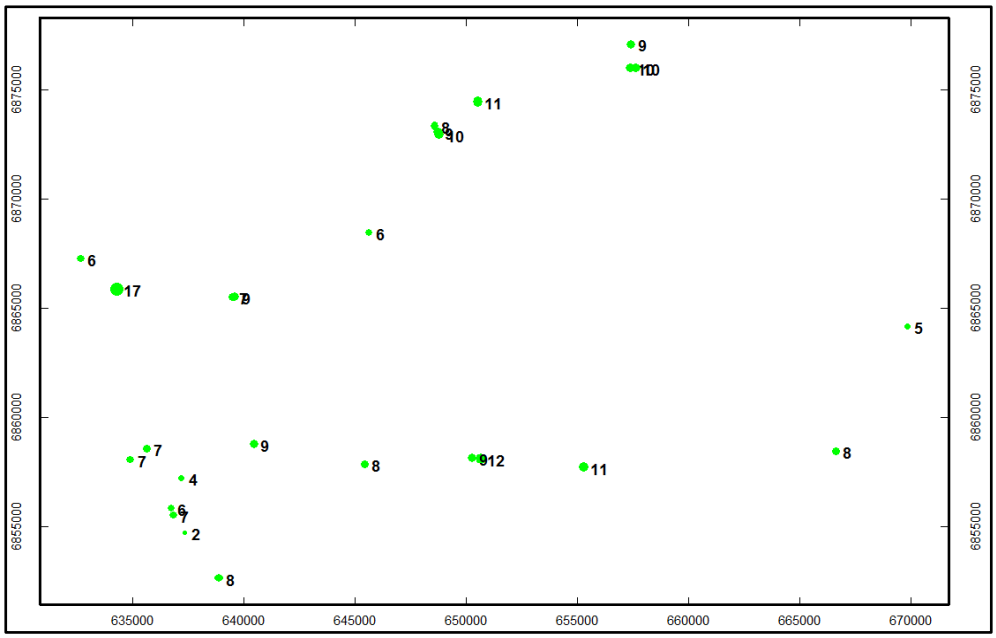


Figure 14 : Agglomération C - Carte des données pour As.

À noter que les données ne sont pas réparties de façon homogène, mais plus ou moins alignées selon 3 « branches » orientées respectivement N45, est-ouest et N150.

2.1.1. Histogramme

L'histogramme Figure 15 ne montre pas de répartition trop dissymétrique ni trop de valeurs extrêmes. La valeur la plus forte (17 mg/kg) se situe à l'ouest de la zone et n'est pas entourée de valeurs fortes. Elle constitue donc un point fort isolé.

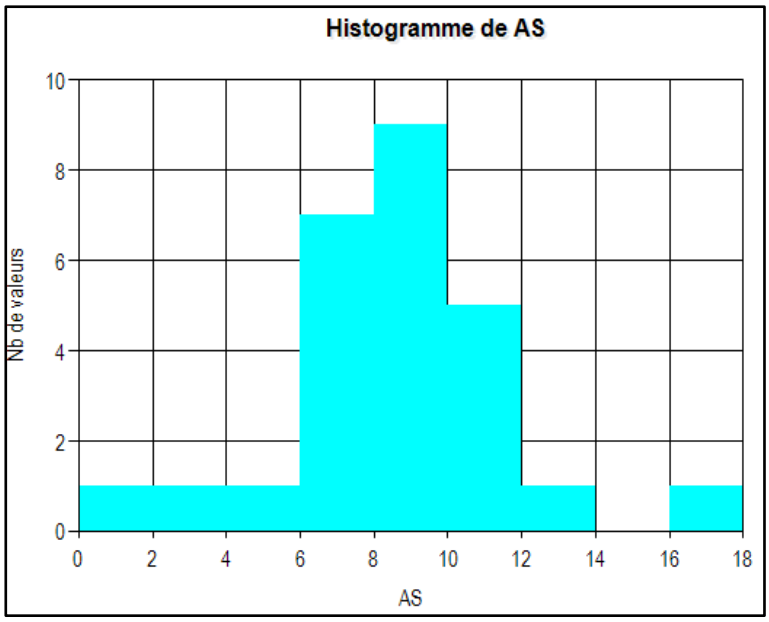


Figure 15 : Agglomération C - Histogramme de As (mg/kg).

2.1.2. Variogramme

Le variogramme directionnel ne fait pas apparaître de comportement anisotrope (Fig. 16). Toutefois ce variogramme repose sur très peu de données et de couples de points de calcul.

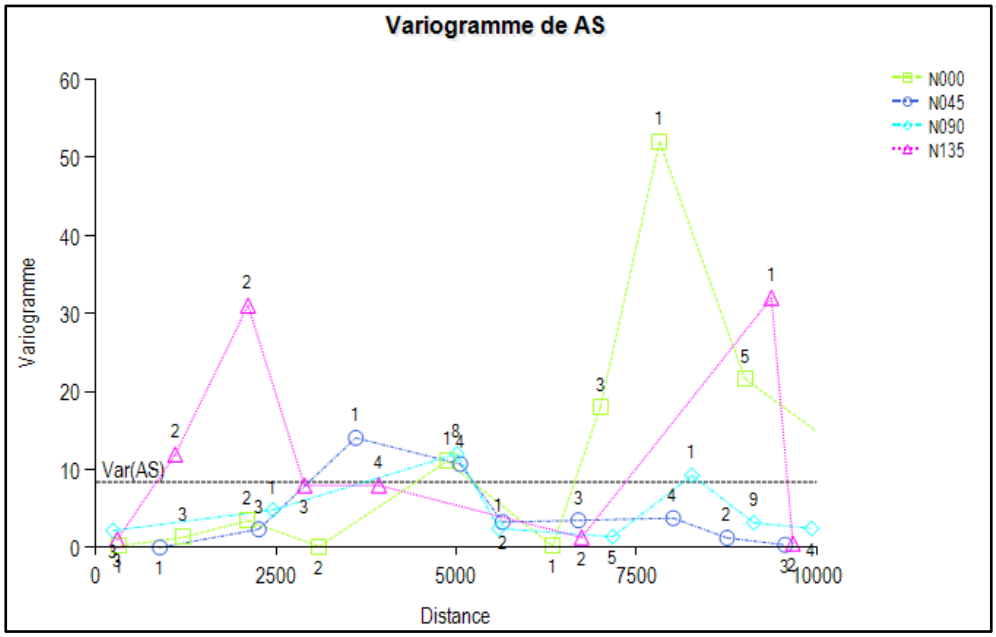


Figure 16 : Agglomération C. Variogramme directionnel de As.

En considérant le comportement comme isotrope, on obtient le variogramme toutes directions confondues représenté Figures 17 & 18. Ce variogramme apparait comme structuré, avec une portée voisine de 2 500 m. Sa représentativité est toutefois discutable en raison du faible nombre de données.

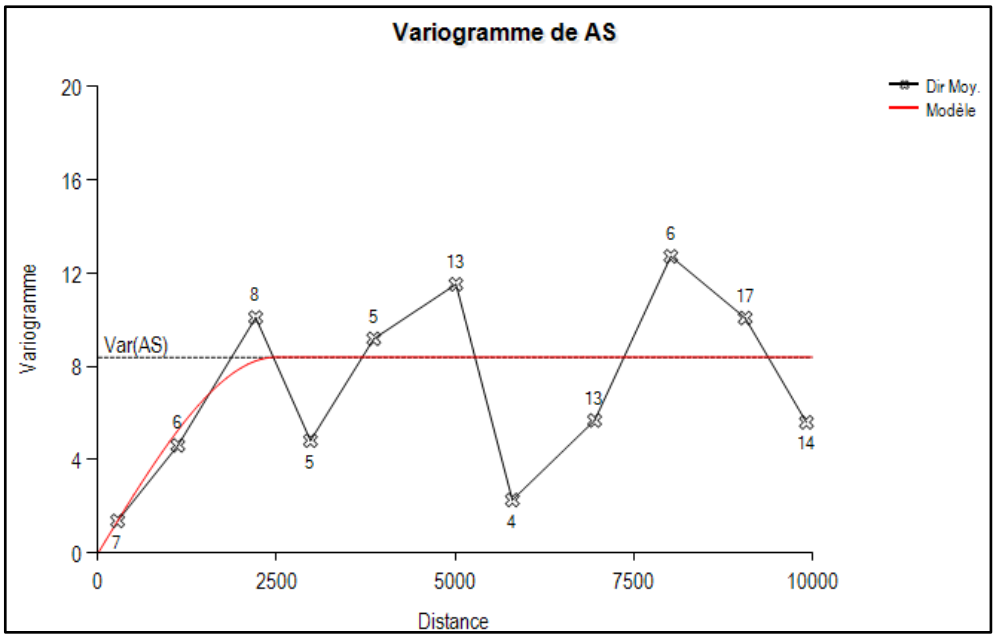


Figure 17 : Agglomération C - Variogramme de As toutes directions confondues.

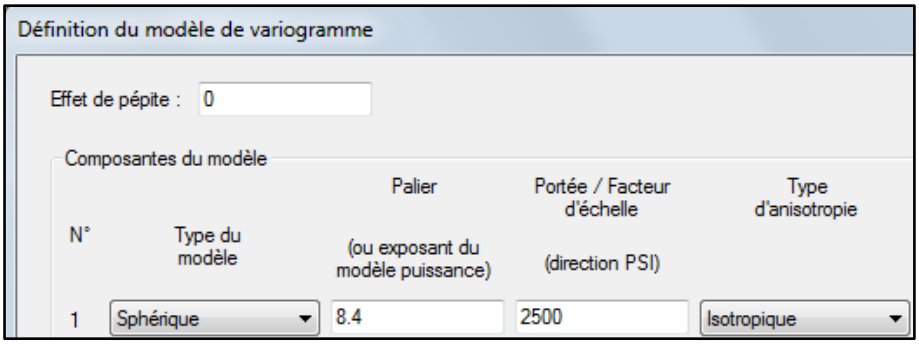


Figure 18 : Agglomération A - Modèle de variogramme pour As.

2.1.3. Validation croisée

Une validation croisée du modèle de variogramme est testée pour les données arsenic. Celle-ci donne d'assez bons résultats. Le point « non robuste » est le point isolé de concentration en As égale à 17 mg/kg, point qui avait déjà été signalé précédemment.

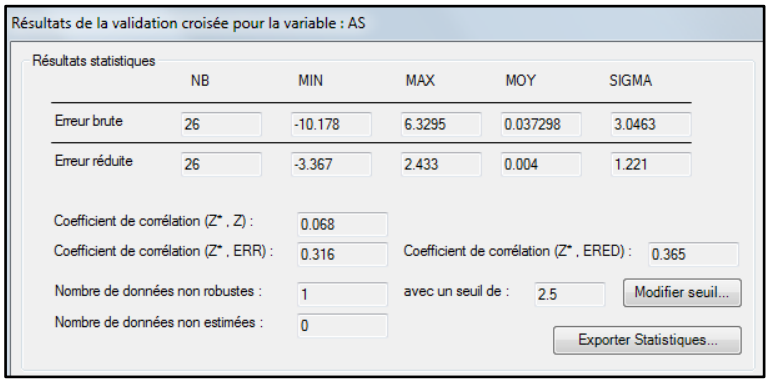


Figure 19 : Agglomération C - Validation croisée de As.

2.1.4. Interpolation

Une interpolation des concentrations en arsenic est tentée. L'interpolation est limitée par un polygone arbitraire entourant les données.

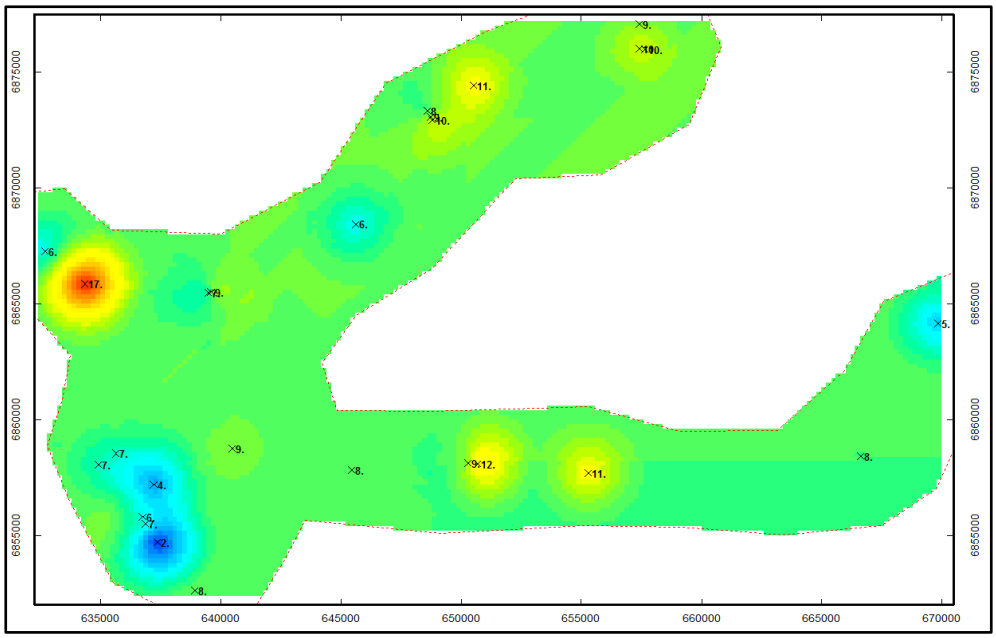


Figure 20 : Agglomération C - Interpolation de As.

2.2. Traitement des données dioxine OCDD

Pour OCDD, on ne dispose que de 14 points de mesure.

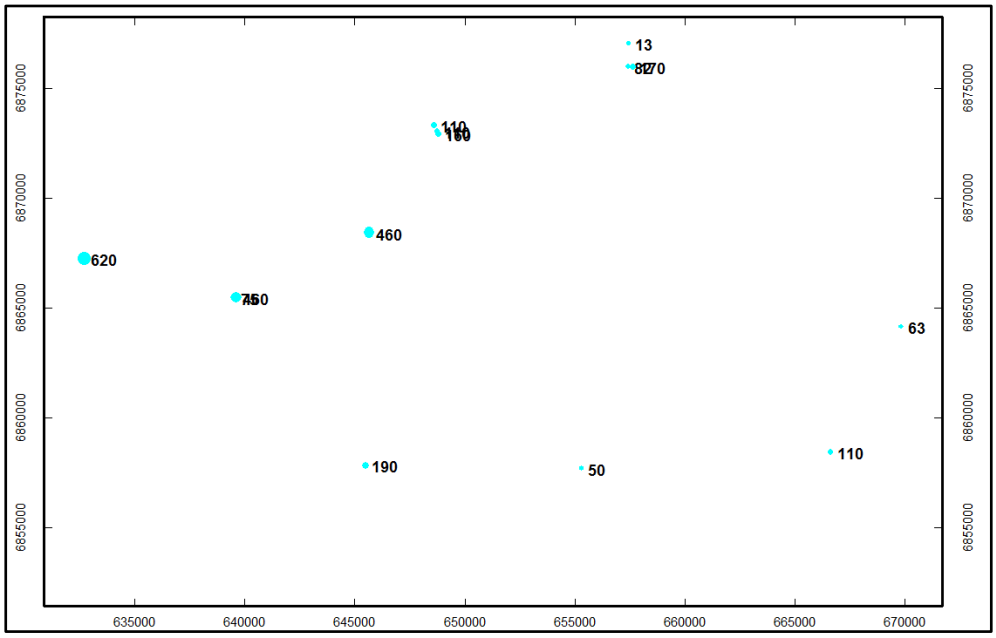


Figure 21 : Agglomération C - Carte des données pour OCDD.

2.2.1. Histogramme

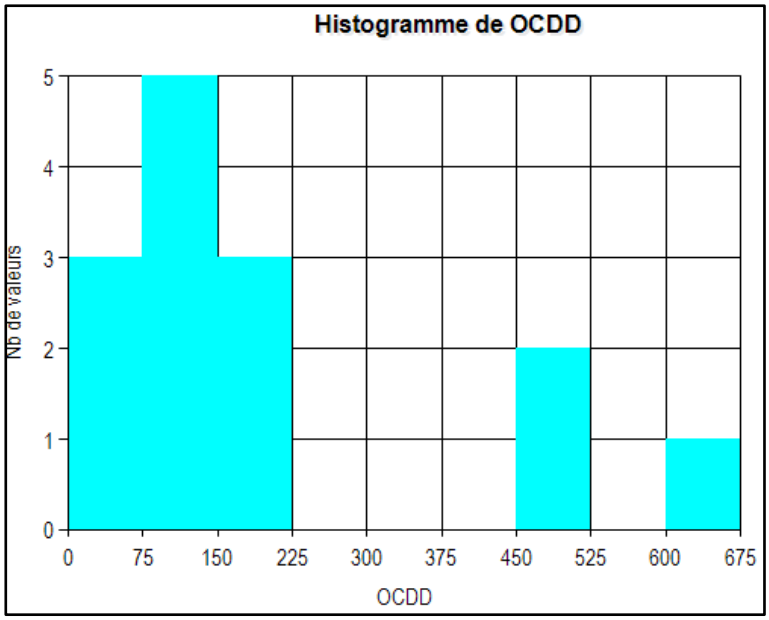


Figure 22 : Agglomération C - Histogramme de OCDD (ng/kg).

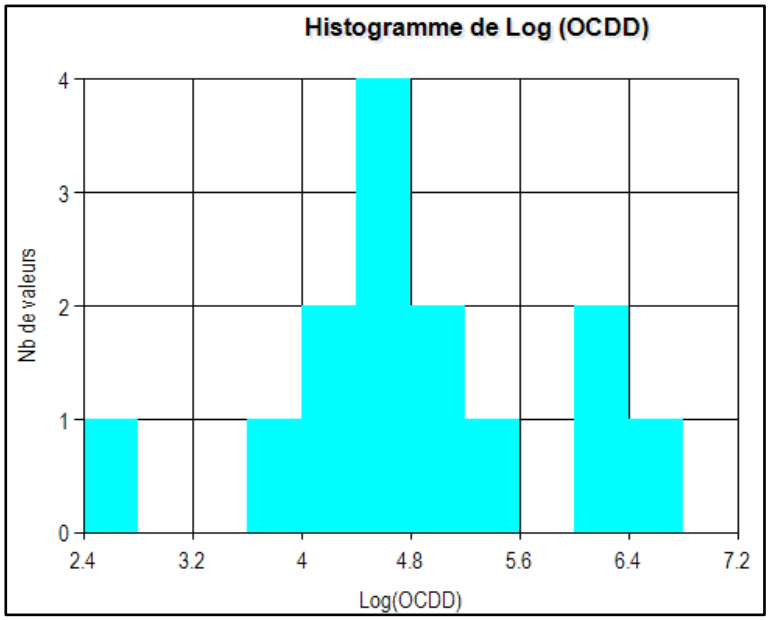


Figure 23 : Agglomération C - Histogramme de Log(OCDD).

L'histogramme en OCDD montre une répartition dissymétrique avec la présence de quelques fortes valeurs. En passant au logarithme, on obtient un histogramme plus symétrique.

2.2.2. Variogramme

Un variogramme est calculé. En raison du faible nombre de points, ce variogramme s'avère inexploitable. Il paraît donc impossible de proposer une cartographie de l'OCDD.

Annexe 2

Rapport de stage sur l'établissement d'un protocole de traitement statistique

POLYTECH PARIS UPMC - BRGM

Protocole d'Analyse Statistique pour la Construction d'un Fond Pédo-Géochimique Anthropisé des Sols Urbains

Mémoire de Stage

Sancho Ludovic

Soutenu le 13/09/2016

Durée du stage : 01/04/2016 – 30/09/2016

Tuteur Entreprise : Jean-François Brunet – Direction Eau, Environnement & Ecotechnologies – Unité Sites, Sols et Sédiments Pollués

Tuteur Ecole : Cyril Schamper – UPMC Sorbonne Universités / UMR 7619 Metis (ex Sisyphe)

Abstract

The quality of urban soils is a subject of concern in more and more countries. Necessary for rural areas, which provide food products from soils, it is also important for the urban sphere where a great part of the population lives. This is why the ministry in charge of environment started a program to identify the geochemical background[1] of French urban soils. It will improve diagnosis of pollution, management of health and excavated soils.

The aim of this paper is to provide a statistical protocol for the determination of soil background values thanks to a database managed by the French Geological Survey. Due to the sampling plan, the data collected (analysis of frequent geochemical substances in urban soils) present some characteristics that will affect the statistical treatment. The size of the dataset, values below the detection limit and the presence of outliers are the principal parameters to consider when dealing with geochemical data.

After a bibliographical study, some methods are proposed here for the treatment of data with such characteristics. It appears that the Kaplan-Meier, Maximum Likelihood Estimation and Regression on Order Statistics procedures allow the estimation of summary statistics in accordance with censored, small or skewed data distributions. However, checking normality of the data must be done before any of these solutions can be applied. Indeed, normality is a determining factor of most of the statistical tests/computations presented here.

As well as the computation of summary statistics, plotting the data must be done with care when dealing with the characteristics of the dataset mentioned. A combination of the histogram, density trace, boxplot and scatterplot is recommended to do so.

Only after these two steps, the computation of a background value can be considered. Likewise, many methods exist and are still tested, while this report is written, as well as methods to compare statistical populations and to determine the spatial range of a geochemical background's validity.

[1] *In french : « fond géochimique »*

Table des matières

Introduction

La connaissance de la qualité physico-chimique des sols est une préoccupation de plus en plus partagée par de nombreux pays. Jugée nécessaire depuis longtemps pour le milieu rural qui nous nourrit, elle devient indispensable aussi pour le milieu urbain où vit une part majoritaire de la population. La question de la qualité géochimique des sols urbains est accentuée pour de nombreuses agglomérations qui ont connu l'essor de la révolution industrielle ainsi que les destructions des deux guerres mondiales. Ces villes doivent aujourd'hui assumer un passif environnemental parfois lourd résultant de leurs activités artisanales et industrielles anciennes mais aussi de leur développement sur des sols mal connus, souvent constitués de remblais de qualité incertaine.

Ce passif environnemental a généré, au cours des siècles, des sols géochimiquement très divers résultant (1) de l'altération des roches puis (2) de leur évolution autonome en sols sous l'action de facteurs climatiques et biologiques, (3) de retombées atmosphériques diffuses d'origine anthropique et (4) de sources de pollution ponctuelles (Figure 1). Cette diversité définit la pluralité des fonds-géochimiques.

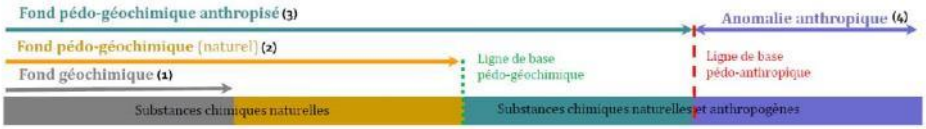


Figure 1 : Terminologie employée dans la détermination de valeurs de fond de la qualité des sols d'après GT BdF 2016

Cette notion permet de distinguer le fond pédo-géochimique anthropisé des anomalies géochimiques locales provenant de contaminations ou pollutions attribuables aux activités de sites urbains pollués. En effet, les sites industriels sont le plus souvent regroupés en zones industrielles et/ou au cœur d'un tissu urbain où un fond anthropique se superpose au fond géochimique naturel. Il convient alors de recueillir des données sur ce fond pédo-géochimique anthropisé.

La démarche française de gestion ces sites s'appuie sur la note ministérielle du 08 février 2007 éditée par le MEEM[2] [1]. La méthodologie mise en place est fondée sur la prévention et sur la gestion des risques suivant l'usage pour les pollutions passées. Toutefois, la France n'est pas dotée de valeurs réglementaires concernant la qualité des sols (potentiellement) pollués. Ainsi, en cas de suspicion de pollution, la démarche privilégie la comparaison de l'état du sol considéré à celui des sols « sains » voisins de la zone d'investigation.

Cette démarche intervient en cohérence avec la politique de gestion des terres excavées et susceptibles d'être réutilisées. Celles-ci doivent être caractérisées afin de vérifier si leurs propriétés chimiques sont compatibles avec le fond géochimique naturel local du site receveur [2]. Une terre est considérée exempte de pollution dès lors que ses caractéristiques sont cohérentes avec le fond géochimique naturel local [1]. Ainsi, une denrée alimentaire peut être consommée sans risque pour la population générale si elle satisfait aux exigences des critères de comestibilité retenus au niveau européen par les pouvoirs publics. De la même façon, un sol peut être considéré sans danger pour les

[2] *Ministère de l'Environnement, de l'Energie et de la Mer*

1

populations lorsqu'il est conforme à son état naturel initial, ou lorsqu'il est conforme à l'état d'un sol dont il est admis qu'il ne pose pas de problème particulier pour l'usage envisagé.

1 Contexte de travail

1.1 Structure d'accueil

Le BRGM[3] est l'établissement public de référence dans les applications des sciences de la terre pour gérer les ressources et les risques du sol et du sous-sol. Son action est orientée vers la recherche scientifique, l'appui aux politiques publiques et la coopération internationale. Autour de son cœur de métier qu'est la géologie, le BRGM développe une expertise dans le secteur de la gestion des ressources, de la maîtrise des risques et des écotechnologies innovantes. Cette activité s'articule en plusieurs thématiques, destinées à répondre aux différents enjeux industriels et sociétaux [3].

La Direction Eau, Environnement et Ecotechnologies (D3E) endosse deux de ces thématiques. La première couvre la gestion de l'eau, l'étude du fonctionnement et la préservation des hydrosystèmes, l'identification de nouvelles ressources ou encore l'impact du changement climatique sur leur qualité et les besoins des usagers. La deuxième thématique porte sur la surveillance et la réhabilitation des sites, sols et sédiments pollués (unité 3SP), la gestion des déchets ménagers, industriels et miniers.

1.2 Contexte technique

1.2.1 Bases de données existantes

En l'absence de valeurs de référence réglementaires pour les sols, la comparaison de la qualité des sols investigués avec celle des sols voisins doit être confortée par les informations contenues dans les bases de données suivantes [2]:

- BDETM : Base de Données des Éléments Traces Métalliques, établie dans le cadre de l'épandage des boues de station d'épuration (milieu rural) ;
- RMQS : Réseau de Mesure de la Qualité des Sols, établie dans le cadre de la surveillance des sols agricoles (milieu rural) ;
- ASPITET : Apport d'une Stratification Pédologique pour l'Interprétation des Teneurs en Éléments Traces ;

Ces bases sont établies et gérées par l'Institut National de la Recherche Agronomique (INRA[4]) et consultables sur le site du Groupement d'Intérêt Scientifique Sol (GisSol[5])

[3] *Bureau de Recherches Géologiques et Minières*

[4] *Institut National de la Recherche Agronomique* – institut.inra.fr

[5] *Groupement d'Intérêt Scientifique Sol* – www.gissol.fr

2

- IMN - Inventaire minier national, établi dans le cadre de l'exploration minière en milieu alluvial, géré par le BRGM et consultables sur le site InfoTerre[6].

Leur utilisation présente un certain nombre d'inconvénients. En effet, elles ont été construites dans des contextes et objectifs différents de ceux de la démarche nationale de gestion des sites et sols (potentiellement) pollués. En plus de ne pas couvrir la France entière, les prélèvements sont réalisés selon des protocoles variés, couvrent très peu les substances organiques et sont majoritairement localisés en milieu rural. Ces caractéristiques soulèvent des questions de comparabilité inter-bases mais également de comparabilité avec les particularités du domaine urbain.

1.2.2 Problèmes soulevés par le contexte urbain

La détermination d'un fond pédo-géochimique en milieu urbain implique la prise en compte de plusieurs spécificités. En plus des problèmes de comparabilité avec des données acquises en milieu rural, plusieurs spécificités sont à prendre en compte :

- En plus de l'absence de valeurs réglementaires pour les sols français, le domaine urbain est très peu couvert par les études relatant la qualité des sols. Les bases de données actuellement disponibles ne permettent pas d'établir un fond géochimique en domaine urbain.
- Au sein même d'une agglomération, la distinction de la pollution d'un site de celle imputable aux sites voisins et à l'activité de toute la ville est également une notion difficilement définissable et pourtant indispensable pour l'objectif recherché
- De plus, les remblais, très présents au sein des sites urbains investigués, n'ont rien en commun avec la roche mère locale à l'origine des sols initialement présents. Ils peuvent contenir des quantités importantes de substances indésirables (ex. : bitumes, scories, mâchefers). Le choix de lieux de prélèvement représentatifs de la pédo-géochimie locale doit tenir compte de leur présence.
- Enfin, de plus en plus d'aménageurs urbains émettent le besoin de mieux connaître la qualité des sols sur leur territoire. Dans le contexte de développement de l'économie circulaire, d'importants enjeux sont associés à la valorisation des terres excavées. Même si elles sont peu contaminées, celles-ci sont jusqu'à présent considérées comme des déchets et souvent mises en décharge. Par exemple, le projet du Grand Paris Express prévoit de « *valoriser le maximum de terres excavées en transformant les déchets en matière première selon le principe de l'économie circulaire* » (site internet : https://www.societedugrandparis.fr/).

[6] *InfoTerre* – infoterre.brgm.fr

3

1.3 Le projet FGU

1.3.1 Le projet Etablissements Sensibles

La mise en œuvre des « Diagnostics des sols dans les lieux accueillant des enfants ou des adolescents » (nom abrégé « ETS[7] ») par le ministère en charge de l'écologie, correspond à une opération préventive qui apparait au programme des PNSE[8] 2 et 3 (PNSE 2. 2009-2013 et PNSE 3. 2014-2016 – Actions 19 et 61, voir [4]). Au cours cette opération, plus de 2000 établissements font l'objet au cas par cas, de visites et de prélèvements spécifiques pour évaluer la qualité des milieux de vie des populations sensibles. ETS porte une attention particulière à l'exposition directe des populations les plus jeunes aux polluants par ingestion de sol suite à un porté main-bouche.

L'opération menée sur tout le territoire français concerne les établissements situés à proximité immédiate d'anciens sites industriels ou d'activités de service recensés dans l'inventaire *BASIAS*[9] (site internet : http://basias.brgm.fr/). Mais, si BASIAS fournit des informations sur les activités des sites industriels du passé, cette base de données ne permet pas en revanche de connaître l'état réel des sols.

Fidèles à la méthodologie nationale, les diagnostics ETS doivent faire appel à plusieurs prélèvements dits « témoins » réalisés sur des sites voisins, pour comparer les résultats des analyses de sols obtenues au droit des établissements.

Le diagnostic d'un établissement donne lieu à un prélèvement dans deux ou trois espaces verts situés à proximité (moins d'1km) du lieu d'échantillonnage. Les échantillons de sols au sein des établissements scolaires sont nommés SLE pour « SoL des Établissements » et ceux des espaces verts, servant de « témoins », sont nommés SLU pour « SoL Urbain ». Ces diagnostics ont impliqué neuf bureaux d'études et cinq laboratoires retenus par le BRGM, maître d'œuvre de l'opération ETS pour le MEEM.

[7] *Etablissements Sensibles*

[8] *Plans Nationaux Santé Environnement*

[9] *Base de données des Anciens Sites Industriels et Activités de Service*

4

1.3.2 Organisation du projet FGU

L'ADEME[10] et le BRGM ont signé en 2010 et 2014 deux conventions consécutives, d'une durée de 4 et 3 ans, visant l'établissement de référentiels des teneurs habituelles des principales substances minérales et organiques présentes dans les sols urbains en s'appuyant sur le projet ETS. Il s'agit du projet « Fonds Géochimiques Urbains » (FGU).

Les prélèvements SLU, réalisés au cours des diagnostics ETS, correspondent à la démarche visée par le projet FGU :

- Ils s'inscrivent dans une approche dite « thématique[11] » puisqu'il s'agit de bancariser les analyses de sols urbains exempts de toute pollution locale (spot) et uniquement impactés par une contamination diffuse. Les espaces verts, et préférentiellement les jardins publics, ont été retenus pour la réalisation de ces prélèvements car ils sont jugés *a priori* exempts d'impact polluant ponctuel, mais représentatifs du cumul des dépôts atmosphériques diffus urbains. En outre, ils sont plus accessibles aux équipes de préleveurs.
- Ils se focalisent sur les sols de surface (entre 0 et 5 cm de profondeur). Les effets dits « pépites » sont minimisés par des échantillonnages composites réalisés par la réunion de 5 prélèvements aux coins et au centre de carrés de trois mètres de côté.
- Ils sont prélevés dans les villes de plus de 5 000 habitants.

Les analyses de sols effectuées concernent les principaux éléments traces métalliques (cuivre, chrome, plomb, zinc, nickel, cadmium, mercure), un métalloïde (arsenic) et des substances persistantes organiques (cyanures totaux, hydrocarbures aromatiques polycycliques (HAP), polychlorobiphényles (PCB), polychlorodibenzo-p-dioxines (PCDD), polychlorodibenzo-furanes (PCDF). Ces résultats sont bancarisés grâce à l'outil BDSolU[12] constitué dans le cadre du projet et géré par le BRGM.

Outre l'amélioration des connaissances de la pédo-géochimie des sols en milieu urbain, cette base de données a pour objectif de servir aux différents acteurs impliqués dans la gestion des sites (potentiellement) pollués, notamment, dans le cadre :

- de la gestion sanitaire de l'exposition des populations aux sols ;
- du diagnostic de pollution ;
- de détermination de seuils de dépollution (en tenant compte également de l'usage envisagé du site) ;
- et de la gestion des terres excavées.

[10] *Agence de L'Environnement et de la Maîtrise de L'Energie*

[11] *Par opposition à un échantillonnage systématique qui consiste à prélever les sols systématiquement sur l'ensemble du territoire étudié, selon un maillage préétabli.*

[12] *Base de Données des Sols Urbains*

5

1.3.3 Présentation des résultats de la première convention

En Mai 2016, les prélèvements dans 278 villes de France métropolitaine avaient permis la bancarisation de 635 échantillons SLU. A ce stade trois agglomérations présentent un nombre d'analyses suffisant pour une exploitation. Par souci de confidentialité, le BRGM a choisi de noter ces agglomérations A, B et C. Les résultats obtenus semblent confirmer les hypothèses faites au commencement du projet :

- Premièrement, pour plusieurs substances, les valeurs obtenues semblent significativement différentes dans les trois agglomérations. Cela tend à indiquer que les agglomérations présentent un fond géochimique différent en fonction de leur climat, de leur histoire et des caractéristiques de leurs activités présentes ou passées.
- Deuxièmement, la confrontation des valeurs obtenues à celles des référentiels locaux (RMQS et BDETM), disponibles en milieu rural pour les éléments traces métalliques, montre que les teneurs des substances recherchées sont globalement plus élevées dans les agglomérations que dans les zones rurales environnantes.

Cette première convention aura aussi pointé plusieurs problèmes apparus à chaque étape de la démarche :

- la pertinence des choix méthodologiques effectués ;
- la représentativité des prélèvements ;
- et le traitement statistique des résultats.

En effet, les hypothèses ci-dessus restent à valider au moyen de teste statistiques fiables, robustes et adaptés au contexte de la démarche.

1.3.4 Objectifs du stage

Le stage effectué a pour vocation de chercher des réponses aux problèmes soulevés lors du traitement statistique des données recueillies.

Dans un premier temps, une étude bibliographique est menée afin de déterminer et analyser de façon critique les différentes pratiques statistiques en France et à l'étranger (essentiellement en Europe) dans le domaine de la détermination de la qualité des sols urbains.

Dans un second temps, les recherches effectuées doivent permettre de construire un protocole de traitement statistique pour la construction d'un fond pédo-géochimique anthropisé en suivant les grandes étapes suivantes :

- Préparation des données
- Réalisation de statistiques descriptives

- Interprétation des données et analyses des résultats

Chaque méthode proposée doit être en accord avec le contexte de réalisation d'un fond géochimique en milieu urbain ainsi qu'avec les caractéristiques des données recueillies.

Le stage ne vise pas le traitement géostatistique des données ni leur représentation cartographique. Si les données disponibles le permettent, ces deux étapes interviendront ultérieurement. La répartition spatiale des prélèvements doit néanmoins être prise en compte lors de la réflexion.

1.4 Analyse critique des traitements statistiques usuels

Afin de mener à bien une étude statistique, plusieurs paramètres doivent être réunis. La représentativité des données disponibles doit être scrupuleusement vérifiée, autrement elles n'auront aucune signification. Or, la construction du projet FGU, fortement conditionnée par celle du projet ETS, entraîne certaines conséquences sur la qualité des données bancarisées.

1.4.1 Effectif et répartition spatiale

Dans la littérature, les études statistiques classiques sont généralement considérées comme étant représentatives pour une population donnée à partir d'un effectif de 100 individus. La construction du plan d'échantillonnage du projet FGU étant thématique (voir 1.3.2), le nombre de prélèvements est lié au nombre d'établissements scolaires retenu par la démarche ETS. Or on estime que la démarche de prélèvement adoptée donnera lieu à cinq agglomérations présentant un potentiel d'échantillons de sols SLU de plus de 30 individus (limite inférieure acceptable de l'effectif des populations pour qu'un fond géochimique puisse être déterminé [5], [6], [7]). D'emblée on sait que le projet ETS, apportera des données à orientation plutôt sanitaire. Il ne pourra pas, à lui seul, fournir le volume de données espéré pour déterminer le ou les référentiels recherchés à l'échelle de chaque agglomération française par le projet FGU.

Actuellement, seule l'agglomération C présente un effectif proche du minimum requis pour des statistiques classiques avec 97 individus. Toutefois, les agglomérations A et B, présentant les effectifs les plus élevés après l'agglomération C, contiennent respectivement 30 et 48 échantillons.

De plus, ces effectifs réduits ont une influence non négligeable sur le calcul d'un fond géochimique spatialisé. En effet, la majorité des jeux de données dans le domaine des sciences de la terre appliquées diffèrent de ceux rencontrés dans les autres disciplines scientifiques parce qu'ils possèdent une composante spatiale. Chaque individu/échantillon est caractérisé par des résultats d'analyses mais aussi par des coordonnées géographiques. Le plan d'échantillonnage étant, par définition, thématique, la répartition spatiale des points de prélèvements est hétérogène. Donc

7

certains secteurs posséderont *in fine* une concentration de points plus élevée que d'autres (exemple Figure 2).

Figure 2 : Exemple de répartition spatiale des points de prélèvements

Les résultats montrent également une forte variabilité des teneurs entre des points séparés par seulement quelques centaines de mètres. Cette observation corrobore celles souvent décrites dans la littérature pour le milieu urbain.

Le résultat d'un fond géochimique se présentant habituellement par une valeur ou un ensemble de valeurs numériques, peut-on calculer des paramètres statistiques représentatifs d'une zone donnée ? La délimitation de cette zone serait également à définir. Jusqu'à maintenant, les agglomérations sont utilisées en tant que délimitation géographique [2]. Cette décision peut-elle être confirmée/infirmée par des tests statistiques ? Le fond pédo-géochimique anthropique est-il statistiquement différent d'une agglomération à une autre ?

1.4.2 Limite de Quantification

Un autre problème récurrent de l'analyse statistique de données géochimiques vient du fait que chaque méthode analytique comporte une limite de quantification (LQ) inférieure et une LQ supérieure. Le fichier reçu du laboratoire contiendra parfois un certain pourcentage de données inférieures à la LQ.

La limite de quantification inférieure[13] est généralement comprise comme étant la plus faible concentration pouvant être mesurée de manière représentative avec une méthode d'analyse donnée [5]. *Un jeu de données est dit* ***censuré*** *s'il contient des valeurs inférieures à la limite de*

[13] La limite de quantification supérieure est peu courante dans le contexte de l'établissement d'un fond géochimique mais est définie en intervertissant les termes suivants : « inférieure » avec « supérieure » ; « faible » avec « élevée »

*quantification. On utilisera les termes « **censure inférieure** » pour qualifier un jeu de données censuré uniquement pour des valeurs faibles.*

Une même méthode peut présenter différentes LQ en fonction des opérations de préparation nécessaires pour s'adapter aux concentrations variables des échantillons. La LQ changera également en fonction de l'instrument de mesure utilisé et du technicien. Pour le même élément, différentes méthodes d'analyse sont possibles, ce qui conduit inéluctablement à des LQ différentes. Ces paramètres sont valables lorsqu'un laboratoire unique est impliqué dans le fonctionnement du projet. Avec des laboratoires multiples, comme c'est le cas du projet FGU, les facteurs de variabilité des LQ reportées augmentent. Une même technique d'analyse peut engendrer des LQ différentes pour différents laboratoires.

Tableau 1 : Teneur en Cadmium des sols de l'agglomération A

Cd (mg/kg)
1.37
1.2
0.82
0.76
0.73
0.7
0.7
0.67
0.6
0.57
0.56
0.55
0.5
<0.5
0.48
0.46
0.36
0.36
0.31
<0.2
0.2
<0.2
0.19
0.15
<0.1
<0.1
<0.1
<0.1
<0.1
<0.1

En géochimie environnementale, la censure inférieure fait plus souvent l'objet d'attention que la censure supérieure [5] puisque les substances analysées sont normalement présentes en faible concentration, voir des concentrations inquantifiables, dans les sols (exceptée en cas d'étude d'un minerai ou d'une forte pollution). Le jeu de données résultant est donc censuré inférieurement (exemple Tableau 1).

Le Tableau 2 se rapporte au jeu de données de l'agglomération A. Il présente les effectifs et le pourcentage de valeurs inférieures à la LQ pour quelques substances sélectionnées de façon à montrer la diversité entre et au sein des familles. Les pourcentages varient plus ou moins fortement en fonction des substances/familles analysées. A la vue de ces résultats, il devient évident que les données (celles relatives à l'arsenic et à l'acénaphtylène, par exemple) ne peuvent être traitées statistiquement de manière identique. Cela risquerait de créer un biais lors du calcul des statistiques basiques, ce qui entraînerait par la suite un résultat erroné lors des tests multidimensionnels comme l'ACP[14].

De manière générale, ce problème est résolu soit en substituant toutes les valeurs inférieures à la LQ par une fraction de la valeur initiale (en général ½) soit en supprimant les valeurs inférieures à la LQ ([5], [8]). Cette habitude provient sûrement du domaine minier, pionnier dans la réflexion sur les valeurs censurées, où l'intérêt est porté aux valeurs élevées indicatrices d'un gisement. La substitution des valeurs faibles n'a que peu d'influence sur ces extrêmes recherchés. Il n'est pas acquis que cette méthode n'ait aucune influence sur la recherche d'un fond géochimique où ces valeurs faibles sont au contraire très étudiées.

Examinons l'influence de cette méthode sur les analyses de plomb de l'agglomération C. L'effectif de la population est de 97, on s'affranchit donc des problèmes que pourraient créer un effectif trop faible. Pour les besoins du test, le jeu de données est censuré artificiellement : les 30 valeurs les plus faibles sont remplacées par les valeurs de LQ/2 reportées

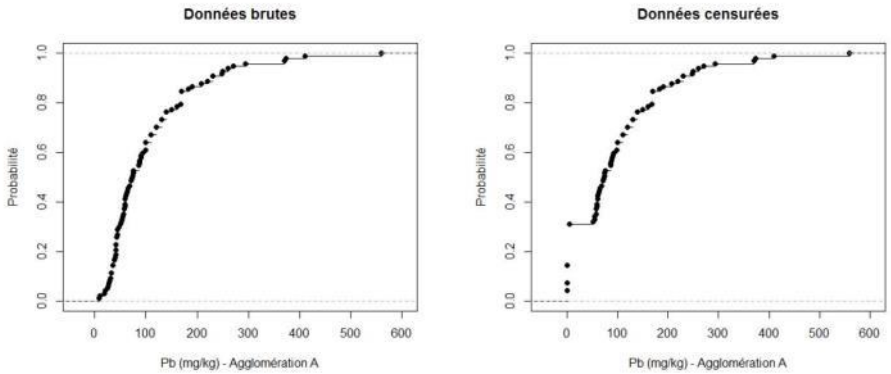
[14] Analyse en Composantes Principales

pour l'échantillon par les laboratoires. On obtient donc un pourcentage de 31% de censure, comparable à l'exemple du cadmium ou du 2,3,7,8-TCDF de l'agglomération A (voir Tableau 2).

Tableau 2 : Pourcentage de valeurs inférieures à la limite de quantification pour quelques substances

Famille	Composé	Nombre d'échantillons	Nombre de valeurs inférieures à la LQ	Pourcentage de valeurs inférieures à la LQ
Métaux-Métalloïdes	Arsenic	30	3	10%
Métaux-Métalloïdes	Plomb	30	0	0%
Métaux-Métalloïdes	Cadmium	30	9	30%
HAP	Acénaphtylène	30	24	80%
PCB indicateurs	PCB n°138	30	21	70%
PCDD / PCDF	1,2,3,7,8,9-HxCDD	12	5	42%
PCDD / PCDF	2,3,7,8-TCDF	12	4	33%
PCDD / PCDF	OCDF	12	2	17%

Les fonctions de répartition empiriques (équivalent d'une fonction de répartition mais sans biais, voir Fiche 8) sont tracées en Figure 3 pour étudier la distribution des populations sélectionnées. Les deux courbes sont identiques à l'exception de la portion inférieure à 50 mg/kg. La censure tronque une partie de l'information. Les points restant correspondent aux valeurs des LQ, qui sont ici multiples (0.5 ; 0.7 ; 1 ; 10). Les 16 valeurs associées à la LQ de 10, par exemple, sont donc superposées sur le point représentant la LQ.



(a) (b)

Figure 3 : Fonction de répartition empirique (a) des données de plomb de l'agglomération A (b) des mêmes données censurées avec traitement par substitution à 50% de la LQ

L'information perdue a des conséquences sur les paramètres statistiques de la distribution. Leur calcul (voir Tableau 3) permet ici de mettre en évidence l'influence de la censure sur la moyenne et le premier quartile. Il en est de même pour tout estimateur découlant de ces paramètres descriptifs mais également pour tout test statistique ultérieur. On rappelle qu'ici les données utilisées

correspondent à une population avec un effectif proche de 100 et une censure de 30%, l'impact de cette perte d'informations est encore plus prononcé quand il s'agit d'effectifs réduits et de taux de censure supérieurs. La poursuite de l'étude ne peut être réalisée sans trouver une alternative au problème posé par les LQ.

Tableau 3 : Statistiques basiques des données de plomb de l'agglomération A

	Taux de censure	Minimum	1er Quartile	Médiane	Moyenne	3ème Quartile	Maximum
Données brutes	0%	8.7	43	74	108.6	140	560
Données censurées	31%	0.25	5	74	98.58	140	560

Effectif : 97 échantillons

1.4.3 Distribution et normalité

La distribution statistique d'une population est une caractéristique permettant de rendre compte de la répartition des données. La majeure partie des tests/méthodes statistiques reposent sur l'hypothèse que les variables décrivant les individus suivent une distribution normale ; ce qui n'est pas forcément le cas pour les distributions des différentes substances analysées dans le cadre du projet FGU. Les tests non-paramétriques à l'inverse s'affranchissent de cette hypothèse et d'autres encore sont qualifiés de robustes, c'est-à-dire que même si l'on s'écarte légèrement des conditions d'applications initiales, ils restent valables [9]. Nous aborderons plus loin quelques-uns de ces tests en détail.

Le théorème central limite stipule que la somme de variables aléatoires de même moyenne et écart type tend vers la loi normale (ou loi de Gauss) et ce quel que soit la distribution initiale. La courbe de distribution de la loi normale, appelée courbe de Gauss ou plus communément « courbe en cloche » (Figure 4), est symétrique. Lorsque l'effectif d'une population augmente, ce théorème prend toute son importance. En effet, il explique la popularité des traitements statistiques adaptés à des populations dont la distribution suit une loi normale [9].

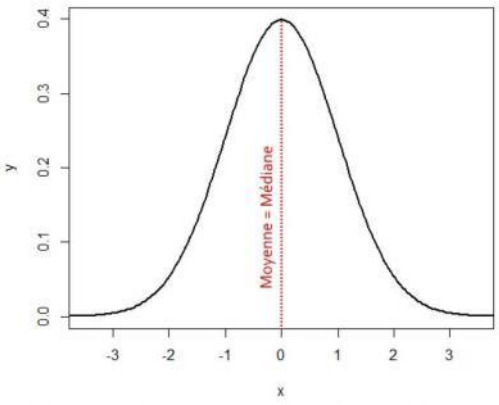


Figure 4 : Courbe de distribution de la loi normale $\mathcal{N}(0,1)$

Un test de normalité (Shapiro-Wilk voir Fiche 1) est réalisé sur les analyses de plomb des agglomérations A, B (malgré leur faible effectif) et C (Tableau 4).

La p-value[15] résultante est de 0,042 pour l'agglomération A et est inférieure à 1.10^{-4}

[15] *La p-value doit être supérieure à « α=0,05 » pour valider la normalité (Fiche 1)*

pour les deux autres agglomérations. La normalité est donc rejetée pour les trois agglomérations, autrement dit la distribution de la population « plomb » n'est pas une distribution normale.

Tableau 4 : Résultat du test de normalité pour les analyses de plomb des agglomérations A, B et C

Agglomération	p-value ($>\alpha = 0.05$)
A (30 échantillons)	0.042
B (48 échantillons)	< 0.0001
C (97 échantillons)	< 0.0001

(rouge : distribution non normale ; vert : distribution normale)

Comme nous l'avons vu ci-dessus, la majeure partie des tests/méthodes statistiques reposent sur l'hypothèse que les variables décrivant les individus suivent une distribution normale. Ce résultat aura donc une conséquence sur la suite de l'étude statistique.

Le test de Shapiro-Wilk utilisé ici possède des équivalents. Ils sont sensibles à l'effectif de la population [10] et à la proportion de valeurs inférieures à la LQ [11]. Dans l'objectif de ne traiter qu'une problématique à la fois, les calculs effectués pour obtenir les résultats du Tableau 4 ont été réalisés avec les analyses de plomb qui ne présentent pas de valeurs inférieures à la LQ. Même avec cette précaution, on observe l'influence de l'effectif réduit (30 échantillons) des analyses de l'agglomération A sur le calcul de la p-value. En effet, la puissance du test de normalité diminue avec le nombre de valeurs disponibles [10]. Par exemple, une population de 30 individus suivant une distribution log-normale aura une p-value plus élevée, ici plus proche de satisfaire le critère « supérieur à α », qu'une population similaire (toujours log-normale) avec un effectif de 100 individus.

Le second effet, concernant la proportion de valeurs inférieures à la LQ, est visible sur la Figure 5. Andersson et Burberg ont réalisé des simulations avec une population de 20 observations suivant une distribution normale [11]. Des valeurs de LQ sont imposées artificiellement afin d'obtenir des jeux de données présentant différents degrés de censures inférieures[16] (5, 20, 40, 60 et 80 pourcent) ; les valeurs censurées sont remplacées par la valeur de LQ divisée par 2 (substitution à 50%). La normalité de la population est testée de façon itérative (10000 répétitions). La distribution de la population d'origine étant normale un rejet de la normalité correspond à une erreur du test. Chaque rejet est comptabilisé et l'axe vertical de la Figure 5 représente le taux d'erreur du test de normalité c.à.d. le résultat de la fraction suivante :

$$Taux\ d'erreur = \frac{Nombre\ de\ rejet\ de\ l'hypothèse\ nulle}{Nombre\ de\ validation\ de\ l'hypothèse\ nulle}$$

[16] *Voir définition* à la *section 1.4.2*

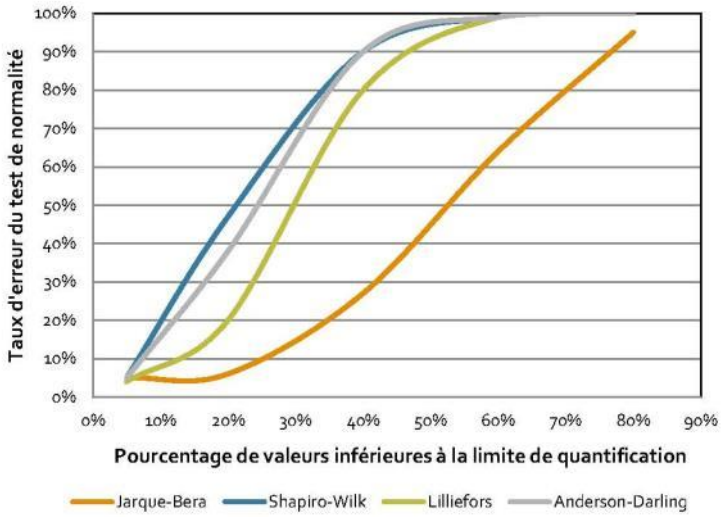


Figure 5 : Influence de la proportion de valeurs inférieures à la LQ sur les tests de normalité ([11], modifié)

Le test de Shapiro-Wilk est le plus sensible au pourcentage de valeurs inférieures à la LQ alors que le test de Jarque-Bera est le plus tolérant. Les tests de Lilliefors et Anderson-Darling sont compris entre les deux précédents. Ce graphique mettant en évidence la variabilité de la robustesse des tests statistiques face à la censure d'un jeu de données soulève la question de la sélection du test le mieux adapté au traitement de données environnementales.

1.4.4 Détection des outliers

La détection des outliers est une des tâches clef de l'analyse statistique de données géochimiques classiques. Elle permet de détecter des processus géochimiques rares souvent indicateurs de gisement potentiellement exploitable.

En géochimie environnementale, les outliers sont définis statistiquement (Hampel et al., 1986 ; Barnett and Lewis, 1994 ; Maronna et al., 2006 cité dans [5]) comme : « *valeurs appartenant à une population différente car elles sont originaires d'un autre processus/source, i.e elles proviennent d'une distribution contaminée* ». On peut ajouter à cette définition les valeurs provenant d'une erreur opératoire au cours de l'acquisition/traitement des données tel qu'une erreur de frappe (typiquement un ajout de zéro supplémentaire). Les outliers sont souvent très élevés et provoquent l'asymétrie positive (voir Figure 6) de la distribution de la population étudiée. Dans l'optique de définir un fond pédo-géochimique, les outliers représentent plutôt une pollution dont il faudrait s'affranchir dans la distribution de la population.

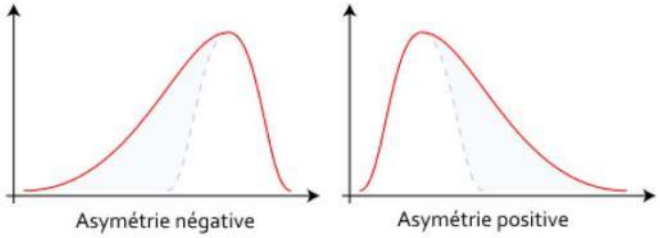


Figure 6 : Distribution asymétrique négative (gauche) et distribution asymétrique positive (droite)

Une solution des statistiques classiques (Gaussiennes) souvent utilisée est celle de la $MOY \pm 2.\mathrm{ET}$[17] [12] produisant une estimation d'un seuil séparant les valeurs appartenant à la population des outliers. Cependant, les outliers peuvent apparaître de façon erratique dans une distribution donnée, pas uniquement aux extrémités, et peuvent être difficile à identifier. La méthode $MOY \pm 2.\mathrm{ET}$ semble fonctionner parce que fort heureusement, les outliers sont souvent très élevés (ou faibles) par rapport à la majeure partie des données [12]. En pratique le terme « outlier » est souvent utilisé pour n'importe quel type de valeur s'écartant de la distribution étudiée.

Cependant cette méthode repose sur une distribution normale des données. Or, en géochimie environnementale, les distributions asymétriques positives sont prépondérantes à cause du problème posé par les outliers. Il est donc nécessaire de trouver une ou des méthode(s) adaptées aux caractéristiques des données disponibles.

[17] *Moyenne plus ou moins deux fois l'écart-type ($MEAN \pm 2.SD$ en anglais)*

2 Etude bibliographique des méthodes statistiques en domaine environnemental

L'ensemble des résultats d'analyses d'échantillons de sols SLU bancarisés par le BRGM dans le cadre du projet FGU s'élève à 8915 en mai 2016. Ces résultats doivent être étudiés et comparés tout en considérant leurs caractéristiques particulières. De plus, les points ayant été prélevés avec une répartition géographique spécifique à chaque agglomération, deux études statistiques doivent être menées séparément : l'une concernant la distribution spatiale des données sur la zone d'étude et l'autre concernant la distribution statistique des valeurs mesurées. Plusieurs méthodes sont décrites dans la littérature, il convient de les étudier et de valider ou non leur utilisation dans l'objectif d'établir un fond pédo-géochimique.

2.1 Préparation des données

2.1.1 Distribution et normalité

En géochimie environnementale, la transformation-log est largement utilisée dans l'objectif approcher une distribution normale [5] [8]. Elle correspond à une réexpression des données dans une nouvelle unité. Cette unité altère les distances entre les observations une fois tracées sur un graphique. L'effet est soit d'augmenter, soit de diminuer les distances entre les valeurs extrêmes et la médiane, ce qui rend la courbe de distribution du jeu de données utilisé plus symétrique [13].

Toutefois la transformation-log n'est qu'un cas particulier de l'Echelle des puissances de Velleman et Hoaglin [13] qui caractérise les fonctions de la forme :

$$y \leftarrow \begin{cases} x^{\lambda} & (\lambda \neq 0) \\ \log(x) & (\lambda = 0) \end{cases}$$

où x est la donnée brute, y la donnée transformée et λ la puissance.

Comme on peut le voir dans le Tableau 5, les transformations utilisant une puissance $\lambda < 1$ sont utiles pour rendre symétriques les courbes de distributions étirées vers les valeurs hautes *« asymétrique-positive »*. Tandis qu'une puissance $\lambda > 1$ permettra de corriger les distributions étirées vers les valeurs faibles *« asymétrique-négative »*.

Etant donné que la majorité des populations statistiques en domaine environnemental présentent une asymétrie positive, les puissances inférieures à 1 présentent une utilité. La puissance $\lambda = 0$ correspond à la transformation logarithmique. Mais la transformation $x^{1/2}$ est également utilisée dans le domaine environnemental.

Tableau 5: Echelle des puissances de Velleman et Hoaglin (modifié depuis [13])

Utilisation	λ	Transformation	Nom	Commentaire
		...		Des puissances supérieures peuvent être utilisées
Asymétrie (-)	3	x^3	Cube	
	2	x^2	Carré	
	1	x	Unité d'origine	Pas de transformation
	½	$\sqrt{x}$	Racine carrée	Fréquemment utilisée
	1/3	$\sqrt[3]{x}$	Racine cubique	Fréquemment utilisée
	0	$\log(x)$	Logarithme	Fréquemment utilisée
Asymétrie (+)	-1/2	$-1/\sqrt{x}$	Racine carrée de l'inverse	Le signe moins préserve l'ordre des observations
	-1	$-1/x$	Inverse	
	-2	$-1/x^2$		
		...		Des puissances inférieures peuvent être utilisées

Le principe de la transformation Box-Cox est identique à celui de l'échelle de Velleman et Hoaglin. Plusieurs valeurs de λ sont testées à travers l'équation suivante :

$$y \leftarrow \begin{cases} \frac{x^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log(x) & (\lambda = 0) \end{cases} \quad [14]$$

où x est la donnée brute, y la donnée transformée et λ la puissance.

La puissance λ estimée permet d'atteindre la distribution se rapprochant au maximum de la normalité du jeu de données transformé. La transformation Box-Cox constitue une amélioration de la transformation logarithmique. Cette dernière est toujours représentée lorsque le résultat de l'estimation est $\lambda = 0$.

L'effet de la transformation Box-Cox sur les analyses de plomb de l'agglomération B est visible en comparant les statistiques graphiques produites avec les données brutes, les données log-transformées et Box-Cox-transformées (Figure 7). On peut voir :

- le centrage de la courbe de distribution autour de la médiane ;
- ainsi que la diminution de la distance entre la moyenne et la médiane (respectivement la barre verticale et la croix rouge sur les boxplots).

16

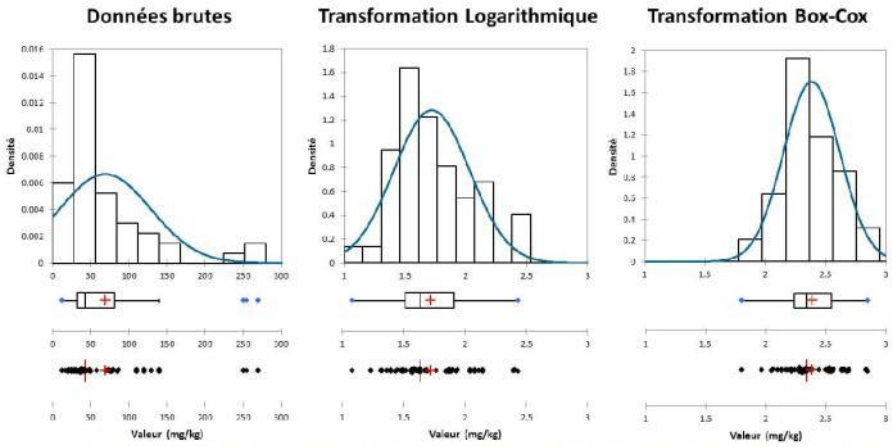


Figure 7 : Statistiques descriptives et tests de normalité pour les données de plomb de l'agglomération B (48 échantillons)

Pour confirmer l'efficacité de la transformation Box-Cox, les données de l'agglomération B sont soumises à un test de normalité (Shapiro-Wilk) avant et après transformation (Tableau 6). Par exemple, la p-value des données brutes de plomb de l'agglomération B est **<0,001** (on rappelle qu'elle doit être supérieure à 0,05) les données ne suivent donc pas une distribution normale. La transformation logarithmique permet d'atteindre une p-value de **0,128**, on obtient donc une distribution normale. En utilisant la transformation Box-Cox ($\lambda = -0{,}276$) la normalité de la distribution est améliorée, on obtient une p-value de **0,584**.

Tableau 6 : Comparaison des différentes méthodes de transformation par rapport à la normalité (analyses de plomb des agglomérations A, B et C)

Agglomération	Données brutes	Transformation Logarithmique	Transformation Box-Cox
A (30 échantillons)	0.042	0.137	0.421
B (48 échantillons)	< 0.0001	0.128	0.584
C (97 échantillons)	< 0.0001	0.635	0.637

rouge : distribution non normale; vert : distribution normale)

Les mêmes transformations réalisées sur l'agglomération C contenant 97 échantillons produit des résultats confortant le fait qu'à partir d'un effectif de 100 échantillons, une population statistique peut être traitée grâce aux traitements classiques. En effet, la p-value du test de normalité n'est modifiée qu'au centième près entre la transformation logarithmique et la

17

transformation Box-Cox (Tableau 6). Tandis que pour les données de l'agglomération A avec 30 échantillons, la p-value est multipliée par 3.

Ce résultat met donc en valeur l'importance d'utiliser un outil plus précis que la transformation logarithmique dans le cas de traitement d'échantillons à faibles effectifs. Cette précision sera nécessaire plus loin lors des calculs/tests statistiques multidimensionnels.

2.1.2 Centrage et réduction

Le traitement de données environnementales nécessite de prendre en compte des substances différentes simultanément, c'est le cas du projet FGU. Pour assurer la comparabilité de ces variables les unes avec les autres, il est parfois utile de recourir à des transformations.

Avant de réaliser certains tests multidimensionnels comme l'ACP (voir section 2.2.3), il est préférable que les données soient centrées autour d'une même valeur. Cette étape intervient après le traitement de la normalité des données par une transformation Box-Cox [5]. L'opération de centrage consiste en la soustraction de la moyenne (ou la médiane) d'un jeu de données d'une substance à chaque valeur initiale.

Les données peuvent cependant rester incomparable directement de par leur écart-type [5]. En effet, certaines familles d'éléments, présentent des gammes de concentrations plus faibles ou plus élevées que d'autres. Ce paramètre est très influent sur le résultat des tests multidimensionnels. La réduction permettra de palier cet effet. Elle consiste en la division de chaque valeur du jeu de données considéré par la valeur de l'écart-type.

Les transformations de données sont utilisées très souvent en analyse statistique. Elles permettent de réduire l'influence de certaines valeurs/caractéristiques du jeu de données comme les outliers et l'asymétrie positive fréquente dans le domaine environnemental.

2.2 Interprétation des données

2.2.1 Statistiques descriptives pour données censurées

Comme vu précédemment (section 1.4), les différents problèmes que pose le jeu de données doivent être étudiés avant de continuer le traitement statistique en utilisant des méthodes Gaussiennes.

Denis Helsel [8] propose plusieurs réponses à la question de la gestion des valeurs inférieures à la LQ. Le principe de base réside dans l'information que représente la proportion de valeurs inférieures à la LQ par rapport à celle des valeurs supérieures. Considérons deux jeux de données, contenant respectivement 75 et 10% de valeurs inférieures à une LQ unique et identique pour les deux populations. Le premier jeu de données contient de manière évidente plus de valeurs faibles que le

deuxième. En utilisant les valeurs au-dessus de la LQ et la proportion de données sous cette LQ, il est possible d'étudier la véritable distribution des données.

Helsel propose quatre approches ayant un intérêt pour notre protocole d'analyse :

- Les méthodes non-paramétriques après avoir re-censuré au niveau de la valeur inférieure à la LQ la plus élevée ;
- La MLE (*Maximum Likelihood Estimation*), méthode d'analyse de survie se basant sur une distribution supposée ;
- Autres méthodes d'analyse de survie non paramétriques ;
- La méthode ROS (*Regression on Order Statistics*).

La première approche ne possède pas la même puissance que les deux autres mais présente une bonne alternative à la substitution lorsque l'objectif est la simplicité/rapidité. Elle se décompose elle-même en deux groupes :

- La **méthode binaire** consiste à recoder les valeurs en deux catégories : « supérieure à la LQ » ou « inférieure à la LQ » si la LQ est unique. Ceci n'étant pas souvent le cas il est parfois nécessaire d'imposer une censure aux valeurs détectées en se basant sur la LQ la plus élevée (voir exemple ci-dessous).

Valeurs :	<1	<1	3	<5	7	8	8	8	12	15	22
Codage :	INF	INF	INF	INF	SUP	SUP	SUP	SUP	SUP	SUP	SUP

Ici, une moyenne, une médiane, un écart-type ne peuvent être calculés. Cependant, il est possible de produire des statistiques descriptives, des tests d'hypothèse et de construire des modèles de régression en utilisant une variable à réponse binaire. Des méthodes connues pour traiter ce genre de données sont les tests de proportions (plus connus sous le nom de tables de contingence).

- Les **méthodes ordinales** permettent au contraire de la méthode binaire d'utiliser une plus grande partie de l'information contenue dans les valeurs détectées. Elles utilisent le rang de chaque valeur sans utiliser la valeur numérique véritable. De nouveau, toutes les valeurs inférieures à la LQ la plus élevée seront censurées même si elles étaient détectées. Considérons le même jeu de données que précédemment :

Valeurs :	<1	<1	3	<5	7	8	8	8	12	15	22
Rangs :	2.5	2.5	2.5	2.5	5	7	7	7	9	10	11

Le rang des valeurs répétées est égal à la médiane des rangs si elles étaient différentes. Aux trois 8, dont les rangs auraient dû être 6, 7 et 8, est assigné le rang de 7, la médiane des trois rangs. La somme des rangs est ainsi préservée, règle statistique utilisée dans un grand nombre de tests. De même, les quatre plus faibles valeurs sont re-censurées comme inférieures à la LQ de 5, puis le rang

de 2.5 leur est attribué en tant que médiane des rangs 1-4 (qui aurait été attribué à des valeurs non censurées). Nous savons que « 3 » est une valeur détectée mais ne sachant pas si la valeur « <5 » est égale, par exemple, à 4 ou à 2.5 nous ne pouvons pas considérer le « 3 » comme une valeur détectée.

Tout en restant simple d'utilisation, ces méthodes permettent d'utiliser les données sans supposer plus qu'il n'est possible de le faire avec l'information disponible. En comparaison, utiliser la substitution revient à fabriquer des valeurs infondées qui pourraient mener à un faux positif lors d'un futur test statistique.

La MLE robuste[18], de plus en plus utilisée en études environnementales (Owen and DeRouen, 1980 ; Miesch, 1967 cité dans [8]), repose sur l'utilisation de trois informations :

- Les valeurs détectées au-dessus de la (des) valeur(s) inférieure(s) à la LQ ;
- La proportion des valeurs en dessous de chaque LQ ;
- La formule mathématique de la distribution supposée, log-normale dans notre cas.

Des statistiques descriptives sont calculées, selon les valeurs supérieures et inférieures à la LQ, tout en correspondant le plus possible à la distribution sélectionnée. Faire intervenir une distribution dans la méthode implique cependant de savoir si les données suivent ou non la distribution supposée. Ce qui peut être un problème pour des jeux de données à faible effectif, la crédibilité des paramètres estimés étant ainsi remise en cause. De plus, la MLE a été démontrée peu efficace pour des populations dont l'effectif est inférieur à 25-50 individus (Gleit, 1985 ; Shumway *et al.*, 2002 cité dans [8]). En revanche pour des effectifs supérieurs à 50, la MLE est toute indiquée.

Les procédures d'analyse de survie, la troisième approche, font partie de la famille des tests non-paramétriques mentionnés plus haut (section 1.4.3). Ces tests sont nommés à juste titre puisqu'ils n'imposent pas l'utilisation de paramètres comme la moyenne ou l'écart-type provenant d'une distribution supposée. A la place, ils utilisent les positions relatives (rangs) des valeurs qui sont un équivalent des centiles. Ces méthodes se révèlent très utiles pour les jeux de données censurés puisqu'ils utilisent uniquement l'information disponible sans utiliser une hypothèse pouvant être fausse, ainsi tout résultat calculé grâce à une méthode non-paramétrique est plus crédible qu'un résultat calculé en substituant par de fausses valeurs.

Les procédures d'analyse de survie non-paramétriques proviennent des études de survie d'une population de plusieurs individus lors de tests biologiques. Les individus sont placés sous différentes conditions de vie et leur durabilité dans le temps est reportée à chaque étape du test. Cependant, le test ne peut durer jusqu'à ce que tous les individus arrivent en fin de vie, certains d'entre eux présentant une durée de vie supérieure à une valeur donnée lorsque le test s'arrête. Le jeu de données contient donc plusieurs résultats censurés vers les valeurs élevées. L'analyse de survie peut

[18] Il existe une version totalement paramétrique de la MLE non présentée ici parce que présentant peu d'intérêt pour l'objectif visé

néanmoins, en utilisant les informations contenues dans les rangs et les proportions de données au-dessus et en-dessous des LQ multiples, calculer les paramètres statistiques basiques de la population étudiée : médiane, moyenne, écart-type, erreur standard, centiles. Deux méthodes permettent d'arriver à ce résultat :

- La méthode Kaplan-Meier (KM) est adaptée aux jeux de données contenant une ou des censure(s) vers les valeurs supérieures
- La méthode Turnbull est adaptée à la « double censure » c.à.d. vers les valeurs supérieures et les valeurs inférieures, elle n'est pas utile dans l'objectif de construire un fond pédo-géochimique anthropisé des sols urbains et donc ne sera pas développée ici.

En effet, étant donné que dans ce contexte, les substances analysées sont présentes « normalement » en faibles quantités, le type de censure rencontré concerne les valeurs faibles contrairement au domaine minier ou un pic de concentration (gisement) est recherché ; la méthode Turnbull peut être utilisée dans ce second contexte mais aussi pour la détermination de fond géochimique des eaux souterraines où le problème se rencontre parfois [15]. Des algorithmes d'analyse de survie sont disponibles dans les logiciels d'analyse statistique classique codés pour le modèle biologique, c.à.d. avec une censure vers les valeurs élevées. Ainsi pour pouvoir les utiliser il faudra effectuer préalablement une transformation des données. Il existe une solution alternative sous R© adaptée aux données censurées vers les valeurs faibles : la fonction `cenfit` de la librairie NADA présentée plus loin (section 0).

Une dernière approche, la méthode ROS[19] robuste[20] (Fiche 2), permet, comme les précédentes, de calculer les estimateurs basiques d'une population statistique. Les valeurs détectées sont utilisées pour imputer des valeurs à la portion censurée de la distribution. A l'aide d'un PP-plot[21], une régression est réalisée entre les quantiles des valeurs brutes et ceux d'une distribution théorique choisie, normale ou log-normale par exemple. Elle est applicable aux jeux de données à effectifs faibles (n<30), domaine ou les paramètres calculés par la MLE sont remis en cause.

Ces approches prennent leur importance ici uniquement à cause du faible effectif de la population et des taux de censure élevés. Il est évident que l'influence de 5 valeurs inférieures à la LQ est minime pour un jeu de données de 500 individus, par exemple. Il faut bien comprendre ici qu'il s'agit de méthodologies applicables à des étapes différentes du traitement statistique : (1) le calcul de paramètres statistiques descriptifs mis en avant ici pour faciliter la compréhension (2) la détermination d'intervalles de confiances pour les valeurs calculées (3) la comparaison et la corrélation intergroupes (4) la réalisation de traitements multidimensionnels. Elles permettent

[19] *Regression on Order Statistics*

[20] *Deux versions de cette méthode existent, l'une paramétrique et l'autre robuste. Seule la deuxième méthode est développée ici, elle est la plus adaptée pour l'objectif visé. Pour plus d'informations, voir [8]*

[21] *Probability-Probability plot (= Diagramme probabilité-probabilité)*

d'éviter la substitution, très répandue parce que simple d'utilisation mais insérant un biais dans les calculs effectués.

2.2.2 Représentations graphiques

L'histogramme est sûrement l'outil le plus utilisé pour étudier la distribution d'une population statistique classique. Il permet de visualiser concrètement la répartition des données et donc d'identifier rapidement si la distribution est symétrique ou non. Cependant, un effectif des données réduit (<50) peut être une source d'erreurs quant au choix du nombre et de la taille des classes (Fiche 5). Certaines informations concernant le jeu de données peuvent ne pas apparaître, la représentation peut donc être source d'erreurs. Denis Helsel considère l'histogramme comme inapproprié à l'étude de données censurées principalement pour ces raisons qui se résume en un manque d'unicité de la représentation graphique [8].

En étant accompagné d'une série de graphiques supplémentaires, l'histogramme peut s'avérer utile même en cas d'effectif faible. Une combinaison idéale de méthodes graphiques pour étudier la distribution des données comprendrait l'histogramme, la densité, le dispersogramme unidimensionnel et le boxplot (Figure 8) (voir Fiche 4, Fiche 5, Fiche 6, Fiche 7 de l'Annexe A pour une description des méthodes si nécessaire) [12].

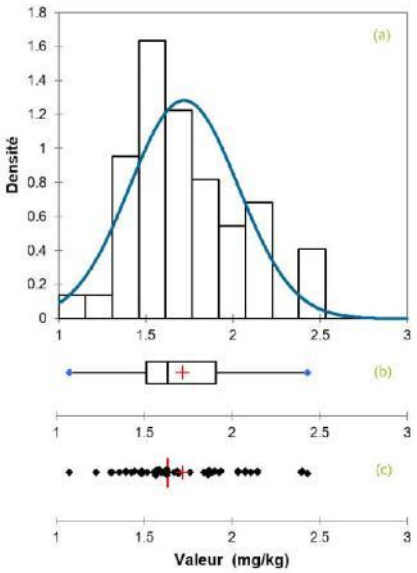


Figure 8 : Combinaison de graphiques descriptifs pour l'étude d'une distribution statistique – (a) Courbe de densité superposée à l'histogramme (b) Boxplot (c) Dispersogramme unidimensionnel

En plus de ces représentations, il est souvent utile de savoir quelle proportion de données se trouve au-dessus ou en-dessous d'une concentration x donnée. Il existe plusieurs graphiques permettant d'étudier cet aspect d'une distribution, le plus intéressant pour notre approche étant la fonction de répartition empirique ou ECDF[22] (Fiche 8) déjà présentée (Figure 3).

[22] *EmpiriCal Distribution Function*

2.2.3 Statistiques multidimensionnelles

Les études environnementales font souvent intervenir plusieurs substances analysées et donc plusieurs facteurs explicatifs. Les statistiques multidimensionnelles fournissent des méthodes donnant un aperçu des tendances et relations entre cet ensemble de données. Des tendances identiques détectées entre deux substances indiquent que leur apparition est gouvernée par les mêmes facteurs.

L'ACP fait partie intégrante de ces tests (Fiche 3). Son principe est basé sur la réduction des composantes générées pour les variables, une composante par variable. L'intérêt est de pouvoir visualiser le maximum d'informations concernant la variabilité des données sur un graphe en deux dimensions. La première dimension, notée « Dim 1 » (Figure 9) contient le taux maximal de variabilité. Le deuxième, « Dim 2 », est perpendiculaire au premier et contient le maximum de la variabilité restante. Les dimensions suivantes sont calculées de façon identique.

Sur le *graphe des individus* (Figure 9), tous les points sont projetés sur la première dimension. Il en résulte de nouveaux points avec de nouvelles coordonnées. Ces points sont à leur tour projetés sur la deuxième dimension, ce qui engendre encore de nouveaux points et ainsi de suite pour les projections suivantes. On peut choisir d'afficher les projections supérieures mais l'intérêt est limité puisque les deux premières contiennent, par construction, le maximum d'informations nécessaire à l'interprétation. En Figure 9 la première dimension (ou composante) explique 48,58% de la variabilité du jeu de données tandis que la deuxième dimension en exprime 17,31%.

L'ACP fournit également un graphe (Figure 10) permettant d'analyser les corrélations entre les variables impliquées, ici les différentes substances analysées. Son interprétation permet de détecter les associations de substances dont la variabilité est influencée par les mêmes facteurs. Il est également nécessaire à la compréhension du graphe des individus : plus l'abscisse d'un point est élevée sur la dimension 1, plus sa concentration en métaux sera élevée (puisqu'il s'agit ici des éléments traces métalliques). A l'inverse, plus son abscisse est faible, plus sa concentration est faible. Il en est de même pour l'interprétation de la dimension 2.

En intégrant une variable qualitative à l'étude on peut désormais afficher les groupes comme c'est le cas sur la Figure 9. Ainsi des tendances peuvent être détectées : par exemple, l'abscisse du centre du nuage de la population FGU étant supérieure celle de la population BDETM, on en déduit que la population FGU présente globalement des concentrations plus élevées sur les différentes substances analysées. De plus il est possible d'identifier quatre pôles regroupant (1) le cuivre, le chrome et le zinc, (2) le cadmium et le nickel, (3) le plomb, (4) le mercure.

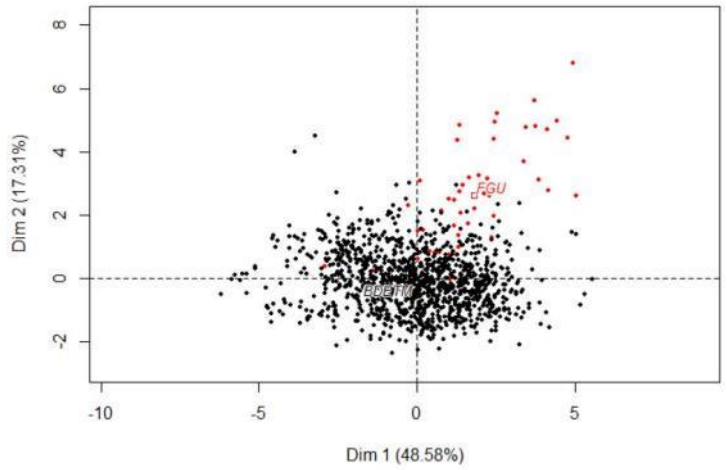


Figure 9 : ACP des données FGU (en rouge) et BDETM (en noir) - Graphe des individus

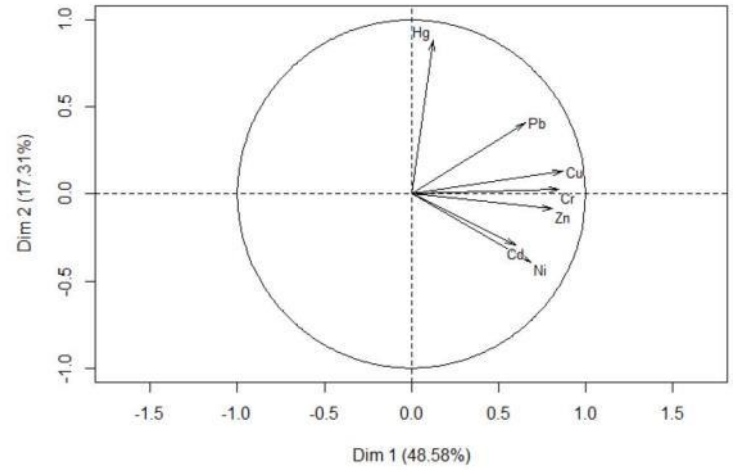


Figure 10 : ACP des données FGU (en rouge) et BDETM (en noir) - Graphe des variables

Comme toutes les méthodes statistiques, l'ACP est sensible aux jeux de données censurés. Helsel propose néanmoins des solutions ayant recours aux approches présentées plus haut (section 2.1.1). Il est possible de s'affranchir de ces problèmes en recodant le jeu de données de façon binaire ou en utilisant les rangs par exemples. En effet, le test de l'ACP est également conçu pour les variables qualitatives. Le résultat gagnera en crédibilité contrairement à la solution de la substitution qui insère un biais dans la distribution de la population considérée, ce qui peut entraîner l'apparition de fausses corrélations entre variables et facteurs.

3 Mise en application du protocole de traitement établi

3.1 Présentation de l'arbre de décision

Afin de mettre en évidence les difficultés rencontrées et la logique suivie au cours de l'établissement de ce protocole, prenons l'exemple du traitement des outliers : pour cette étape, la méthode de représentation par un boxplot est la plus indiquée [5]. Cependant, en raison de sa construction basée sur des statistiques Gaussiennes (Fiche 7), nous avons vu que le tracé d'un boxplot avec les valeurs brutes peut entraîner une surestimation du nombre de valeurs supérieures à la vibrisse supérieure. Il faudrait donc tracer le boxplot logarithmique afin d'obtenir une meilleure estimation sur une population moins asymétrique et enfin étudier les points dépassant la vibrisse supérieure sûrement moins nombreux qu'au premier tracé. Le tracé d'un boxplot bien que non limité par l'effectif de la population étudiée nécessite des valeurs numériques, les valeurs censurées sont donc un problème. Recourir à une substitution devient intéressant puisqu'elle fournit des valeurs numériques rapidement utilisables. Nous avons vu les conséquences statistiques que peut avoir cette solution (section 1.4.2). De plus en choisissant de tracer un boxplot logarithmique, l'hypothèse sous-jacente utilisée est « la distribution de la population étudiée suit une loi logarithmique ». De même que pour la substitution, nous savons maintenant qu'un mauvais choix de distribution hypothétique peut créer un biais lors de la suite des calculs, surtout lorsque les effectifs sont faibles (section 1.4.3).

Le protocole suivant est rédigé de façon chronologique. Pour un traitement optimal des données aucune étape ne doit être écartée. La suppression d'une étape pourrait avoir des conséquences sur la suite du protocole et mener à des résultats non valables.

Toute analyse statistique doit débuter avec le calcul des paramètres basiques de la population étudiée : nombre d'observations, pourcentage d'observations censurées, minimum, moyenne, médiane, 1er et 3ème quartile, maximum. Le calcul s'effectue ici avec les données brutes en conservant les LQ (Tableau 7).

Tableau 7 : Statistiques descriptives des analyses de cadmium (agglomération A)

Effectif	Pourcentage de censure	Minimum	1er Quartile	Médiane	Moyenne	3ème Quartile	Maximum
30	30%	0,1	0,19	0,47	0,46	0,65	1,37

Unité de mesure en mg/kg

Les valeurs calculées ne sont pas complètement interprétables à ce stade, mais permettent néanmoins d'obtenir une vision grossière des caractéristiques de la population. De plus, les erreurs de stockage ou de manipulation (copié-collé) des données, très fréquentes, sont facilement repérables.

Etape 1

Base de données BDSolU

↓

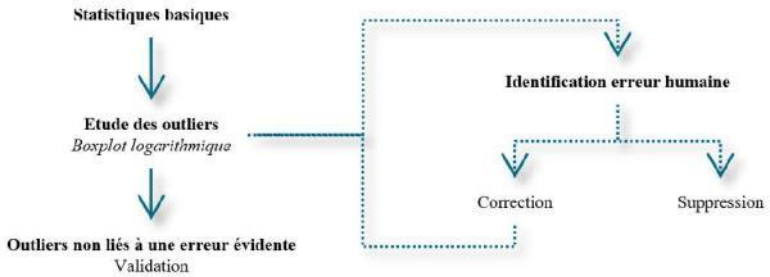
Statistiques basiques

Normalement, l'étape précédente permet de s'affranchir des erreurs humaines les plus flagrantes. Néanmoins une étude des outliers simple et rapide permet de s'assurer que l'on possède un jeu de données non biaisé. Une étape de traitement plus fine est prévue dans la suite du protocole afin de détecter les outliers restants.

Un boxplot logarithmique doit être tracé avec l'ensemble des analyses de chaque substance. Les valeurs inférieures à la LQ peuvent être substituées par la valeur de la LQ dans un premier temps. Chaque valeur au-dessus de la vibrisse supérieure doit être étudiée minutieusement afin de valider son statut : (a) erreur humaine à corriger/supprimer (b) valeur extrême appartenant à la distribution étudiée et devant être conservée (c) outlier à supprimer : valeur se distinguant des autres valeurs extrêmes pouvant être expliquée par exemple par un spot de pollution identifiable et proche du point de prélèvement. **Une valeur, même incompréhensible, ne doit en aucun cas être modifiée/supprimée sans justification précise.**

Le choix du boxplot logarithmique est motivé par le fait que les valeurs extrêmes (potentiellement outliers) détectées seront moins nombreuses qu'avec un boxplot classique [5]. Ainsi, seules les valeurs réellement déviantes sont étudiées. En effet, elles peuvent être nombreuses et ainsi impliquer une étude plus longue que nécessaire.

Etape 2



Après cette étape, le jeu de données est le moins biaisé possible. On peut procéder à une ACP afin de détecter les tendances de groupe. La variable qualitative à tester doit provenir d'une hypothèse de type : « la variabilité spatiale des substances analysées est observable à l'échelle des

agglomérations » ou « la gamme de concentration des éléments traces métalliques en zone urbaine est supérieure à celle des zones rurales ».

Le graphe des individus (Figure 10) permet déjà de confirmer ou d'infirmer l'hypothèse émise sur la variable qualitative sélectionnée grâce à la répartition du nuage de points. Si les deux dimensions conjecturées expriment (en sommant le pourcentage des deux dimensions) moins de 50% des données, les variables sélectionnées ne sont pas adéquates pour exprimer la variabilité des individus.

Etape 3

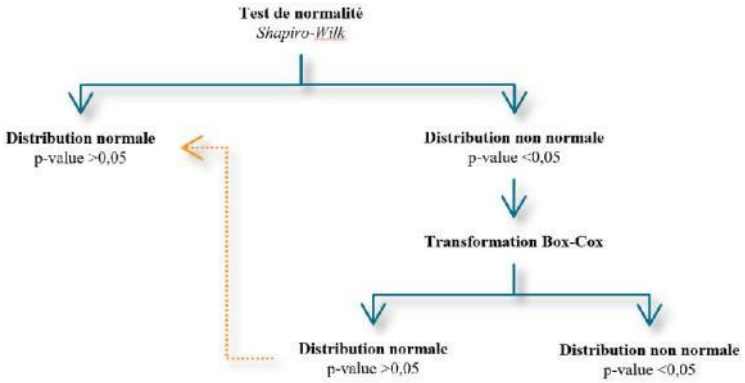
Données originales contenant uniquement les outliers validés

↓

Analyse en Composantes Principales
logarithmique

Avant de poursuivre l'étude statistique plus loin, il est nécessaire de réaliser un test de normalité sur le jeu de données étudié. La normalité ou non des données est déterminante sur le choix des tests à utiliser. Elle permet de différencier les populations qui nécessite ou non une transformation avant de continuer le traitement. Le test de Shapiro-Wilk (Fiche 1), considéré comme le plus fiable dans la littérature [11], est recommandé pour ce protocole. A partir de ce point, les données sont scindées en deux groupes en fonction de leur normalité. Le groupe présentant un résultat négatif au test est soumis à une transformation Box-Cox puis le test de normalité lui est de nouveau appliqué. Les données dont la p-value est supérieure à 0,05 sont reclassées dans la catégorie principale « Distribution normale ».

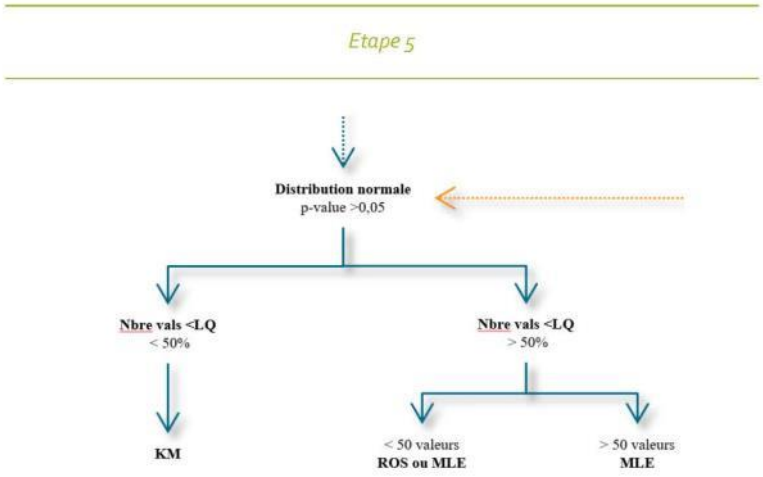
Etape 4



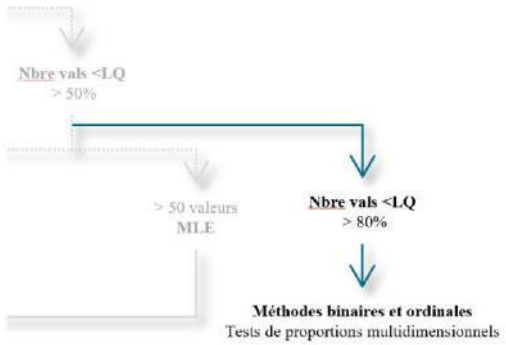
27

L'étape suivante fait intervenir les approches proposées par Helsel (section 2.2.1) pour la détermination de valeurs descriptives pour des données censurées. La méthode Kaplan-Meier, la méthode MLE et la méthode ROS présentent des avantages/inconvénients relatifs à la qualité des données statistiques ainsi que des domaines de validité. Ces caractéristiques permettent de proposer le protocole suivant [8]:

La méthode Kaplan Meier, non paramétrique, est privilégiée dans le cas des jeux de données censurés à hauteur de 50% et moins. Elle permet d'obtenir des estimations très fiables pour un pourcentage de censure relativement élevé. En comparaison, les méthodes ROS robuste et MLE robuste sont performantes à des taux de censure plus élevés grâce à l'emploi d'une distribution hypothétique.



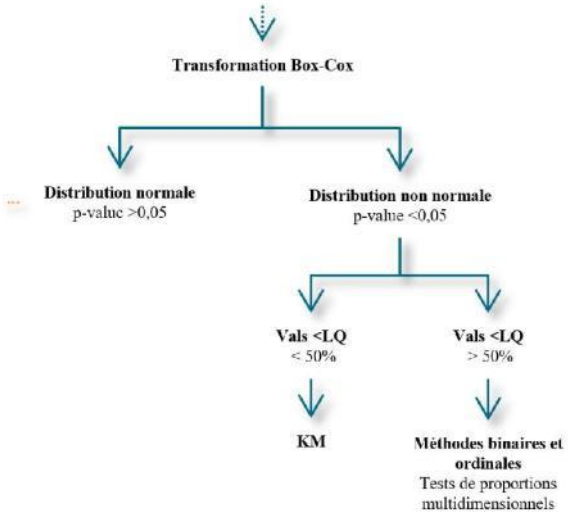
Une mention particulière est accordée aux jeux de données présentant plus de 80% de censure. Les méthodes proposées ne peuvent fournir des estimations valables dans ce cas, il est donc recommandé de se reporter aux méthodes binaires et ordinales qui permettront éventuellement un traitement par tests de proportions (ex : ACP qualitative).



28

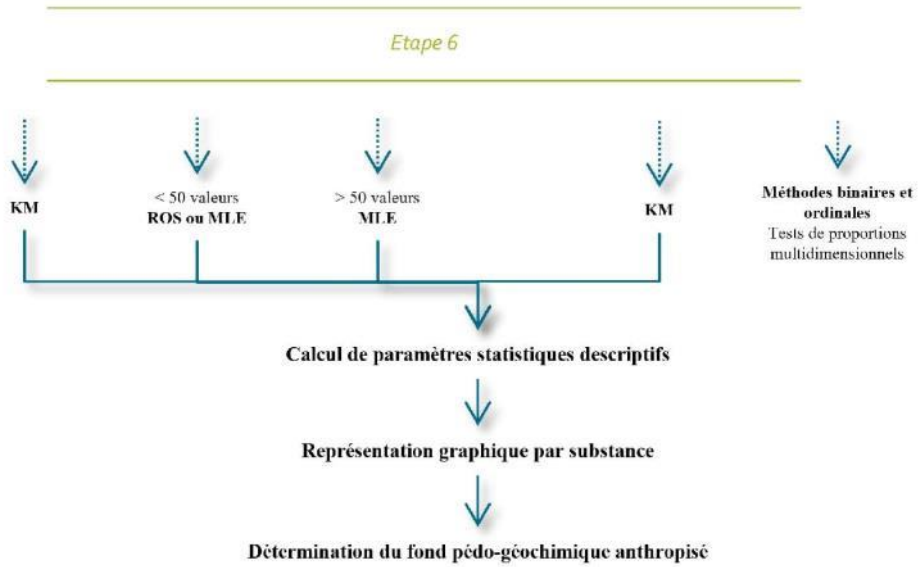
Au terme de l'étape 4, certains jeux de données sont jugés non normaux même après une transformation Box-Cox. Ceux d'entre eux présentant un pourcentage de censure inférieur à 50% peuvent être traités par la méthode Kaplan-Meier, puisqu'elle ne se base sur aucune distribution hypothétique. En revanche les autres jeux de données (pourcentage de censure supérieur à 50%) devront être traités par proportions comme dans l'étape précédente.

Etape 5 bis



Après la détermination des paramètres descriptifs de chaque population statistique, l'étape de traitement graphique peut être abordée. Pour ce faire, la combinaison histogramme, densité, boxplot et dispersogramme unidimensionnel est fortement recommandée (section 2.2.2). Un ensemble de graphique doit être tracé pour chaque substance analysée.

29



Ce n'est qu'après cet ensemble d'étapes que l'établissement d'un fond pédo-géochimique anthropisé peut être envisagé. Son calcul est réalisable par diverses méthodes largement commentées dans la littérature [5]. Ces méthodes, en cours de test, ne sont pas présentées dans cette version du protocole. Un schéma récapitulatif du protocole est disponible en Annexe B.

3.2 L'outil R dans le traitement des données

Il existe plusieurs logiciels permettant de tracer les différents graphiques présentés au cours de cette étude. Sur chacun d'eux plusieurs méthodes d'analyse graphique de la distribution d'une population statistique sont disponibles. Pour ce mémoire, XLStat© (https://www.xlstat.com/fr/) a été utilisé dans un premier temps, R© (https://www.r-project.org/) est rapidement indispensable à la vue des problématiques abordées. De façon générale, les méthodes permettant de traiter des jeux de données censurés, à faibles effectifs, à distributions biaisées sont beaucoup plus rares que les méthodes classiques.

Denis Helsel propose en collaboration avec Lee Lopaka la librairie R suivante [16] : NADA[23] qui propose, comme son nom l'indique, des solutions de traitements statistiques pour données contenant des valeurs inférieures à la LQ. Cette libraire ne répond pas à toutes les questions mais fournit des améliorations des outils graphiques cité en section 2.2.2. Par exemple, la fonction `cenboxplot` trace un boxplot dont la LQ la plus élevée est représentée par une ligne horizontale.

[23] *Nondetects And Data Analysis for environmental data*

Contrairement au boxplot classique, cette représentation permet la visualisation immédiate du niveau de censure maximal du jeu de données. Tout calcul sur des données inférieures à cette limite doit être réalisé avec des méthodes adaptées aux données censurées. La fonction `cenboxplot` présente néanmoins un inconvénient pour le traçage simultané de boxplots ayant des LQ différentes, une seule valeur n'est prise en compte pour l'affichage de la LQ la plus élevée. Cette question a été résolue par la fonction `cenboxplot2` développée par l'USEPA[24]. Cet outil permet de visualiser rapidement le degré de censure de chaque population étudiée et ce simultanément (Figure 11)

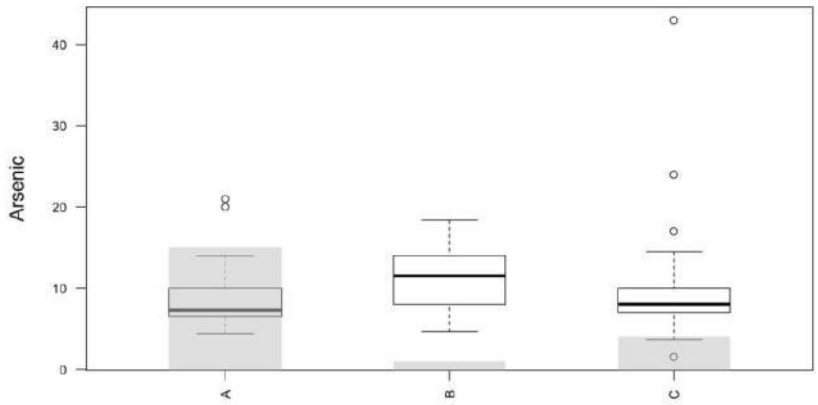


Figure 11: Boxplots des données arsenic (mg/kg) pour les agglomérations A, B et C ; la surface grisée représente la LQ la plus élevée de la population considérée

[24] *United States Environmental Protection Agency*

31

4 Conclusion

L'étude bibliographique des méthodes de traitement statistique de données environnementales pour la détermination d'un fond géochimique urbain a dominé cette étude. Elle a permis de sélectionner plusieurs méthodes de traitement adaptées aux données censurées, asymétriques et à effectif réduit. Le protocole proposé tient compte des conditions d'applications de ces méthodes ainsi que de leur efficacité par rapport aux caractéristiques des données de l'utilisateur. A ce jour, il permet de produire des statistiques descriptives et graphiques en ayant vérifié la normalité des populations, paramètre déterminant dans l'utilisation des solutions proposées.

L'application du protocole est proposée avec le logiciel R, ce qui a impliqué une autoformation au langage au cours du stage. A terme, un guide d'utilisation sera rédigé à partir des fonctions testées afin de faciliter la prise en main du protocole pour des utilisateurs non-initiés.

Plusieurs tests doivent encore être effectués afin de valider la suite du protocole portant sur le calcul de valeurs de fond géochimique. D'autres possibilités concernant l'analyse des résultats ont également été abordées : la comparaison des données FGU avec les autres bases de données disponibles (notamment BDETM), la confirmation de l'existence d'un fond géochimique par agglomération et la mise en évidence de la différence de fond géochimique entre le domaine urbain et le domaine rural.

Il faut être conscient que le protocole établi présente un intérêt à cause du faible effectif des populations étudiées. On peut espérer qu'à l'avenir la base de données BDSolU sera complétée par d'autres projets et donc contiendra assez d'échantillons pour se ramener au traitement statistique classique.

Annexe A

Fiches descriptives des méthodes statistiques et représentations graphiques présentées

33

Fiche 1

Test de Shapiro Wilk	
Objectif	Détermination du degré de normalité de la distribution d'une population statistique
Principe	Il compare un jeu de données de distribution inconnue avec une distribution de référence de même variance et même écart-type (dans ce cas la distribution normale). Il est dit « test d'hypothèse » : il requiert la formulation d'une hypothèse dite « nulle » dont la validité sera évaluée au cours du test. Par exemple, l'hypothèse nulle peut être : « La distribution hypothétique des données arsenic de l'agglomération A est une distribution normale ». En parallèle, une hypothèse alternative est également formulée : « La distribution hypothétique est différente de celle supposée ». Bien évidemment les « vraies » données ne suivront jamais exactement la distribution considérée, et en pratique une certaine déviation est tolérée. Si cette déviation est plus élevée que la limite définie par l'utilisateur pour certifier un minimum de significativité du test, l'hypothèse nulle ne peut être acceptée. Il en résulte l'acceptation de l'hypothèse alternative. Le résultat d'un test d'hypothèse est la p-value qui indique l'acceptation ou non de l'hypothèse nulle. Si cette p-value est inférieure à α, seuil de significativité prédéfini, l'hypothèse nulle est rejetée. Usuellement, $\alpha = 5\%$, ce qui implique que la p-value doit être supérieure à 0,05 pour que l'hypothèse nulle soit acceptée. Un seuil de significativité de 5% suppose que le résultat du test est fiable à 95%.
Conditions d'application	- Il n'est pas conseillé de diminuer α parce que la probabilité d'accepter la mauvaise hypothèse augmente
Bibliographie	[5]

34

Fiche 2

Regression on Order Statistics (ROS) – Version robuste	
Objectif	Calcul de statistiques descriptives d'une population censurée
Principe	A l'aide d'un PP-plot (Fiche 9), une régression par la méthode des moindres carrés est réalisée entre les centiles des données brutes (ou transformées) et les centiles d'une distribution hypothétique normale. Les paramètres de régression (pente et ordonnée à l'origine) sont calculés grâce aux observations non censurées. Par définition, si les points représentés sont proches de la droite régressée cela signifie que la population suit une loi normale. Des valeurs sont imputées à la partie censurée de la distribution d'origine en utilisant le modèle théorique. Les paramètres descriptifs de la distribution peuvent être calculés comme si elle n'avait pas été censurée. Si les données utilisées ont subi une transformation logarithmique, une transformation inverse doit être appliquée avant le calcul des paramètres statistiques de la distribution étudiée dans les unités d'origine. La déviation standard est égale à la pente de la droite de régression.
Conditions d'application	- Jeux de données présentant une censure inférieure - Recommandée pour un effectif n<50 et un taux de censure de 50-80%
Avantages	- Utilisable avec un jeu de données réduit jusqu'à n < 30
Inconvénients	- Intervention d'une distribution théorique - Moins efficace que la MLE robuste
Bibliographie	[8]

35

Fiche 3

Analyse en Composantes Principales (ACP)	
Objectif	Etude graphique de la structure d'un jeu de données multidimensionnel
Principe	L'ACP permet de traiter simultanément un nombre quelconque de variables quantitatives et qualitatives. Chaque variable est associée à une dimension et l'intérêt est de transcrire la variabilité du jeu de données multidimensionnel en un nuage de points bidimensionnel.
Conditions d'application	- L'unité de mesure des données doit être identique - Les données doivent être transformées avant de procéder à une ACP (transformation logarithmique puis centrage et réduction)
Avantages	- Traitement simultané de plusieurs jeux de données - La normalité des données étudiées n'est pas essentielle
Inconvénients	- Ne traduit qu'un certain pourcentage de la variabilité totale des données - La substitution des valeurs inférieures à la LQ induit un biais non négligeable sur l'interprétation des résultats - Le résultat de l'ACP est énormément influencé par le choix des variables incluses/exclues et par la présence de valeurs extrêmes/outliers - Un jeu de données non homogène conduit à des résultats instables. - Fiabilité diminuée pour des pourcentages de censure supérieurs à 30%
Bibliographie	[5], [8]

Fiche 4

Histogramme	
Objectif	Etude de la distribution d'un jeu de données
Principe	Les données sont représentées sous la forme de barres contiguës et de même largeur selon l'axe des abscisses. Chaque barre représente une classe contenant les données. La hauteur de la barre correspond à la fréquence d'apparition des données dans la classe. Dans le cas d'une distribution très asymétrique ou biaisée à cause de l'apparition de valeurs extrêmes ou outliers, le graphe peut être réalisé à partir des données log-transformées.
Conditions d'application	Aucune condition particulière
Exemple	Figure 8a
Avantages	- Visualisation rapide de la répartition des données
Inconvénients	- Nombre et taille des classes variables - Devient difficilement interprétable quand l'effectif diminue - Distribution asymétrique positive ou présence d'outliers dans la population réduit l'interprétabilité
Bibliographie	[5]

37

Fiche 5

Densité	
Objectif	Etude de la distribution d'un jeu de données
Principe	La densité est de manière simplifiée « l'histogramme représenté par une courbe lissée ». Elle représente une approximation de la distribution inhérente des données. Chaque point est calculé selon une certaine bande passante en utilisant une fonction de pondération. Dans le cas d'une distribution très asymétrique ou biaisée à cause de l'apparition de valeurs extrêmes ou outliers, le graphe peut être réalisé à partir des données log-transformées.
Conditions d'application	Aucune condition particulière
Exemple	Figure 8a
Avantages	- Plusieurs densités peuvent être superposées sur le même graphique afin de comparer la distribution de jeu de données différents (impossible avec des histogrammes) - Manipulable facilement
Inconvénients	- Le tracé est fortement conditionné par la bande passante sélectionnée (paramètre modifiable sous R©)
Bibliographie	[5]

38

Fiche 6

Dispersogramme unidimensionnel (ou *scatterplot*)	
Objectif	Etude de la distribution des données
Principe	Les valeurs sont tracées uniquement selon l'axe des abscisses. Une deuxième représentation est possible en disposant les points sur l'axe des ordonnées de manière aléatoire
Conditions d'application	L'effectif du jeu de données doit être assez faible pour que les points ne soient pas trop proches et ne se superposent pas
Exemple	Figure 8c
Avantages	Permet une visualisation simple et rapide de la structure des données
Inconvénients	La superposition des points peut induire l'utilisateur en erreur
Bibliographie	[5]

39

Fiche 7

Boxplot de Tukey (ou boîte à moustaches)	
Objectif	Etude de la distribution d'une population
Principe	La construction du boxplot se démontre facilement avec un petit jeu de données, par exemple : 2.3 2.7 1.7 1.9 2.1 2.8 1.8 2.4 5.9 Les données sont classées par ordre croissant afin de trouver la médiane[25] qui sera, ici « 2.3 » : 1.7 1.8 1.9 2.1 **2.3** 2.4 2.7 2.8 5.9 La médiane des deux portions restantes est également calculée : 1.7 1.8 **1.9** 2.1 **2.3** 2.4 **2.7** 2.8 5.9 1.9 et 2.7 correspondent respectivement au premier quartile ($Q1$) et au troisième quartile ($Q3$). Ils définissent la boîte centrale, qui contient approximativement 50% des données et permet d'apprécier la symétrie par rapport à la médiane. La longueur de la boîte est définie comme la différence entre les quartiles, approximativement la distance interquartile (DI) : $DI = Q_3 - Q_1 = 0{,}8$ Elle représente une estimation de la dispersion des données autour de la médiane. Des frontières permettent de définir la limite au-delà de laquelle les individus sont considérés comme valeurs extrêmes/outliers. Elles sont définies comme suit : $Frontière\ supérieure = Q3 + 1{,}5 \times DI = 3{,}9$ $Frontière\ inférieure = Q3 - 1{,}5 \times DI = 0{,}7$ Les frontières permettent de calculer les vibrisses (ou moustaches) de la boîte centrale : $Moustache\ supérieure = \max(x[x \leq Frontière\ supérieure]) = 2{,}8$ $Moustache\ inférieure = \min(x[x \geq Frontière\ inférieure]) = 1{,}7$ Les moustaches permettent d'apprécier la symétrie de la distribution. (Exemple Figure 8)
Conditions d'application	Aucune condition particulière
Avantages	- Représentation graphique la complète permettant de décrypter la distribution d'un ensemble de valeurs statistiques
Inconvénients	- Le calcul des vibrisses est basé sur la théorie de la loi normale et donc sur l'hypothèse de symétrie des données.
Bibliographie	[5]

[25] Il existe plusieurs méthodes de calcul de la médiane, ici on fait référence à la médiane de Tukey [5]. Un exemple simple a été choisi pour faciliter la compréhension.

Fiche 8

Fonction de répartition empirique – ECDF	
Objectif	Etude de la distribution d'un jeu de données
Principe	L'ECDF est une fonction de répartition discrète qui attribue la probabilité $1/n$ à chacune des observations (n : nombre d'observations). Plus n augmente, plus l'ECDF se rapproche d'une fonction de répartition continue. L'axe des abscisses représente les données tandis que l'axe des ordonnées représente la probabilité suivante : $$F_n(x) = \frac{nombre\ d'éléments\ dans\ l'échantillon \leq x}{n}$$ Dans le cas d'une distribution très asymétrique ou biaisée à cause de l'apparition de valeurs extrêmes ou outliers, le graphe peut être réalisé à partir des données log-transformées.
Conditions d'application	Aucune condition particulière
Exemple	Figure 3a
Avantages	- Version robuste de la fonction de répartition classique (moins influencée par les différents biais possibles) - Au contraire de la densité, chaque point de mesure est visible.
Inconvénients	- Résultat graphique très influencé par les valeurs extrêmes et outliers
Bibliographie	[5]

41

Fiche 9

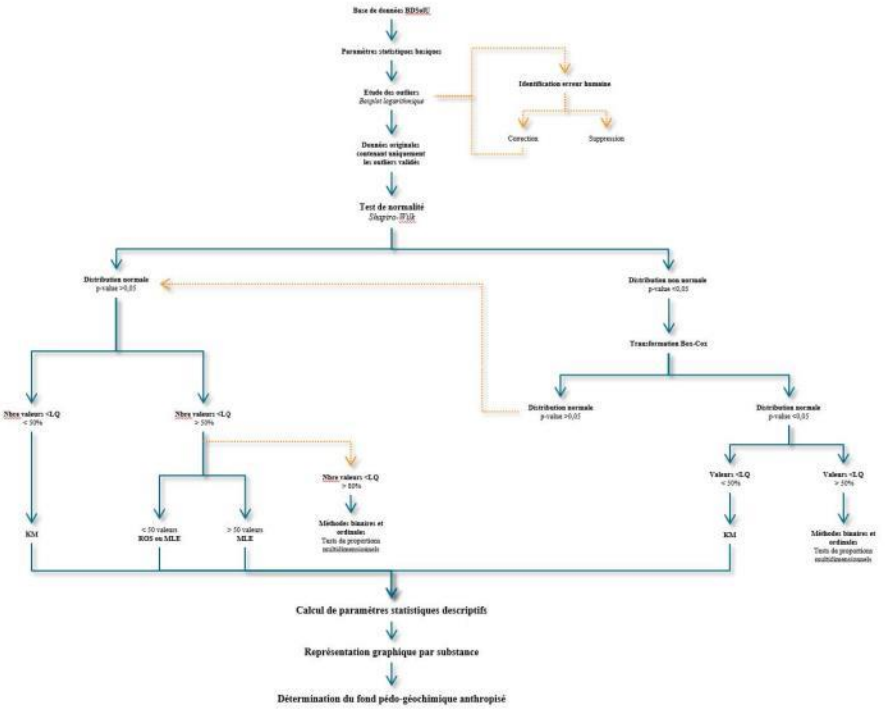
Diagramme Probabilité-Probabilité (*PP-plot*)	
Objectif	Vérification graphique de l'ajustement de la distribution d'une population à une distribution normale hypothétique
Principe	Les probabilités des données brutes sont représentées sur l'axe des abscisses tandis que les probabilités de la distribution hypothétique normale sont tracées sur l'axe des ordonnées.
Conditions d'application	Aucune condition particulière
Avantages	- Comparé aux autres diagrammes de comparaison de deux distributions (CP-plot, QQ-plot, ...) le PP-plot est moins influencé par les valeurs extrêmes/outliers ; en raison de leur faible probabilité.
Inconvénients	- Perte totale de l'échelle originale des données (à la différence de l'ECDF et du CP-plot)
Bibliographie	[5]

42

Annexe B

Schématisation du protocole de sélection proposé

43


Base de données BDSolU
Paramètres statistiques basiques
Etude des outliers
Boxplot logarithmique
Identification erreur humaine
Correction
Suppression
Données originales contenant uniquement les outliers validés
Test de normalité
Shapiro-Wilk
Distribution normale
p-value >0,05
Distribution non normale
p-value <0,05
Transformation Box-Cox
Distribution normale
p-value >0,05
Distribution normale
p-value <0,05
Nbre valeurs <LQ
< 50%
Nbre valeurs <LQ
> 50%
Nbre valeurs <LQ
> 80%
Valeurs <LQ
< 50%
Valeurs <LQ
> 50%
KM
≤ 50 valeurs
ROS ou MLE
> 50 valeurs
MLE
Méthodes binaires et ordinales
Tests de proportions multidimensionnels
KM
Méthodes binaires et ordinales
Tests de proportions multidimensionnels
Calcul de paramètres statistiques descriptifs
Représentation graphique par substance
Détermination du fond pédo-géochimique anthropisé

44

Bibliographie

[1] MEEM, «Méthodologie de gestion - Deux démarches bien distinctes,» 11 04 2011. [En ligne]. Available: http://www.developpement-durable.gouv.fr/Deux-demarches-bien-distinctes.html. [Accès le 19 08 2016].

[2] J.-F. Brunet, F. Guiet, C. Blanc, V. Laperche, P. Balon et N. Aubert, «Etablissement de fonds pédo-géolchimiques urbains et industriels en parallèle à l'Opération ETS du Ministère du Développement durable,» 2015.

[3] BRGM, «Le BRGM, service géologique national,» 06 07 2016. [En ligne]. Available: http://www.brgm.fr/brgm/le-brgm-service-geologique-national/brgm-service-geologique-national. [Accès le 08 08 2016].

[4] MEEM, «Etablissements sensibles - Diagnostiquer les lieux accueillant les enfants et les adolescents,» 15 04 2011. [En ligne]. Available: http://www.developpement-durable.gouv.fr/Diagnostiquer-les-lieux.html. [Accès le 04 2016].

[5] C. Reimann, P. Filzmoser and R. Dutter, Statistical Data Analysis Explained : Applied Environmental Statistics with R, John Wiley & Sons, Ltd, 2008, p. 359.

[6] E. L. Ander, C. C. Johnson, M. R. Cave, B. Palumbo-Roe, C. P. Nathanail and R. M. Lark, "Methodology for the determination of normal background concentrations of contaminants in English soil," *Science of The Total Environment*, vol. 454–455, pp. 604-618, 01 06 2013.

[7] ISO, Qualité du sol — Guide pour la détermination des valeurs de fond, ISO, 2016, p. 33.

[8] D. R. Helsel, Statistics for censored environmental data using Minitab and R, 2nd ed., Denver, Colorado: John Wiley & Sons, Inc., 2012, p. 343.

[9] R. Rakotomalala, «Tests de normalité : Techniques empiriques et tests statistiques,» Juin 2008. [En ligne]. Available: http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf. [Accès le 19 Mai 2016].

[10] N. M. Razali et Y. . B. Wah, «Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov,Lilliefors and Anderson-Darling tests,» *Journal of Statistical Modeling and Analytics*, vol. Vol.2, n° %1No.1, pp. 21-33, 2011.

[11] J. Andersson et M. Burberg, «Testing For Normality of Censored Data,» Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Social Sciences, Department of Statistics, Uppsala, 2015.

[12] C. Riemann, M. Birke et P. Filzmoser, «Data Analysis for Urban Geochemical Data,» chez *Mapping the Chemical Environment of Urban Areas*, C. Johnson, A. Demetriades, J. Locutura et R. T. Ottensen, Éds., John Wiley & Sons, Ltd., 2011, p. 616.

[13] D. Helsel and R. Hirsch, Statistical Methods in Water Resources Techniques of Water Resources Investigations, vol. 4, U.S. Geological Survey, 2002, p. 522.

[14] G. E. P. Box et D. R. Cox, «An Analysis of Transformations,» *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, n° %12, pp. 211-252, 1964.

[15] N. Devau, J. Lions et A. Mauffret, Interviewees, *Discussion sur les méthodes et bonnes pratiques pour la détermination d'un fond géochimique, cas des eaux souterraines.*. [Interview]. Mai 2016.

[16] L. Lee and D. Helsel, *NADA: Nondetects And Data Analysis for environmental data*, R Package, 2015.

[17] R Development Core Team, «R: A language and environment for statistical computing,» 2008. [En ligne]. Available: http://www.R-project.org. [Accès le 07 2016].

[18] A. Bolks, A. DeWire et J. Harcum, «Baseline assessment of left-censored environmental data using R. Tech Notes 10 Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA, 28p.,» 06 2014. [En ligne]. Available: www.bae.ncsu.edu/programs/extension/wqg/319monitoring/tech_notes.htm.. [Accès le 07 05 2016].

[19] BRGM, «Gestion des environnements pollués : une approche intégrée,» 11 02 2013. [En ligne]. Available: http://www.brgm.fr/activites/environnement-ecotechnologies/gestion-environnements-pollues-approche-integree. [Accès le 08 08 2016].

[20] MEDDE, «Bases de données relatives à la qualité des sols : contenu et utilisation dans le cadre de la gestion des sols pollués,» Avril 2008. [En ligne]. Available: www.developpement-durable.gouv.fr/spip.php?page=doc&id_article=19946. [Accès le 23 Juin 2015].

[21] Wikipédia, «Asymétrie (statistiques) en français,» 17 08 2016. [En ligne]. Available: http://fr.wikipedia.org/w/index.php?title=Asym%C3%A9trie_(statistiques)&oldid=128729438. [Accès le 22 08 2016].

46

5 Index des acronymes et définitions

ACP	Analyse en Composantes Principales	p.23
ADEME	Agence de l'Environnement et de la Maîtrise de l'Energie	p.5
BASIAS	Base de données des Anciens Sites Industriels et Activités de Service	p.4
BRGM	Bureau de Recherches Géologiques et Minières	p.2
Censure	Un jeu de données est dit censuré s'il contient des valeurs inférieures à la limite de quantification	p.8
ECDF	EmpiriCal Distribution Function	p.22
ETS	Etablissements Sensibles	p.4
FGU	Fond Géochimique Urbain (raccourci de Fond Pédo-Géochimique Urbain Anthropique)	p.4
INRA	Institut National de la Recherche Agronomique	p.2
LQ	Limite de quantification	p.8
MEEM	Ministère de l'Environnement, de l'Energie et de la Mer	p.1
MLE	Maximum Likelihood Estimation	p.24
NADA	Nondetects And Data Analysis for environmental data	p.30
Outliers	Valeurs appartenant à une population différente car elles sont originaires d'un autre processus/source, i.e elles proviennent d'une distribution contaminée	p.13
Plan d'échantillonnage thématique	Prélever des sols choisis pour leurs caractéristiques (par exemple : localisation, activité accueillie, nature...)	p.5
PNSE	Plans Nationaux Santé Environnement	p.4

47

p-value	Le résultat du test de normalité qui indique l'acceptation ou non de l'hypothèse nulle. Si cette p-value est inférieure à α, seuil de significativité prédéfini, l'hypothèse nulle est rejetée...	p.11
ROS	Regression on Order Statistics.....	p.24
Terres excavées	Sol excavé, qui peut comporter des remblais hétérogènes apportés au fil des ans.....	p.1
USEPA	United States Environmental Agency.....	p.31
Vibrisse	Voir Fiche 7.....	p.45

48

Centre scientifique et technique
Direction Eau, Environnement & Écotechnologies
3, avenue Claude-Guillemin
BP 36009 – 45060 Orléans Cedex 2 – France – Tél. : 02 38 64 34 34
www.brgm.fr