Document public

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

Rapport final

BRGM/RP-66501-FR
décembre 2016

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

Géosciences pour une Terre durable
brgm

Document public

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

Rapport final

BRGM/RP-66501-FR
décembre 2016

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

L. Sancho
Avec la collaboration de
J.F. Brunet, N. Saby, V. Derycke, R. Albinet, J. Lions, N. Deveau, A. Mauffret, N. Croiset, H. Berthier, B. Bourgine

Vérificateur :	**Approbateur :**
Nom : BRUNET Jean-François	Nom : LÉPROND Hubert
Fonction : Chef de Projet	Fonction : Responsable de l'unité Sites, Sols et Sédiments Pollués
Date : 04 avril 2017	Date : 19 juin 2017
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 géochimique, Arsenic, Cuivre, Cadmium, Pyrène.

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

Sancho L. (2016) - Protocole d'Analyse Statistique pour la construction d'un Fond Pédo-Géochimique Anthropisé des Sols Urbains. Rapport final. BRGM/RP-66501-FR, 111 p., 12 tab., 23 fig., 3 ann.

Remerciements

Une reconnaissance particulière est portée à la participation de Monsieur Nicolas Saby (Institut National de la Recherche Agronomique) intervenu dans ce rapport en tant que relecteur et fournisseur des données de la Base de Données des Éléments Traces Métalliques, lesquelles ont permis de réaliser les Analyses en Composantes Principales comparatives entre BDETM et BDSolU (voir section 0).

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. Au cours de l'année 2016, un stage de fin d'études a été réalisé dans l'objectif de fournir un protocole statistique pour la détermination de fonds pédo-géochimiques anthropisés de sols urbains à partir des analyses présentes dans la base de données BDSolU.

Suite aux conditions de prélèvement et d'analyse des sols, les données collectées présentent des caractéristiques qui affectent le traitement statistique. La taille du jeu de données, les valeurs inférieures aux limites de quantification et la présence d'outliers sont les principaux paramètres à prendre en compte lors de l'étude statistique de ces données géochimiques.

Après une étude bibliographique, plusieurs méthodes sont proposées pour le traitement de données présentant ces caractéristiques. Les méthodes Kaplan-Meier, Maximum Likelihood Estimation et Regression on Order Statistics permettent une description statistiques adaptée aux données à faible effectif, censurées et asymétriques.

De même que la détermination de statistiques descriptives, la représentation graphique des résultats doit être réalisée avec précaution lors du traitement de données environnementales. Pour avoir la vision la plus complète possible des caractéristiques des données, la combinaison d'un histogramme, d'une courbe de densité, d'une boîte à moustaches, d'un dispersogramme unidimensionnel et d'une fonction de répartition empirique est recommandée.

Après ces deux étapes, la détermination du fond pédo-géochimique anthropisé (FPGA) peut être réalisée selon les méthodes calculatoires suivantes : « *MEAN* ± *2SD* » et les vibrisses internes de la boîte à moustaches. Les méthodes graphiques, très utilisées dans la littérature, ne s'adaptent pas aux données disponibles principalement à cause des effectifs trop faibles.

Enfin, une carte géochimique doit être réalisée pour visualiser la relation entre la répartition spatiale des valeurs mesurées avec les seuils du FPGA déterminés.

La réalisation d'un programme informatique permettant l'automatisation des tests et calculs proposés faciliterait grandement le traitement des données et permettrait de le rendre utilisable à un usager non statisticien. De plus, une étude géostatistique poussée est à envisager afin d'explorer les possibilités qu'offre la simulation de variable aléatoire dans l'objectif de produire des jeux de données simulés à effectif augmenté. Cette solution permettrait l'utilisation du *Concentration Area plot*, outil statistique requérant des effectifs élevés, mais permettant de déterminer un fond géochimique en considérant la répartition spatiale des points de mesures.

Sommaire

Liste des tableaux

Liste des figures

Liste des Annexes

1. 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 n'est responsable qu'en partie de l'évolution géochimique des sols au cours des derniers siècles. La composition des sols actuelle est le résultat d'une combinaison de facteurs : (1) l'altération des roches (2), leur évolution autonome en sols sous l'action de facteurs climatiques et biologiques, (3) l'existence de retombées atmosphériques diffuses d'origine anthropique et (4) de sources de pollution ponctuelles. Ces diverses situations correspondent à plusieurs types de fonds géochimiques récapitulés par la Figure 1.

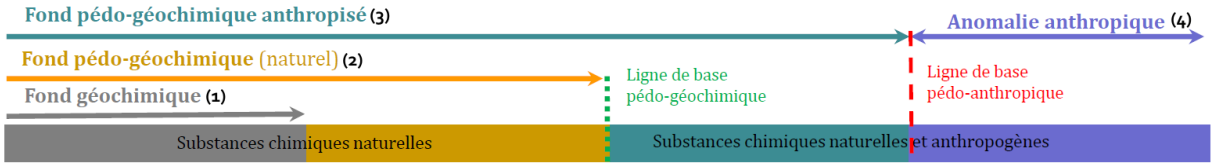


Figure 1 : Terminologie employée dans la détermination de valeurs de fond de la qualité des sols d'après le Groupe de Travail [1] « Valeurs de Fond » 2016.

Cette figure distingue 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. Selon la méthodologie nationale, il convient alors de recueillir des données sur ce fond pédo-géochimique anthropisé (FPGA) pour mieux gérer les sites et sols (potentiellement) pollués.

La démarche française de gestion des sites (potentiellement) pollués s'appuie surtout sur la note ministérielle du 08 février 2007 éditée par le MEEM[2] et révisée en 2017 [1]. La méthodologie mise en place pour les pollutions passées est fondée sur la prévention et sur la gestion des risques suivant l'usage. 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.

[1] L'ADEME a mis en place un Groupe de Travail « Valeurs de Fond » animé par Yncréa (www.yncrea.fr) pour la rédaction d'un « Guide de bonnes pratiques pour la détermination de fonds pédo-géochimiques anthropisés pour la gestion d'un site pollué en milieu urbain, rural ou industriel ».

[2] Ministère de l'Environnement, de l'Énergie et de la Mer.

Cette politique est cohérente avec la démarche de gestion des terres excavées 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]. En effet, une terre est considérée exempte de pollution dès lors que ses caractéristiques sont cohérentes avec le fond géochimique naturel local [3]. On considère qu'une denrée alimentaire peut être consommée sans risque par 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 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.

Dans ce contexte la détermination du fond pédo-géochimique anthropisé donne lieu à des questions méthodologiques statistiques. En effet, les données collectées présentent des caractéristiques particulières découlant des prélèvements et analyses de sols : la taille des jeux de données, la présence de valeurs inférieures aux limites de quantification analytiques et la présence d'outliers. Un stage de fin d'études ingénieur, prévu dans le projet FGU, s'est déroulé entre avril et septembre 2016 pour tenter de répondre à ces questions statistiques. Le résultat de ce travail est disponible en annexe du rapport intermédiaire BRGM/RP-66306-FR.

Le présent rapport, réalisé entre octobre et décembre 2016, complète le travail effectué. Il contient l'analyse critique des traitements statistiques actuels de données environnementales et propose des méthodes adaptées aux caractéristiques des données contenues dans BDSolU. La préparation des données, leur interprétation et la détermination de valeurs seuil sont abordées. L'application pratique de ces méthodes est traitée dans le protocole statistique proposé en s'appuyant sur des fiches méthodologiques décrivant les méthodes utilisées (Annexe 1).

2. Contexte de travail

2.1 CONTEXTE TECHNIQUE

2.1.1 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 peut être confortée par la consultation des bases de données suivantes [3] :

- 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 (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[3]) et consultables sur le site du Groupement d'Intérêt Scientifique Sol (GisSol[4]) ;
- 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[5].

Néanmoins, leur utilisation dans le cadre d'un diagnostic de sols (potentiellement) pollué 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. Les prélèvements ne couvrent pas la France entière, sont réalisés selon des protocoles variés et majoritairement localisés en milieu rural. Ces caractéristiques soulèvent des difficultés potentielles au moment d'établir des comparaisons avec les données spécifiques au milieu urbain.

2.1.2 Problèmes soulevés par le milieu urbain

La détermination d'un fond pédo-géochimique en milieu urbain implique la prise en compte de plusieurs spécificités :

- le domaine urbain est très peu couvert par les études relatant la qualité des sols ;
- au sein d'une agglomération, il est difficile de distinguer la contamination d'un site de celle imputable aux sites voisins et à l'activité de toute la ville, tâche pourtant indispensable à l'objectif recherché ;
- de plus, les remblais, très présents au sein des sites urbains, n'ont généralement 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 ;

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

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

[5] InfoTerre – infoterre.brgm.fr

- 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 : www.societedugrandparis.fr/).

2.2 LE PROJET FOND PÉDOGÉOCHIMQUE URBAIN (FGU)

Le projet FGU conduit par le BRGM en convention avec l'ADEME est adossé au projet « Diagnostic des sols dans les lieux accueillant des enfants et des adolescents » également intitulé « Établissements Sensibles », géré par le BRGM pour le Ministère en charge de l'environnement.

2.2.1 Le projet Établissements Sensibles

La mise en œuvre des « Diagnostics des sols dans les lieux accueillant des enfants ou des adolescents » (nom abrégé « Établissements Sensibles » ou « ETS[6] ») par le ministère en charge de l'environnement, correspond à une opération préventive qui apparait au programme des PNSE[7] 2 et 3 (PNSE 2. 2009-2013 et PNSE 3. 2014-2016 – Actions 19 et 61, voir [4]). Au cours cette opération, plus de 2 000 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 sur, ou à proximité immédiate, d'anciens sites industriels ou d'activités de service recensés dans l'inventaire BASIAS[8] (site internet : 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 idéalement lieu à un prélèvement dans deux ou trois espaces verts situés à proximité (moins d'1 km) du lieu d'échantillonnage. Les échantillons de sols prélevés au sein des établissements scolaires sont nommés SLE pour « SoL des Établissements » et ceux prélevés dans les 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.

[6] Établissements Sensibles.

[7] Plans Nationaux Santé Environnement.

[8] Base de données des Anciens Sites Industriels et Activités de Service.

2.2.2 Organisation du projet FGU

L'ADEME[9] 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 « Établissement de fonds pédo-géochimiques urbains et industriels également appelé « Fonds Géochimiques Urbains » ou 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 « typologique [10]» 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.

Les échantillons retenus pour le projet FGU sont ceux 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 seront bancarisés à terme grâce à l'outil BDSolU[11] 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.

[9] Agence de L'Environnement et de la Maîtrise de L'Énergie.

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

[11] Base de Données des analyses de Sols Urbains www.bdsolu.fr

2.2.3 Présentation des résultats de la première convention

En Mai 2016, les prélèvements du projet ETS dans 278 villes de France métropolitaine avaient permis la bancarisation de 635 échantillons SLU. À ce stade trois agglomérations présentaient 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 questions méthodologiques apparues à chaque étape de la démarche et portant sur :

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

Concernant ce dernier point, les hypothèses ci-dessus restent en effet à valider au moyen de tests statistiques fiables, robustes et adaptés au contexte de la démarche.

2.3 ANALYSE CRITIQUE DES TRAITEMENTS STATISTIQUES ACTUELS

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, totalement conditionnée par celle du projet ETS, entraîne certaines conséquences sur la qualité des données bancarisées.

2.3.1 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 typologique (voir 2.2.2), 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 terme, au maximum 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 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.

En mai 2016, seule l'agglomération C (les agglomérations A,B et C ont été introduites dans le rapport RP-64845-FR (2015) [3]) 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, typologique, la répartition spatiale des points de prélèvements est hétérogène. Donc certains secteurs posséderont *in fine* une concentration de points plus élevée que d'autres (exemple Figure 2).

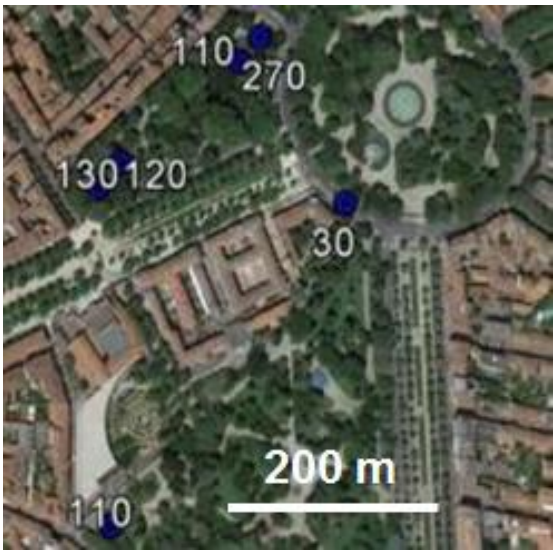


Figure 2 : Exemple de répartition spatiale des points de prélèvements.

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.

Un fond géochimique se présentant habituellement sous forme d'une valeur ou d'un ensemble de valeurs numériques, peut-on calculer des paramètres statistiques prenant en compte l'hétérogénéité spatiale des points ET la forte variabilité des teneurs pour la zone d'étude considérée ?

La délimitation de cette zone serait également à définir. Jusqu'à maintenant, les agglomérations sont utilisées en tant que délimitation géographique [3]. Cette décision peut-elle être confirmée/infirmée par des tests statistiques ?

Le fond pédo-géochimique anthropisé est-il statistiquement différent d'une agglomération à une autre ?

2.3.2 Limite de quantification

Chaque méthode analytique comporte une limite de quantification (LQ) inférieure et une LQ supérieure (aussi appelée limite de saturation). Le fichier reçu du laboratoire contiendra parfois un pourcentage élevé de données inférieures à la LQ. Dans ce cas, l'utilisation de l'ensemble des valeurs brutes introduit un biais lors du traitement statistique, et les résultats deviennent ininterprétables La présence de valeurs inférieures au seuil de quantification peut donc biaiser l'estimation des paramètres des modèles statistiques couramment utilisés (ex : moyenne arithmétique).

La limite de quantification inférieure 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 quantification. On utilisera les termes « **censure inférieure** » pour qualifier un jeu de données censuré uniquement pour des valeurs faibles*[12]*.*

Plus la population de données analytiques présente un taux élevé de résultats inférieurs à la LQ plus sa description statistique se trouve potentiellement biaisée.

Une même méthode d'analyse 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 de l'analyste. 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. L'usage d'une même technique d'analyse par différents laboratoires peut engendrer des LQ différentes.

Quand on recherche des valeurs de fond, la censure inférieure fait plus souvent l'objet d'attention que la censure supérieure [5] puisque les substances analysées dans les sols sont normalement présentes en faible concentration, voir à des concentrations inquantifiables. Le jeu de données résultant est donc censuré inférieurement (exemple Tableau 1).

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

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

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 de substances. Les pourcentages varient plus ou moins fortement en fonction des substances/familles analysées.

[12] La limite de quantification supérieure ou limite de saturation 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 ».

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 %

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

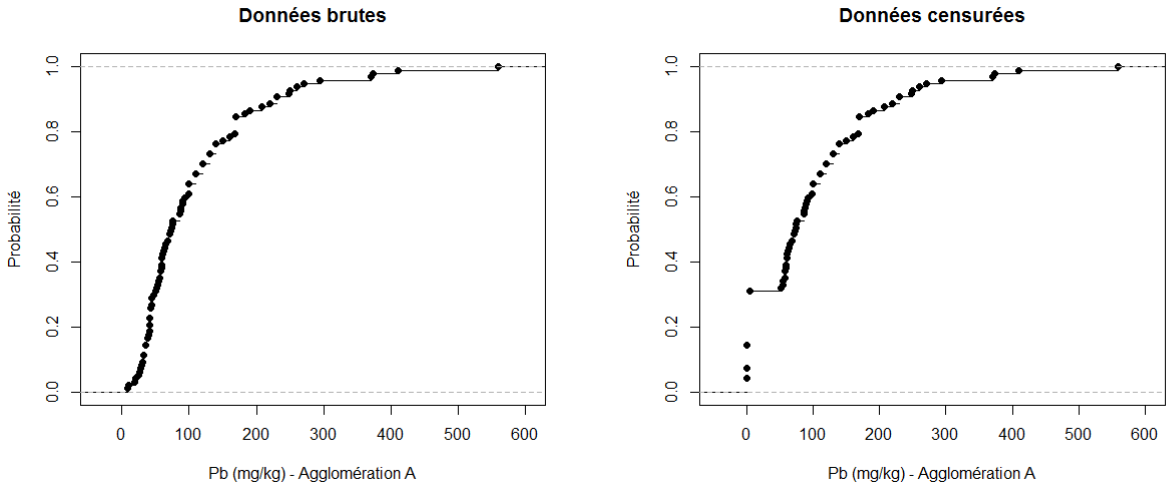
À la vue de ces résultats, il devient évident que les données (celles relatives à l'arsenic et à l'acénaphtylène, par exemple) ne peuvent être traitées statistiquement de manière identique. Cela risquerait de créer un biais lors du calcul des statistiques basiques, ce qui entraînerait par la suite un résultat erroné lors d'analyses multidimensionnelles comme l'ACP[13] (voir 0).

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 la moitié) soit en supprimant les valeurs inférieures à la LQ ([8], [7]). 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 soit adaptée à la recherche d'un fond géochimique pour lequel au contraire ce sont les valeurs faibles qui sont très étudiées. La substitution des données inférieures à la LQ par la moitié de la LQ peut donc entraîner un biais puisque la distribution statistique de la population se trouve complètement faussée :

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 de la démonstration, le jeu de données est censuré artificiellement : les 30 valeurs les plus faibles sont remplacées par les valeurs de LQ/2 reportées 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).

[13] Analyse en Composantes Principales.

Les fonctions de répartition empiriques du plomb (équivalent d'une fonction de répartition mais sans biais, voir Annexe 1 - 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 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

2.3.3 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 repose 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 qu'au-delà d'un certain effectif, la plupart des lois peuvent être approchées par une loi normale (ou loi de Gauss) sous condition d'indépendance. La courbe de distribution de la loi normale, appelée courbe de Gauss ou plus communément « courbe en cloche » (Figure 4), est symétrique. Cette propriété explique la popularité des traitements statistiques adaptés à des populations dont la distribution suit une loi normale [10].

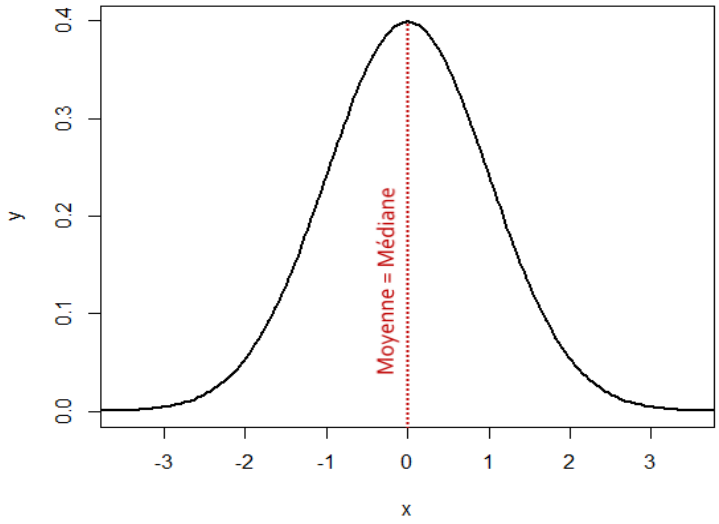


Figure 4 : Courbe de distribution de la loi normale $\mathcal{N}(0,1)$.

Un test de normalité (Shapiro-Wilk voir Annexe 1 - Fiche 1), permettant de déterminer si une population statistique suit une loi normale, est réalisé sur les analyses de plomb des agglomérations A, B (malgré leur faible effectif) et C (Tableau 4).

La p-value[14] résultante est de 0,042 pour l'agglomération A et est inférieure à 1.10^{-4} pour les deux autres agglomérations. L'hypothèse de normalité est donc rejetée au seuil de 5 % pour les trois agglomérations, autrement dit la population « plomb » ne suit pas une distribution normale.

[14] Si la p-value est inférieure au seuil « α=0,05 » l'hypothèse de normalité est rejetée (Annexe 1 – Fiche 1).

Agglomération	p-value
A (30 échantillons)	0.042
B (48 échantillons)	< 0.0001
C (97 échantillons)	< 0.0001

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

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

Comme nous l'avons vu ci-dessus, la majeure partie des tests/méthodes statistiques repose sur l'hypothèse que les variables (ou plus exactement les « résidus[15] ») 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 [9] 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 [9]. 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 [12]. 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 (10 000 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 :

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

[15] Résidus : termes d'une régression qui ne sont pas expliqués par les autres variables.
[16] Voir définition à la section 2.3.2.

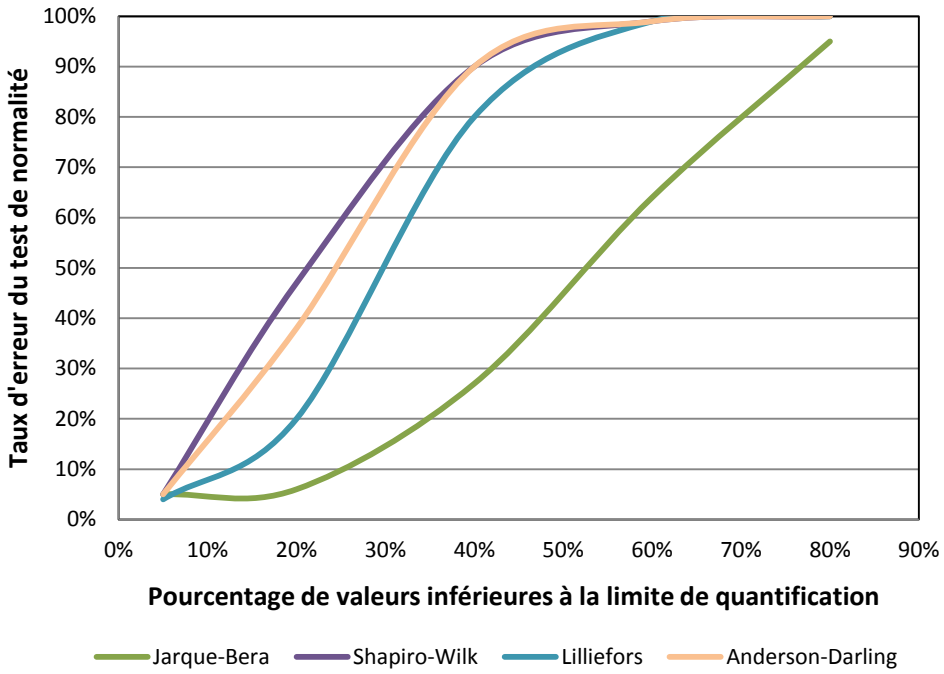


Figure 5 : Influence de la proportion de valeurs inférieures à la LQ sur les tests de normalité ([12] modifié).

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 des données du projet FGU.

2.3.4 Détections des outliers

La détection des outliers[17] est une des tâches clef de l'analyse statistique de données géochimiques dans le domaine minier. 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 des concentrations élevées dont il faudrait s'affranchir pour pouvoir étudier la distribution de la population [7].

[17] En pratique le terme « outlier » est souvent utilisé pour n'importe quel type de valeur s'écartant de la distribution étudiée.

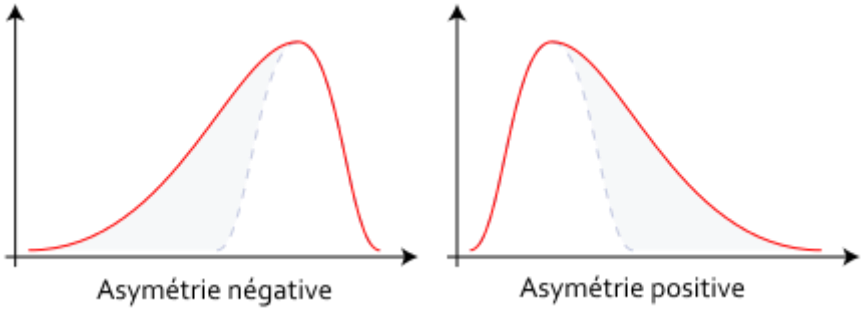


Figure 6 : Distribution asymétrique négative (gauche) et distribution asymétrique positive (droite).

Une solution des statistiques classiques (Gaussiennes), souvent utilisée pour traiter le problème des outliers, est celle de la *MEAN ± 2SD*[18] [5] produisant une estimation du seuil les séparant des valeurs appartenant au fond géochimique. Cependant, les outliers peuvent apparaître de façon erratique dans une distribution donnée, pas uniquement aux extrémités, et peuvent être difficiles à identifier.

La méthode *MEAN ± 2SD* semble fonctionner parce que fort heureusement, les outliers sont souvent très élevés (ou faibles) par rapport à la majeure partie des données [5]. 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.

[18] Moyenne plus ou moins deux fois l'écart-type (Mean ±2.Standard Deviation en anglais).

3. 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 30 635 en mai 2016. Ces résultats doivent être étudiés et comparés en tenant compte des caractéristiques de chaque population de données et de la répartition géographique des points de prélèvement spécifique à chaque agglomération. Deux études statistiques doivent donc être menées conjointement : l'une concernant la distribution spatiale des données sur la zone d'étude et l'autre concernant la distribution statistique des valeurs mesurées. Une étude statistique doit débuter par une étape de préparation des données. S'en suit l'étape d'interprétation à l'aide de représentations graphique et d'une analyse multidimensionnelle. Enfin, la détermination d'une valeur seuil peut être réalisée en utilisant une méthode calculatoire et/ou graphique selon les caractéristiques statistiques des populations étudiées. Les seuils résultants sont ensuite utilisés pour représenter les données spatialement. Pour chaque étape 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 anthropisé.

3.1 PRÉPARATION DES DONNÉES

3.1.1 Distribution et normalité

La plupart des analyses de sols ne se répartissent pas selon une loi normale. La représentation de leur distribution n'est pas une courbe de Gauss parfaitement symétrique autour de la moyenne. Cette courbe de distribution est souvent étirée vers les valeurs hautes. On constate cependant qu'une représentation de cette courbe en utilisant une échelle logarithmique pour les abscisses tend à lui donner un caractère symétrique : on dit que les données suivent une loi log-normale. La transformation logarithmique des données consiste donc à les substituer par la valeur de leur logarithme (base 10). Cette transformation équivaut à exprimer les données dans une nouvelle unité. 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 logarithmique n'est qu'un cas particulier de l'Échelle 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 »*.

Étant donné que la majorité des populations statistiques en domaine environnemental présente une asymétrie positive, les puissances inférieures à 1 présentent une utilité. La puissance $\lambda = 0$ correspond à la transformation logarithmique. En géochimie environnementale, la transformation logarithmique est largement utilisée dans l'objectif d'approcher une distribution normale [5] [7]. Mais la transformation $x^{1/2}$ est également utilisée.

Utilisation	λ	Transformation	Nom	Commentaire
Asymétrie (-)		…		Des puissances supérieures peuvent être utilisées
	3	x^3	Cube	
	2	x^2	Carré	
	1	x	Unité d'origine	Pas de transformation
Asymétrie (+)	½	$\sqrt{x}$	Racine carrée	Fréquemment utilisée
	1/3	$\sqrt[3]{x}$	Racine cubique	Fréquemment utilisée
	0	$\log(x)$	Logarithme	Fréquemment utilisée
	-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

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

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 est estimée par un algorithme statistique qui 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 moyenne ;
- ainsi que la diminution de la distance entre la moyenne et la médiane (respectivement la barre verticale et la croix rouge sur les boxplots).

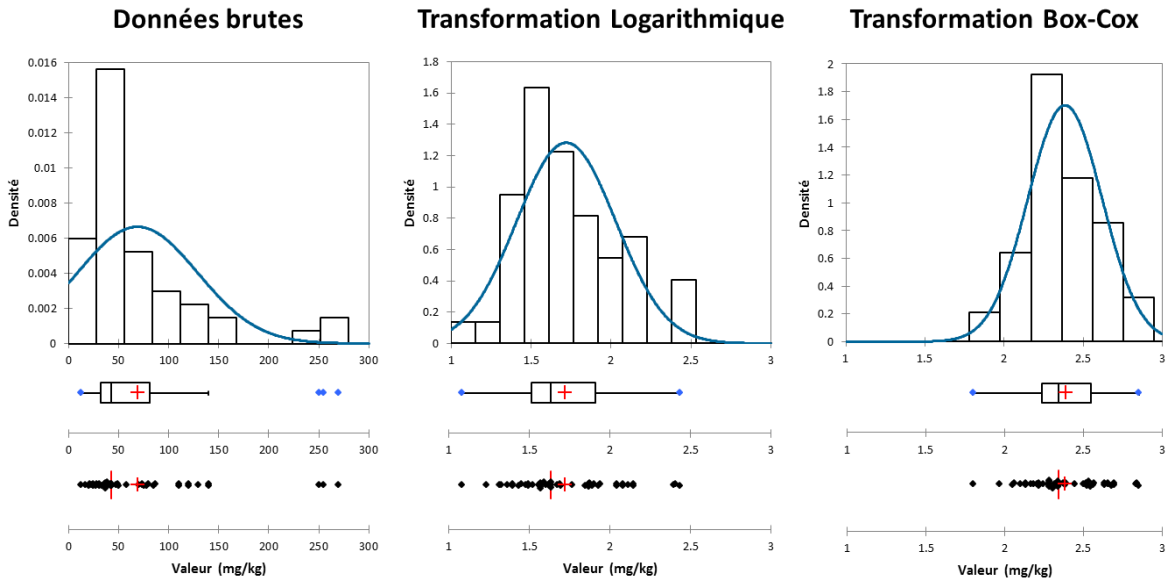


Figure 7 : Statistiques descriptives et tests de normalité pour les données de plomb de l'agglomération B (48 échantillons).

Les résultats du test de normalité réalisé sur les données brutes et transformées sont disponibles dans le Tableau 6 ci-dessous.

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)

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

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.

3.1.2 Centrage et réduction

Le traitement de données environnementales nécessite de prendre en compte des substances différentes simultanément, c'est le cas dans le cadre du projet FGU. Pour assurer la comparabilité des variables entre elles, il est parfois utile de recourir à leur transformation selon une loi centrée réduite.

Avant de réaliser certaines analyses multivariées, comme l'ACP, il peut être intéressant que les données soient centrées autour d'une même valeur [5]. Cette étape intervient après le traitement de la normalité des données par une transformation Box-Cox. 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 incomparables directement à cause de la variabilité des écarts-types [5]. En effet, certaines populations de données, présentent des courbes de distribution plus ou moins larges. Par ailleurs, les résultats des analyses ne s'échelonnent pas dans les mêmes gammes de concentrations en fonction des lieux et des substances ou familles de substances analysées. Cette hétérogénéité de répartition est très influente sur le résultat des tests multidimensionnels. La réduction permettra de palier cet effet. Elle consiste à diviser chaque valeur du jeu de données considéré par la valeur de l'écart-type.

3.2 INTERPRÉTATION DES DONNÉES

3.2.1 Statistiques descriptives pour données censurées

Denis Helsel [10] 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 présentant des dispersions statistiques[19] identiques mais 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 3 familles d'approches ayant un intérêt pour notre protocole d'analyse :

- les méthodes non-paramétriques ;
- la MLE (*Maximum Likelihood Estimation*[20]), méthode d'analyse de survie se basant sur une distribution supposée ;
- la méthode ROS (*Regression on Order Statistics*[21]).

Méthodes non paramétriques

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. À la place, ils utilisent les positions relatives (rangs) des valeurs. Ces méthodes se révèlent très utiles pour les jeux de données censurés puisqu'ils utilisent uniquement l'information disponible sans utiliser une hypothèse qu'il peut être difficile de vérifier.

[19] La dispersion statistique d'une population correspond à la tendance qu'on ses valeurs à se distribuer les unes par rapport aux autres ou par rapport à une valeur centrale.

[20] Estimation par maximum de vraisemblance.

[21] Méthode de régression sur les statistiques d'ordre.

Cette 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. Si les résultats comptent plusieurs LQ, il est 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 et 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'emploi, ces méthodes permettent d'utiliser les données sans supposer plus qu'il n'est possible de le faire avec l'information disponible. En comparaison, faire appel à la substitution revient à estimer des valeurs qui pourraient mener à un faux positif lors d'un futur test statistique.

Les procédures d'analyse de survie

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 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[22].

*MLE (*Maximum Likelihood Estimation*),*

La MLE robuste[23], de plus en plus utilisée en études environnementales (Owen and DeRouen, 1980 ; Miesch, 1967 cité dans [10]), repose sur l'utilisation de trois informations :

- les valeurs détectées au-dessus de la (ou des) 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 supposer que les données suivent 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 [10]). En revanche pour des effectifs supérieurs à 50, la MLE est toute indiquée.

*La méthode ROS (*Regression on Order Statistics*).*

Une dernière approche, la méthode ROS[24] robuste[25] (Annexe 1- 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

[22] Il existe une solution alternative sous le logiciel R© adaptée aux données censurées vers les valeurs faibles : la fonction cenfit de la librairie NADA présentée plus loin.

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

[24] Regression on Order Statistics.

[25] 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 [10].

distribution. À l'aide d'un PP-plot[26], 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 d'éviter la substitution, très répandue parce que simple d'utilisation mais insérant un biais dans les calculs effectués.

3.2.2 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 (Annexe 1 - 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 résument à un manque d'unicité de la représentation graphique [10].

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 [8] (voir Annexe 1 - Fiches 4, 5, 6 et 7 pour une description des méthodes si nécessaire).

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[27] (Annexe 1 - Fiche 8) déjà présentée (Figure 3).

[26] Probability-Probability plot - Diagramme probabilité-probabilité.

[27] EmpiriCal Distribution Function – Fonction de répartition empirique.

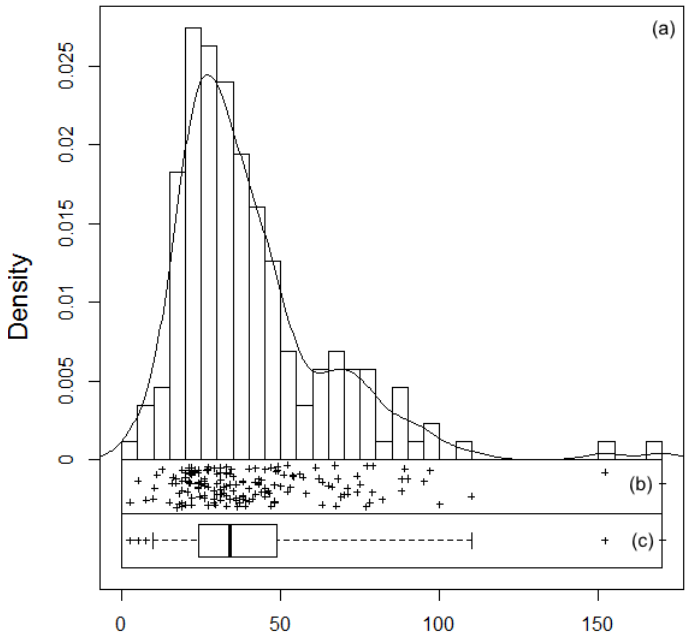


Figure 8 : Combinaison de graphiques descriptifs pour l'étude d'une distribution statistique - (a) Courbe de densité superposée à l'histogramme (b) Boxplot (c) Dispersogramme unidimensionnel.

3.2.3 Statistiques multidimensionnelles

Les études environnementales donnent souvent lieu à l'analyse de plusieurs substances. Les concentrations obtenues s'expliquent par plusieurs facteurs. Les statistiques multidimensionnelles fournissent des méthodes donnant un aperçu des tendances et relations au sein d'un 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 méthodes (Annexe 1 - Fiche 3). À partir d'un espace à q dimensions défini par q variables initiales, l'ACP a pour but de déterminer un sous espace de dimension moindre, en recherchant de nouvelles variables (les composantes principales), linéairement indépendantes, d'importance décroissante, expliquant au mieux l'ensemble des données.

L'ACP s'applique à un tableau croisant p individus et q variables numériques et conduit à représenter simultanément les points des deux ensembles, individus et variables. L'analyse consiste d'abord à rechercher l'axe passant par le centre de gravité du nuage de points représentant les individus et sur lequel les distances seront en moyenne le mieux conservées. On cherche en fait l'axe (« Dim 1 » voir Figure 9a) sur lequel la moyenne des carrés des distances entre les projections des points est maximale. Il correspond donc à l'axe d'élongation la plus large du nuage. On cherche ensuite un second axe (« Dim 2 » voir Figure 9a) perpendiculaire au premier rendant également maximale la moyenne des carrés des distances entre les projections des points.

Sur le *graphique des individus* (Figure 9a), 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. Chaque point résulte d'une combinaison linéaire des variables initiales. 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 9a la première dimension (ou composante) explique 48,58 % de la variabilité du jeu de données tandis que la deuxième dimension en exprime 17,31 %.

Le graphique des variables (Figure 9b) permet d'interpréter les composantes principales et de repérer rapidement les groupes de variables liées entre elles ou opposées. Ce graphique est obtenu en représentant chaque variable par un point dont les coordonnées sont ses corrélations avec les deux premières composantes principales. Il est nécessaire à la compréhension du graphique des individus : plus l'abscisse d'un point est élevée sur la dimension 1, plus sa concentration sera élevée. À 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 9a.

Ainsi des tendances peuvent être détectées. Dans l'exemple ci-dessous, la méthode est appliquée à la comparaison des données obtenues pour les analyses de métaux en milieu urbain par le projet FGU et des données de la base BDETM (Base de données des éléments traces métalliques) de l'INRA dans les zones rurales entourant les agglomérations considérées. Les données ne sont pas obtenues exactement selon les mêmes protocoles. Toutefois, on a pris soin ici de ne considérer que les données de la BDETM obtenues après minéralisation à l'eau régale comme c'est le cas pour les données FGU.

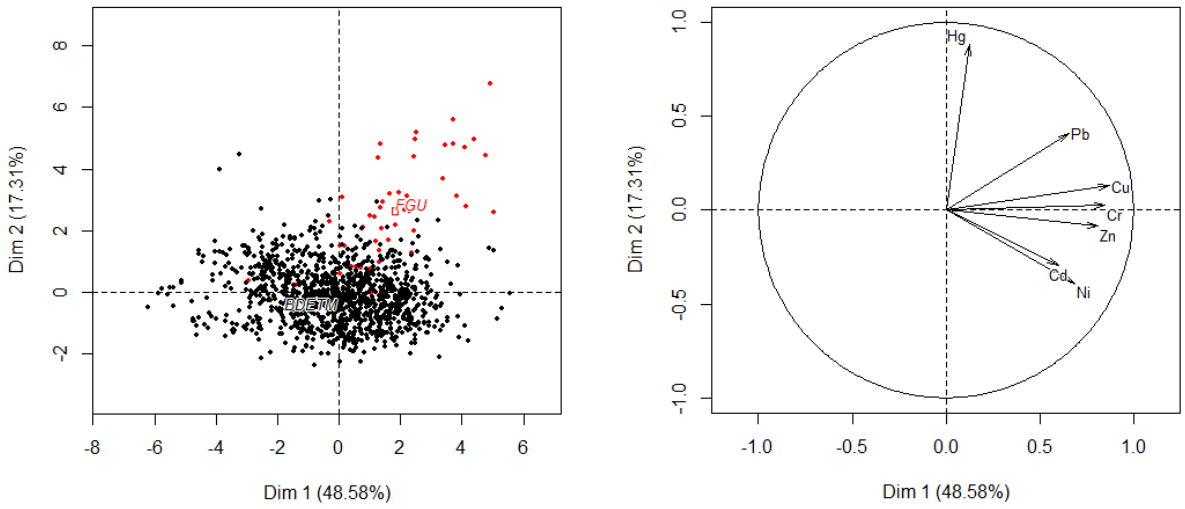


Figure 9 : Agglomération B : ACP des données FGU (en rouge) et BDETM (en noir) : (a) Graphique des individus et (b) Graphique des variables.

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.

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 3.1.1). 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, 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.

3.3 DÉTERMINATION D'UNE VALEUR SEUIL

Si on se base sur la définition fournie par la norme ISO 19258 [7] une **teneur pédo-géochimique** correspond à la « *teneur d'une substance présente dans un sol du fait de processus géologiques et pédologiques naturels, à l'exception des substances introduites dans les sols du fait de l'activité humaine* ». Le fond pédo-géochimique est définit par des **valeurs de bruit de fond pédo-géochimique** qui sont des « *caractéristiques statistiques*[28] *de la teneur pédo-géochimique* ». Ainsi, la détermination d'une valeur seuil n'est pas nécessairement le meilleur moyen de définir un fond pédo-géochimique urbain. Cependant, il peut être utile ou pratique de déterminer un seuil permettant de déterminer si une valeur appartient ou non à la population statistique représentative du fond pédo-géochimique.

Pour cela, il convient d'analyser les diverses méthodes de calcul d'un fond-géochimique disponibles dans la littérature afin de sélectionner la, ou les, plus adaptées au contexte définit par la norme. Les méthodes analysées dans ce rapport proviennent en partie du répertoire établi dans le rapport BRGM RP-64845-FR [3] et également des sources bibliographiques utilisées au cours de l'étude [5], [15], [16], [17], [18]. Toutes sont soumises à diverses interprétations.

Il existe globalement deux types de méthodes :

- les méthodes dites calculatoires comme la *MEAN ± 2.SD*, les vibrisses internes du boxplot de Tukey, les centiles ou la *MED ± 2.MAD*[29]*.* Elles se basent sur des critères de distribution de la population étudiée en supposant que celle-ci est normale ;
- les méthodes dites graphiques comme la droite de Henry, la tangente à la courbe des fréquences cumulées, etc. Elles reposent sur la construction d'une courbe illustrant la répartition des données puis la recherche d'un point d'inflexion mettant en évidence deux (ou plus) populations différentes.

On remarquera que les méthodes calculatoires considèrent que l'échantillon étudié est composé d'une population statistique majoritaire alors que le principe des méthodes graphiques s'appuie sur la présence de deux populations (exemple : une population non anthropisée et une population anthropisée).

De manière générale, la précision des deux types de méthodes est fortement tributaire de l'effectif de la population statistique. L'ensemble des sources consultées, de même que la norme ISO 19258, s'accorde sur une valeur de 30 échantillons minimum pour la détermination d'un fond géochimique [5], [6], [7].

3.3.1 Méthodes calculatoires

Les méthodes calculatoires les plus répandues sont décrites en Annexe 1 :

- moyenne ± 2 x Écart-Type (voir Annexe 1 - Fiche 10) ;
- médiane ± 2 x Écart-Médian-Absolu (voir Annexe 1 - Fiche 12) ;
- centiles 95, 97,5 ou 98 (voir Annexe 1 - Fiche 11) ;
- vibrisse interne supérieure du boxplot de Tukey (voir Annexe 1 - Fiche 11).

[28] Exemples : moyenne, valeur médiane, écart-type ou les percentiles de la distribution de fréquence [7].

[29] Médiane plus ou moins deux fois l'écart médian absolu (Median Absolute Deviation en anglais).

Ces méthodes fournissent deux seuils encadrant le FPGA. Ces seuils permettent ainsi d'identifier les outliers inférieurs et supérieurs et de définir le fond pédo-géochimique comme une gamme des valeurs les plus fréquentes sur une zone d'investigation.

Le Tableau 7 contient le résultat des calculs à partir des données du projet FGU pour le cuivre pour les agglomérations A, B et C (un tableau regroupe en Annexe 2 les résultats pour trois autres substances). Pour chaque méthode, le calcul du seuil est réalisé avec les valeurs brutes puis avec les valeurs log-transformées (sauf pour la méthode des centiles pour laquelle le résultat serait identique). La population de l'élément cuivre est un exemple assez simple puisqu'elle ne comporte qu'une seule valeur censurée.

	MEAN ± 2.SD				**Boxplot**				**MEDIAN ± 2.MAD**				**Centiles**	
	MEAN - 2.SD		MEAN + 2.SD		Vib inf		Vib sup		MED - 2.MAD		MED + 2.MAD		2%	98%
	brut	*log*	*brut*	*log*	*brut*	*log*	*brut*	*log*	*brut*	*log*	*brut*	*log*		
A *(30)*	1,9	14	84	104	10,2	10,2	90	90	-1,9	10	76	132	15	84
B *(48)*	-1,2	11	74	89	7.5	12,6	63	95	6,3	14	**60**	80	10	88
C *(97)*	-14	9,8	99	127	2,6	11	89	**152**	-2,6	11	69	100	11	113
ABC *(175)*	-8,5	11	90	112	2,6	9,8	82	110	1,4	12	67	96	10	99

L'effectif des échantillons est affiché sous le nom de chaque agglomération.

Tableau 7 : Comparaisons des méthodes de détermination de seuil calculatoires pour le cuivre (mg/kg) dans les agglomérations A, B et C et pour la combinaison des trois. (NB : les valeurs obtenues après transformation logarithmique ont été rétro transformées à l'échelle d'origine).

Premièrement, on peut constater la différence des résultats fournis par les différentes méthodes utilisées. La différence entre les seuils hauts le plus élevé (bleu) et le plus faible (orange) est de 92 mg/kg.

Deuxièmement, des différences non négligeables sont visibles entre les valeurs calculées à partir de données brutes et à partir de données log-transformées. Par exemple 63 mg/kg de différence entre les vibrisses supérieures calculées pour l'agglomération C.

La méthode *MEAN ± 2.SD* est considérée par la plupart des auteurs comme inadaptée au contexte de détermination d'un fond géochimique [3], [5], [16]. En effet, elle fournit des estimations basées sur une distribution normale de la population statistique en utilisant des estimateurs sensibles à la présence de valeurs extrêmes, ce qui est souvent le cas pour les données environnementales.

Les centiles donnent une information intéressante puisque l'on peut interpréter que 98 % des données de l'agglomération C, par exemple, se trouve en dessous de la valeur 113 mg/kg. Les valeurs au-dessus de cette limite peuvent donc être sélectionnées pour une inspection plus poussée. Cependant, l'utilisation de cette méthode sous-entend qu'il existe toujours le même pourcentage d'outliers dans une population statistique, cette approche n'est pas forcément valide [5].

Les deux autres méthodes, *MED* ± *2.MAD* et « vibrisse interne supérieure » sont des estimateurs moins couramment implémentés dans les logiciels mais sont basées sur des estimateurs robustes. Toutefois, l'hypothèse de normalité reste nécessaire, une transformation logarithmique (ou Box-Cox) doit donc être appliquée au jeu de données avant le calcul des valeurs seuils. On recommande donc uniquement l'utilisation de ces deux méthodes pour le protocole de traitement.

3.3.2 Méthodes graphiques

Les méthodes graphiques analysées au cours de cette étude sont les suivantes :

- méthode de Lepeltier [17] ;
- méthode d'estimation semi-automatique fondée sur la méthode de la droite de Henry et sur la méthode de Lepeltier [18] ;
- tangente à la courbe de la fonction de répartition empirique [5] ;
- méthode de répartition des concentrations.

Elles reposent sur la construction d'une courbe illustrant la répartition des données puis la recherche d'un point d'inflexion mettant en évidence deux (ou plus) populations différentes. Elles ont toutes un fonctionnement très similaire, donc pour simplifier l'étude, on prendra ici l'exemple de la « Fonction de répartition empirique ».

La Figure 10 correspond à l'application de la méthode « Fonction de répartition empirique » aux jeux de données de cuivre des agglomérations A, B et C.

On observe l'avantage que représente l'utilisation des fonctions de répartition empirique au niveau de la visibilité de chaque point de mesure. Cette caractéristique favorise l'identification des outliers et permet d'observer leur distance par rapport au cœur des données (masse principale).

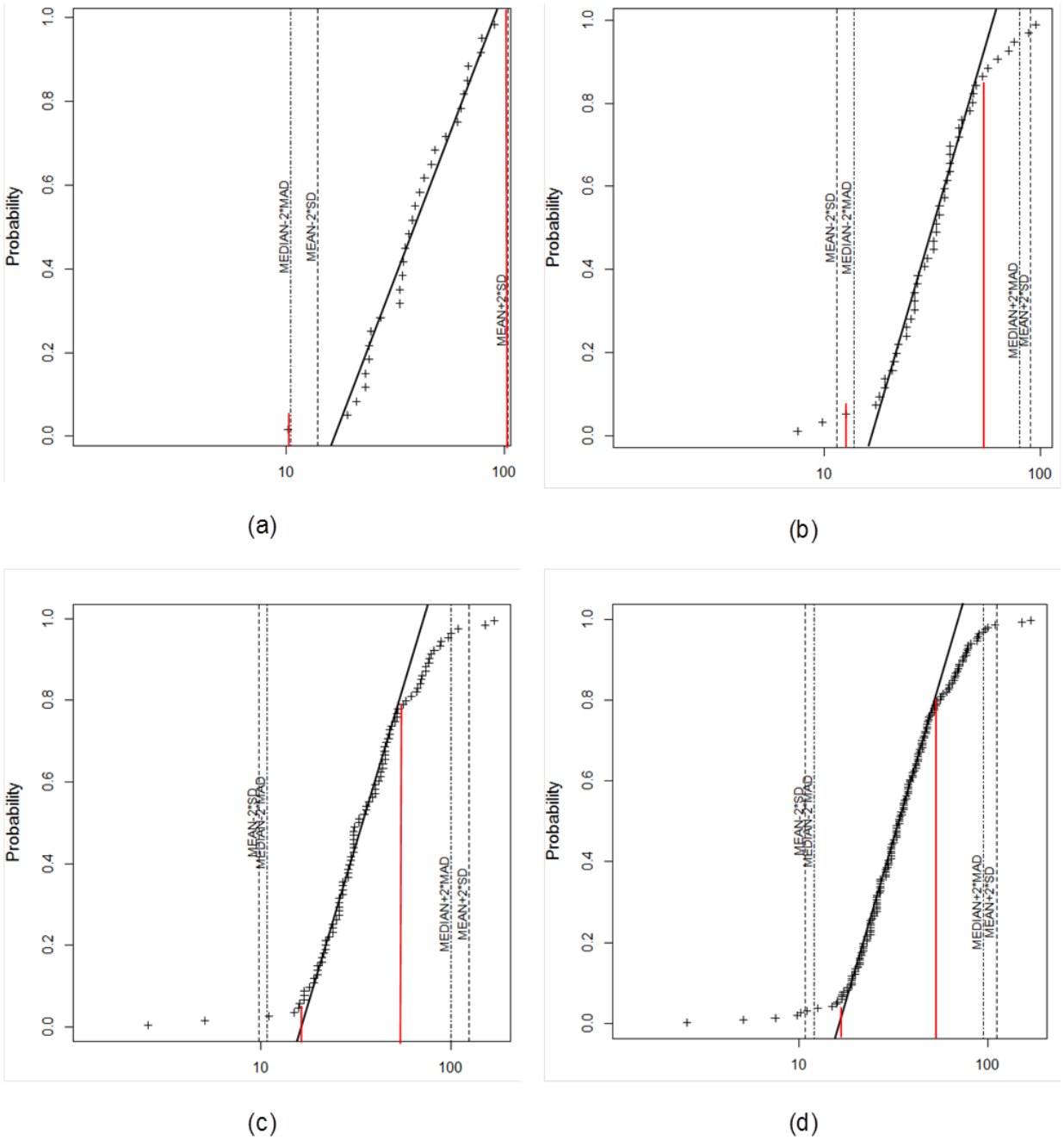


Figure 10 : ECDF de la population cuivre en mg/kg (échelle logarithmique) des agglomérations (a) A, (b) B, (c) C et (d) de la combinaison des données des trois agglomérations - (échelle logarithmique).

En combinant les informations visibles sur l'ECDF, il devient possible de tracer une droite épousant au mieux la partie linéaire de la courbe normalement sigmoïde. La valeur seuil supérieure (respectivement inférieure) correspond à l'abscisse de la première valeur s'écartant de la droite tracée. Sur les graphiques de la Figure 10, les projections permettant la détermination des seuils depuis les points d'inflexion sont représentées par des droites verticales rouges.

Les valeurs seuils calculées sur les données log-transformées par les méthodes « MEAN ± 2.SD » et « MED ± 2.MAD » sont également projetées sur les graphiques (après transformation inverse) à titre de comparaison.

Les résultats de l'interprétation des fonctions de répartition empirique (Tableau 8) montrent que les seuils déterminés sont relativement homogènes excepté pour l'agglomération A. Cette différence est principalement due au faible effectif disponible pour cette agglomération.

Toujours pour l'agglomération A, le nombre de points déviants de la droite principale est quasiment nul ce qui rend impossible la visibilité de points d'inflexion. Pour les autres agglomérations, le nombre d'outliers est un peu plus élevé mais reste cependant très faible. Il devient convenable pour le jeu de données combinant les échantillons des trois agglomérations (effectif avoisinant 200 individus).

	Fonction de répartition empirique	
	Seuil inférieur	Seuil supérieur
A *(30)*	9	90
B *(48)*	12	51
C *(97)*	15	54
ABC *(175)*	12	54

Tableau 8 : Valeurs seuils déterminées graphiquement à partir de la population de cuivre (mg/kg) dans les agglomérations A, B et C et pour la combinaison des trois.

L'utilisation de cette méthode est cependant controversée :

- premièrement pour la difficulté à repérer visuellement les points d'inflexion dans le cas d'effectifs faibles, ce qui est souvent le cas quand l'intérêt est porté sur une agglomération particulière. À plus grande échelle, cette méthode peut s'avérer très utile et plus simple à interpréter ;
- deuxièmement, même quand l'effectif est suffisant ; l'identification des points d'inflexions peut s'avérer compliquée en fonction de la clarté avec laquelle la (ou les) population(s) d'outliers se démarque(nt) du fond pédo-géochimique. En effet, on peut imaginer sur le graphique (d) de la Figure 10 un deuxième point d'inflexion parmi les valeurs hautes qui serait en adéquation avec la valeur de la méthode « MED ± 2.MAD ». Il existe une grande part de subjectivité dans l'interprétation qui peut être réalisée, interprétation qui dépend de l'expérience du statisticien ;
- troisièmement, la précision obtenue en utilisant cette méthode, comme toutes les méthodes graphiques, est réduite comparée aux méthodes calculatoires. En effet, il est difficile de fournir un résultat au dixième près quand la position du seuil présente lui-même une variabilité élevée.

Le Tableau 9 récapitule les résultats obtenus au moyen des méthodes calculatoires en 3.3.1 et la méthode graphique par répartition empirique ci-dessus. D'une agglomération à l'autre (donc d'un jeu de données à l'autre), on remarque que les différences entre deux méthodes ne sont pas systématiquement de même signe ni de mêmes proportions. Ce phénomène a déjà été observé par Reimann (2005) [16].

Enfin, en comparant les résultats obtenus avec la méthode graphique de répartition empirique et ceux obtenus grâce aux méthodes de la vibrisse interne et « MED ± 2MAD », on constate une similitude entre les seuils bas calculés. Une différence beaucoup plus importante (amplitude de 50 mg/kg environ) est visible pour les seuils hauts calculés. Une raison possible serait que les méthodes graphiques sont plus adaptées aux données prélevées selon un plan d'échantillonnage systématique. En effet, dans ce cas le jeu de données résultant contient une proportion d'outliers plus importante [7] ce qui facilite l'apparition d'un point d'inflexion clair.

Pour résumer, les méthodes graphiques nécessitent un effectif relativement élevé (≥ 200) et, s'il y a lieu, doivent être utilisées en complément de méthodes calculatoires afin de s'assurer de la fiabilité du seuil déterminé.

Bornes inférieures		Méthodes					Récapitulatif		
Agglomérations	**n**	**MEAN - 2.SD**	**Vib. inf**	**MED - 2.MAD**	**2e cent.**	**Graphique Repart. Emp.**	**Min.**	**Max.**	**Diff.**
A	**30**	14	10,2	10	15	9	9	15	40
B	**48**	11	12,6	14	10	12	10	14	29
C	**97**	9,8	11	11	11	15	9,8	15	35
ABC	**175**	11	9,8	12	10	12	9,8	12	18

Bornes supérieures		Méthodes					Récapitulatif		
Agglomérations	**n**	**MEAN + 2.SD**	**Vib. sup**	**MED + 2.MAD**	**98e cent.**	**Graphique Repart. Emp.**	**Min.**	**Max.**	**Diff.**
A	**30**	104	90	132	84	90	84	132	**36**
B	**48**	89	95	80	88	51	51	95	46
C	**97**	127	152	100	113	54	54	152	64
ABC	**175**	112	110	96	99	54	54	112	52

Tableau 9 : Comparaison des résultats obtenus avec les méthodes calculatoires et graphiques étudiées pour les agglomérations A, B et C et pour ces trois populations réunies.

3.3.3 Intégration de la variabilité spatiale dans le calcul du FPGA

Jusqu'à maintenant aucune des méthodes proposées dans ce rapport ne tient compte du problème soulevé en section 2.3.1 par la répartition spatiale des échantillons.

Une solution possible proposée par Riemann [5] serait de combiner les résultats obtenus selon les méthodes calculatoires et graphiques avec une représentation cartographique des seuils calculés. En effet, il peut être intéressant pour les usagers de disposer d'une carte représentant la répartition spatiale des mesures supérieures au seuil supérieur du FPGA. Le seuil inférieur n'est pas utilisé ici car l'intérêt est porté sur les outliers hauts. La répartition spatiale de ces derniers permettra d'identifier les zones présentant des teneurs élevées par rapport au FPGA et de valider ainsi le seuil déterminé par les méthodes calculatoires et graphiques.

Des cartes utilisant ce mode de représentation ont été réalisées avec les données du cuivre, du cadmium, de l'arsenic et du pyrène pour les agglomérations A, B et C (Annexe 3). Le rayon des points représentés est directement proportionnel à la teneur mesurée dans l'échantillon. Dans un souci de comparabilité inter-agglomérations, une échelle unique est utilisée pour la taille des points. La vibrisse interne supérieure et la méthode « MED + 2MAD » ont été utilisées pour produire les seuils à partir des populations log-transformées (Annexe 2).

Vibrisse supérieure et MED + 2MAD

La Figure 9 de l'annexe 3 illustre ce mode de représentation pour le cuivre de l'agglomération C. Avec la méthode de la vibrisse supérieure, deux valeurs sont identifiées comme outliers hauts et donc au-delà du seuil haut FPGA de l'agglomération. En revanche la méthode « MED + 2.MAD » (Annexe 3 - Figure 12) est plus contraignante, elle identifie deux outliers en supplément des deux précédents.

La répartition et l'effectif des points identifiés ne sont pas suffisants pour identifier une zone préférentielle qui présenterait des valeurs supérieures au FPGA.

Tangente à la fonction de répartition empirique (ECDF)

Cependant, à titre de comparaison, une carte (Annexe 3 - Figure 16) a été réalisée pour le cuivre de l'agglomération C avec le seuil déterminé par la méthode de la tangente à l'ECDF. Ce seuil, 54 mg/kg (voir Tableau 8), est beaucoup plus faible que ceux calculés à partir des autres méthodes (respectivement 152 mg/kg et 100 mg/kg pour la vibrisse supérieure et la méthode MED + 2MAD, voir Tableau 7). L'ECDF permet donc d'identifier 21 outliers. On peut à présent identifier une zone, au nord de l'agglomération C, présentant une accumulation d'outliers. Bien que la fiabilité de la méthode de la tangente à l'ECDF soit remise en question (voir section 3.3.2), il est intéressant de constater que son utilisation met en évidence une répartition spatiale particulière des outliers hauts. Il serait intéressant de tenter de justifier cette observation par des recherches bibliographiques et/ou archivistiques sur les potentielles sources de contaminations diffuses dans cette zone (ex : industrie métallurgique cuprifère).

Une information pratique est disponible sur les cartes : l'identification des points représentants des mesures censurées. Leur poids statistique étant réduit par rapport aux mesures quantifiées, il est intéressant de les différencier de ces dernières. L'intérêt de cette représentation se comprend aisément en observant les teneurs en cadmium de l'agglomération C (Annexe 3 - Figure 13) qui comprend 52 valeurs censurées sur un effectif de 97 individus.

Plusieurs remarques peuvent être faites sur l'utilisation de ce mode de représentation :

- la carte résultante est fortement influencée par la taille et la localisation de la zone d'investigation. Pour des zones d'études plus étendues et des effectifs plus importants, l'apparition de secteurs particulièrement touchés par la présence d'outliers est beaucoup plus probable [5] ;
- les seuils déterminés par la vibrisse supérieure et la méthode « MED + 2MAD » sont proches sur nos exemples. Les cartes produites selon ces deux méthodes sont donc peu différentes (*i.e.* : le nombre d'outliers identifiés est souvent proche). La différence est beaucoup plus flagrante en comparant ces résultats avec ceux obtenus par la tangente à l'ECDF ;
- sur certaines cartes, les points représentants les teneurs en mg/kg ont un diamètre très faible, ces cartes sont donc illisibles (Annexe 3 - Figure 14). Cet effet est du au choix de l'échelle unique entre les trois agglomérations (pour une même subsance). Le rayon des points étant directement proportionnel à la teneur mesurée, en présence d'une valeur extrême dans une agglomération donnée, la représentation des points dans les autres agglomérations est affectée. Une solution possible est de changer d'échelle et donc de redéfinir le diamètre des points en fonction de l'agglomération désirée (ex : Annexe 3 - Figure 15).

La principale limite de l'utilisation de ce mode de représentation est qu'il ne permet pas de distinguer l'origine naturelle ou anthropique des outliers identifiés.

Concentration Area plot

Un autre outil permettant d'intégrer la variabilité spatiale au calcul du FPGA a été identifié : le CA-plot (*Concentration Area plot*) [5]. Son intérêt réside dans son approche fractale de la structure des données qui consiste à émettre l'hypothèse suivante :

La distribution spatiale des teneurs suit une configuration spatiale « répétitive ».

Le principe de cette méthode permet d'étudier le pourcentage de l'aire occupée par une valeur donnée. Par définition, les valeurs appartenant au fond géochimique seront les plus fréquentes et représenteront un pourcentage élevé de la zone d'investigation. A l'inverse, les zones contenant des valeurs extrêmes (hautes ou basses) représenteront un pourcentage faible sur le CA-plot.

L'étude de la courbe ainsi réalisée (recherche d'un point d'inflexion) permet d'identifier les différentes distributions fractales présentes au sein du jeu de données.

De plus, cette approche nécessite l'intervention d'une interpolation, ce qui permet d'augmenter le nombre de données disponibles sur la zone d'étude. Ce point constitue un avantage non négligeable face aux faibles effectifs disponibles actuellement à l'échelle de l'agglomération.

Cette méthode est difficile à mettre en place pour le projet FGU :

- elle requiert un plan d'échantillonage systématique ;
- la forte variabilité spatiale des teneurs sur de faible distance est déconseillée pour l'application de cette méthode ;
- les tests d'interpolation spatiale déjà réalisés pour les données contenues dans BDSolU se sont, en partie, révélées infructueuses [19].

4. Mise en application du protocole de traitement statistique

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 invalides.

4.1 ÉTAPE 1 - CALCUL DE PARAMÈTRES DESCRIPTIFS

Étape 1

Figure 11 : Étape 1 de l'arbre de décision du traitement statistique.

Toute analyse statistique doit débuter avec le calcul des paramètres basiques descriptifs de la population étudiée (Figure 11) : nombre d'observations, pourcentage d'observations censurées, minimum, moyenne, médiane, 1er et 3ème quartiles, maximum, etc[30]. Le calcul s'effectue ici avec les données brutes en conservant les LQ (Tableau 10 en Annexe 2).

Les résultats fournis par ce tableau doivent être utilisés pour l'application du protocole statistique. 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 des données, très fréquentes, sont facilement repérables.

En fonction des contextes d'études, l'intérêt peut être porté sur des paramètres différents de ceux conseillés ici.

[30] Pour plus de détails sur l'interprétation des paramètres statistiques basiques se référer à [5].

4.2 ÉTAPE 2 - ÉTUDES SIMPLE ET RAPIDE DES OUTLIERS

Étape 2

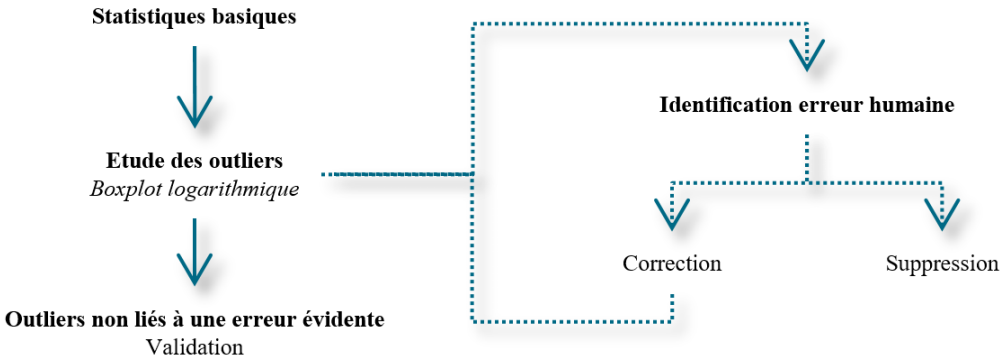


Figure 12 : Étape 2 de l'arbre de décision du traitement statistique.

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 (Figure 12). Une étape de traitement plus fine est prévue dans la suite du protocole afin de détecter les outliers restants.

Un boxplot sur données transformées doit être tracé avec l'ensemble des analyses pour 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 d'administration 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 avec les valeurs log-transformées 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 et leur examen est plus court.

4.3 ÉTAPE 3 - ACP DE DÉTERMINATION DES TENDANCES

Étape 3

Figure 13 : Étape 3 de l'arbre de décision du traitement statistique.

Après cette étape, les estimations suivantes seront moins biaisées. On peut procéder à une ACP afin de détecter les tendances de groupe (Figure 13). La variable qualitative à étudier 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 ».

Il est nécessaire de s'assurer que la somme des inerties des axes considérées est suffisante pour décrire la variabilité du jeu de données. Dans notre cas, cette somme fait 50 %. Ce qui signifie que 50 % de la variabilité du jeu de données sont expliqués par les deux axes retenus.

Ici deux ACP ont été réalisées avec les données des agglomérations B et C afin de confronter les analyses contenues dans la base BDSolU et les analyses fournies par la base BDETM dans les zones rurales environnantes. L'objectif est de vérifier si les concentrations en ETM observées dans une agglomération sont significativement différentes de celles obtenues dans le milieu rural proche. Cette hypothèse a été avancée dans le rapport BRGM RP-64845-FR [3] à l'aide d'histogrammes. L'ACP est proposée ici en vue d'améliorer l'analyse statistique par une approche multivariée.

Les données de la base BDETM de l'INRA sont obtenues selon des protocoles de prélèvements et d'analyses différents de ceux mis en œuvre dans le cadre du projet FGU, en particulier les méthodes de préparation des échantillons. En effet, les résultats BDETM ont été obtenus à la fois après minéralisation à l'acide fluorhydrique et à l'eau régale. Les analyses du projet FGU sont obtenues uniquement par cette dernière méthode. Ici seules les données obtenues par minéralisation à l'eau régale ont été utilisées grâce à une sélection préalable. Cette démarche permet d'améliorer la comparabilité des données confrontées dans les ACP.

Premièrement, on remarque que la somme des deux dimensions de chaque ACP est supérieure à 50 % (Figures 14 et 15), la variabilité des jeux de données est donc correctement exprimée par les calculs des ACP. La première dimension, représentée par l'axe horizontal, est très corrélée positivement avec le zinc, le chrome et le cuivre, dans une moindre mesure avec le plomb et le cadmium et peu avec le nickel et le mercure. La deuxième dimension corrélée positivement avec le nickel et le cadmium et est corrélée négativement avec le plomb et le mercure. Les trois autres éléments sont peu corrélés avec la deuxième dimension. Ainsi, sur le graphique des individus, plus un point se situe à droite plus il s'écarte de la moyenne par des fortes concentrations en zinc, chrome, cuivre et plomb (et dans une moindre mesure pour les autres éléments). De même, plus un point se situe en

haut sur le graphique, plus il s'écarte de la moyenne par de fortes concentrations en nickel et par de faibles concentrations en mercure. Aux points situés en bas à gauche du graphique correspondent des caractéristiques inverses.

Deuxièmement, les deux graphiques des individus montrent deux nuages bien distincts ce qui confirme l'hypothèse que deux fonds géochimiques distincts peuvent être identifiés entre le domaine rural et le domaine urbain.

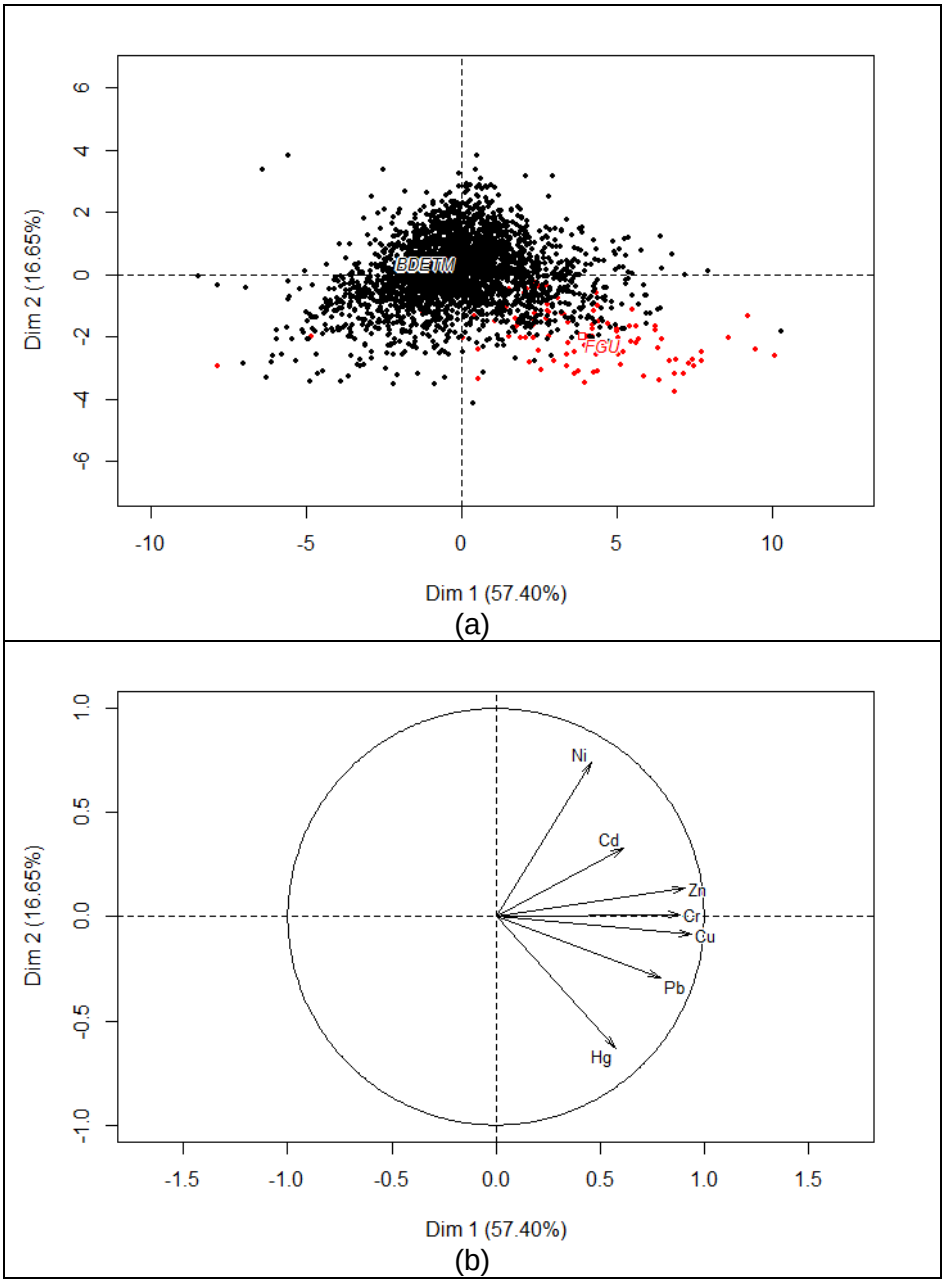
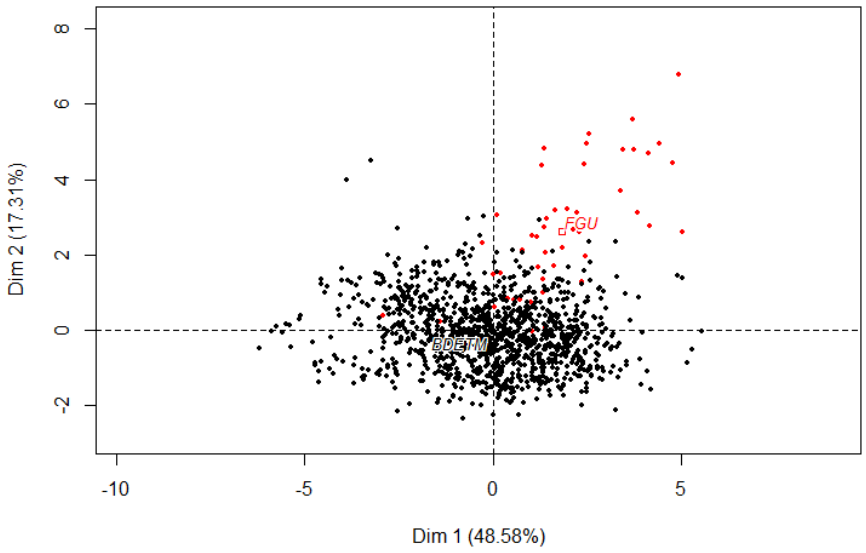
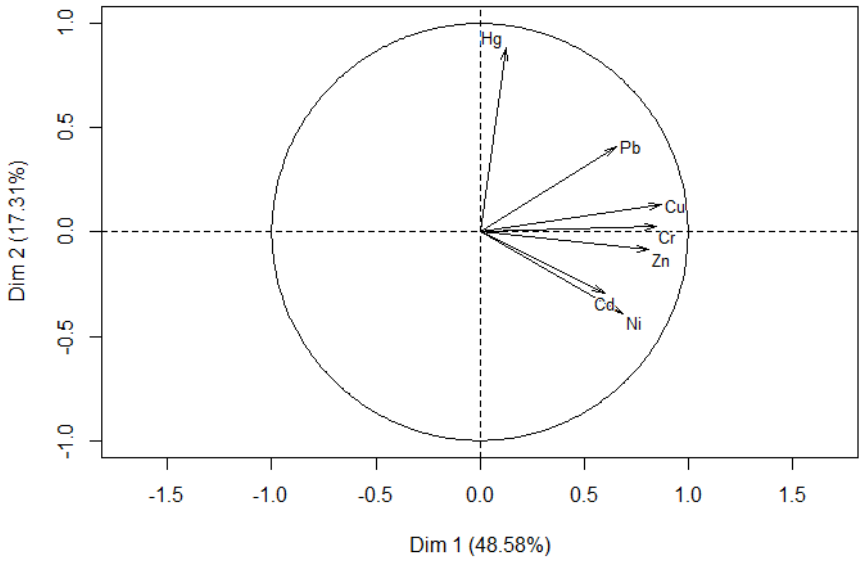


Figure 14 : ACP logarithmique des données FGU (en rouge) et BDETM (en noir) autour de l'agglomération C: (a) Graphique des individus et (b) Graphique des variables.

Troisièmement, la position du centre des nuages urbains (FGU) par rapport au centre des nuages ruraux (BDETM) indique, en se référant aux graphiques des variables, que la teneur des éléments traces métalliques est plus élevée en domaine urbain qu'en domaine rural.



(a)



(b)

Figure 15 : ACP logarithmique des données FGU (en rouge) et BDETM (en noir) autour de l'agglomération B: (a) Graphique des individus et (b) Graphique des variables.

L'ACP a donc permis de confirmer la tendance de groupe différenciant le domaine rural du domaine urbain. Le fait que les données BDETM fournies pour réaliser les comparaisons correspondent au département considéré est critiquable. Les évolutions des teneurs géochimiques dans les sols ne sont pas tributaires des découpages administratifs. Il faudrait tester plusieurs échelles autour d'une agglomération afin d'identifier la meilleure permettant d'exprimer la variabilité spatiale d'un fond géochimique.

Un exemple d'ACP inter-agglomérations est disponible en annexe 1 (Fiche 3) et confirme également l'hypothèse d'une tendance de groupe par agglomération et donc de l'existence d'un fond géochimique distinct par agglomération.

4.4 ÉTAPE 4 - TEST DE NORMALITÉ

Étape 4

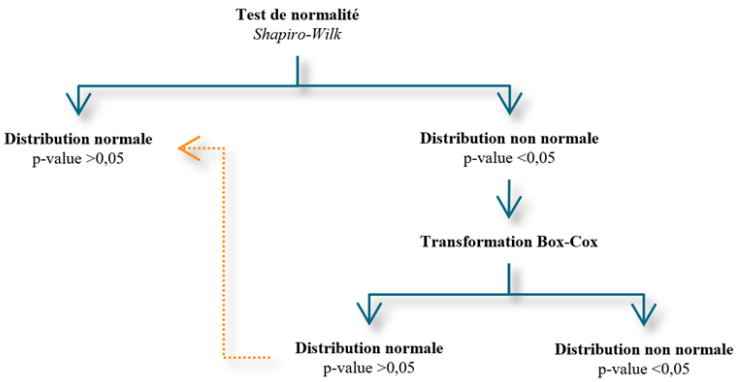


Figure 16 : Étape 4 de l'arbre de décision du traitement statistique.

Avant de poursuivre l'étude statistique plus loin, il est nécessaire de réaliser un test de normalité sur le jeu de données étudié (Figure 16). La normalité ou non des données est déterminante sur le choix des tests à utiliser. Elle permet de différencier les populations qui nécessitent ou non une transformation avant de continuer le traitement. Le test de Shapiro-Wilk (Annexe 1 - Fiche 1), considéré comme le plus fiable dans la littérature [11], est recommandé pour ce protocole. À 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 ».

4.5 ÉTAPE 5 - DÉTERMINATION DE VALEURS DESCRIPTIVES POUR LES DONNÉES CENSURÉES

Étape 5

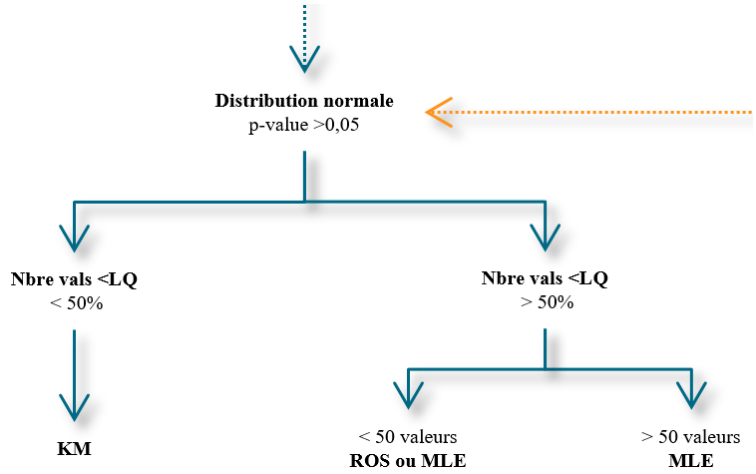


Figure 17 : Étape 5 de l'arbre de décision du traitement statistique.

L'étape 5 fait intervenir les approches proposées par Helsel (section 3.2.1) pour la détermination de valeurs descriptives pour des données censurées (Figure 17). 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 [10] :

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.

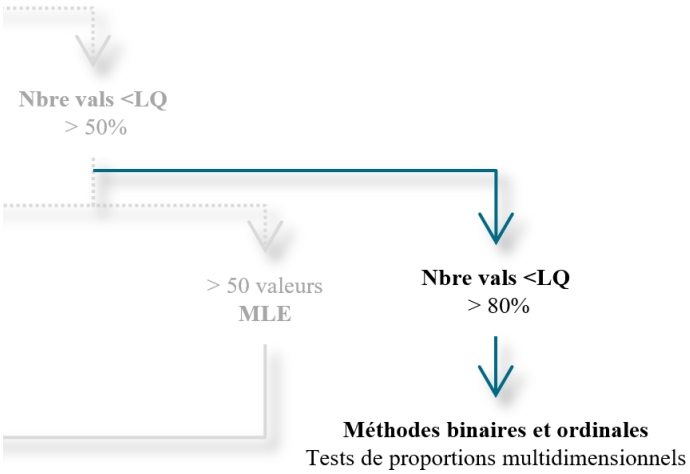


Figure 18 : Détail de l'étape 5 de l'arbre de décision du traitement statistique.

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 (Figure 18).

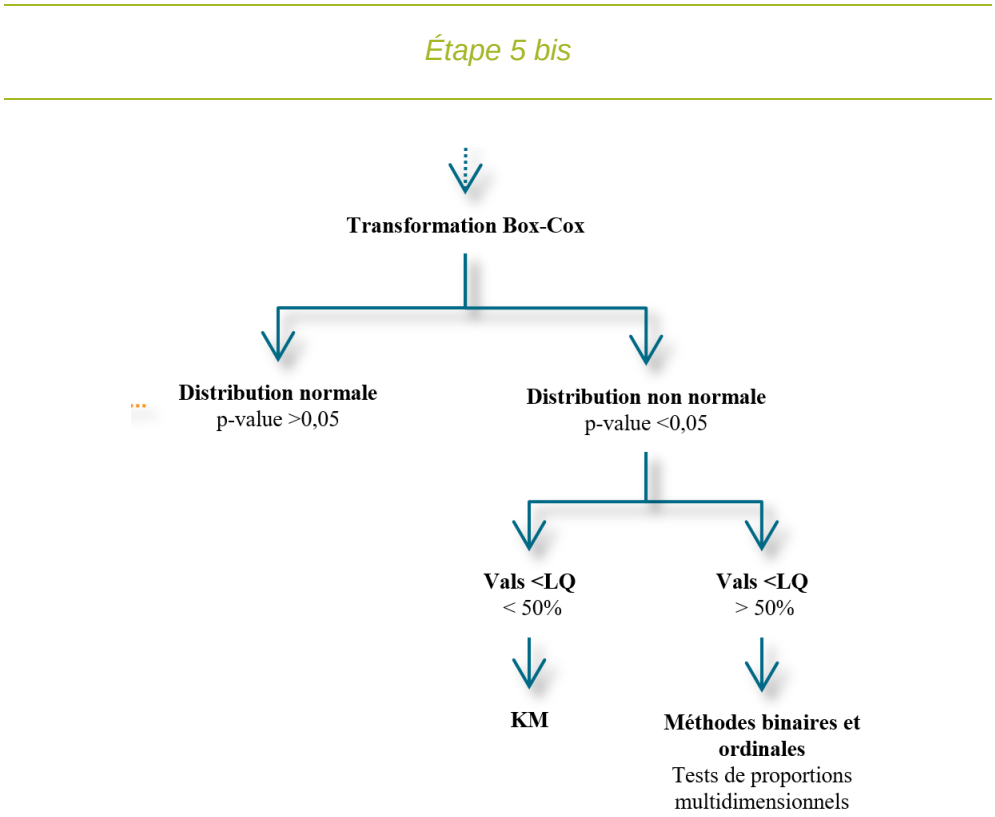


Figure 19 : Étape 5bis de l'arbre de décision du traitement statistique.

Au terme de l'étape 4, la distribution de certains jeux de données est jugée non normale 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 des méthodes de proportions comme dans l'étape précédente [10] (Figure 19).

4.6 ÉTAPE 6 - TRAITEMENT GRAPHIQUE

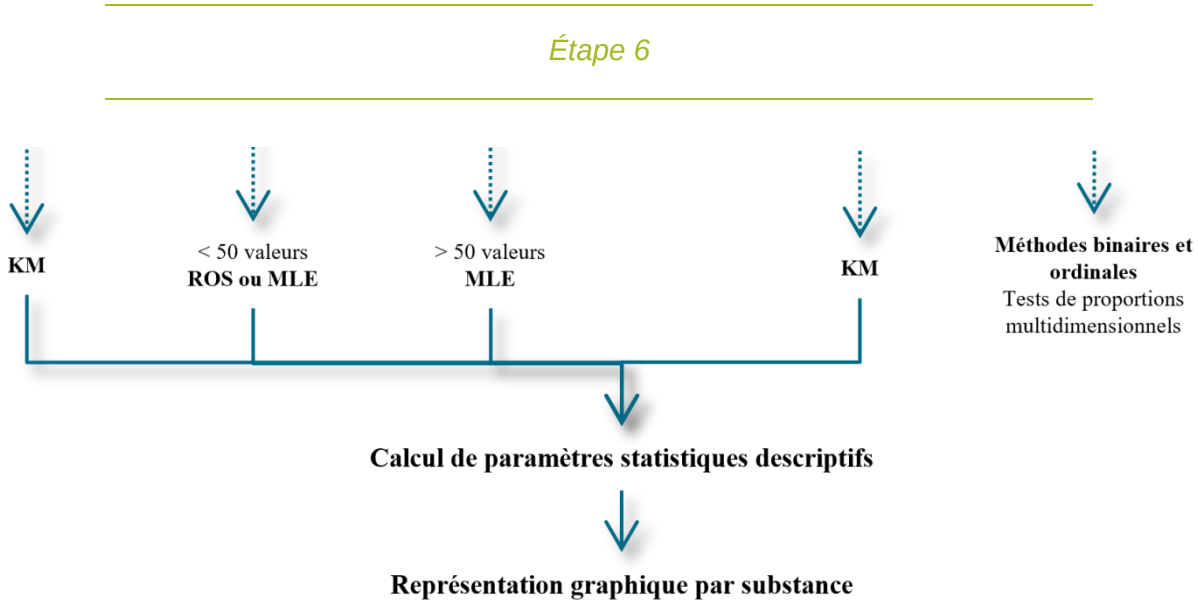


Figure 20 : Étape 6 de l'arbre de décision du traitement statistique.

Après la détermination des paramètres descriptifs de chaque population statistique, l'étape de traitement graphique peut être abordée (Figure 20). Pour ce faire, la combinaison histogramme, densité, boxplot et dispersogramme unidimensionnel est fortement recommandée (section 3.2.2). Un ensemble de graphique doit être tracé pour chaque substance analysée.

Les graphiques supérieurs (Figure 21) représentent la distribution des données de cuivre sans transformation. On observe l'histogramme et la densité caractéristiques d'une distribution asymétrique positive causée en grande partie par le point avec une teneur proche de 500 mg/kg ; teneur clairement identifiable grâce à l'ECDF, le dispersogramme unidimensionnel et le boxplot.

Le boxplot et le dispersogramme unidimensionnel permettent également d'identifier des outliers hauts et bas : le boxplot de manière directe en les représentants sous forme de point isolé en dehors de la boîte centrale et le dispersogramme unidimensionnel de manière intuitive de par leur distance par rapport au nuage central.

La pertinence de l'analyse graphique peut être améliorée grâce à une transformation logarithmique[31]. Le résultat, Figure 21 graphiques inférieurs, permet d'observer la forme caractéristique de la courbe de Gauss épousant la distribution symétrique des barres de l'histogramme. L'effet est également visible sur la courbe de l'ECDF qui possède maintenant une forme sigmoïde caractéristique d'une population normale.

[31] On utilise ici une transformation logarithmique parce que l'on dispose d'un script R complet permettant de produire la Figure 21 [5]. Ce script nécessiterait une amélioration pour intégrer la transformation Box-Cox. Pour l'exemple présenté la transformation logarithmique est suffisante.

L'effet des outliers est également réduit grâce à la transformation, cela est visible sur le boxplot et le dispersogramme unidimensionnel. Le nuage de points, ou la boîte centrale, correspondant aux valeurs rencontrées le plus fréquemment est maintenant plus étendu et permet de mieux juger la répartition de la population. Les données montrent à présent une distribution symétrique nécessaire pour la détermination du fond pédo-géochimique.

En ayant respecté les étapes précédentes du protocole, tous les jeux de données identifiés comme ayant une distribution non normale devrait avoir subi une transformation Box-Cox. Il est quand même nécessaire de tracer les graphiques avec les données brutes. Ils contiennent des informations non négligeables et permettront de faire une comparaison avec les graphiques des données transformées. Enfin, ils possèdent l'échelle originale des données brutes ce qui présente un avantage évident par rapport à une échelle transformée.

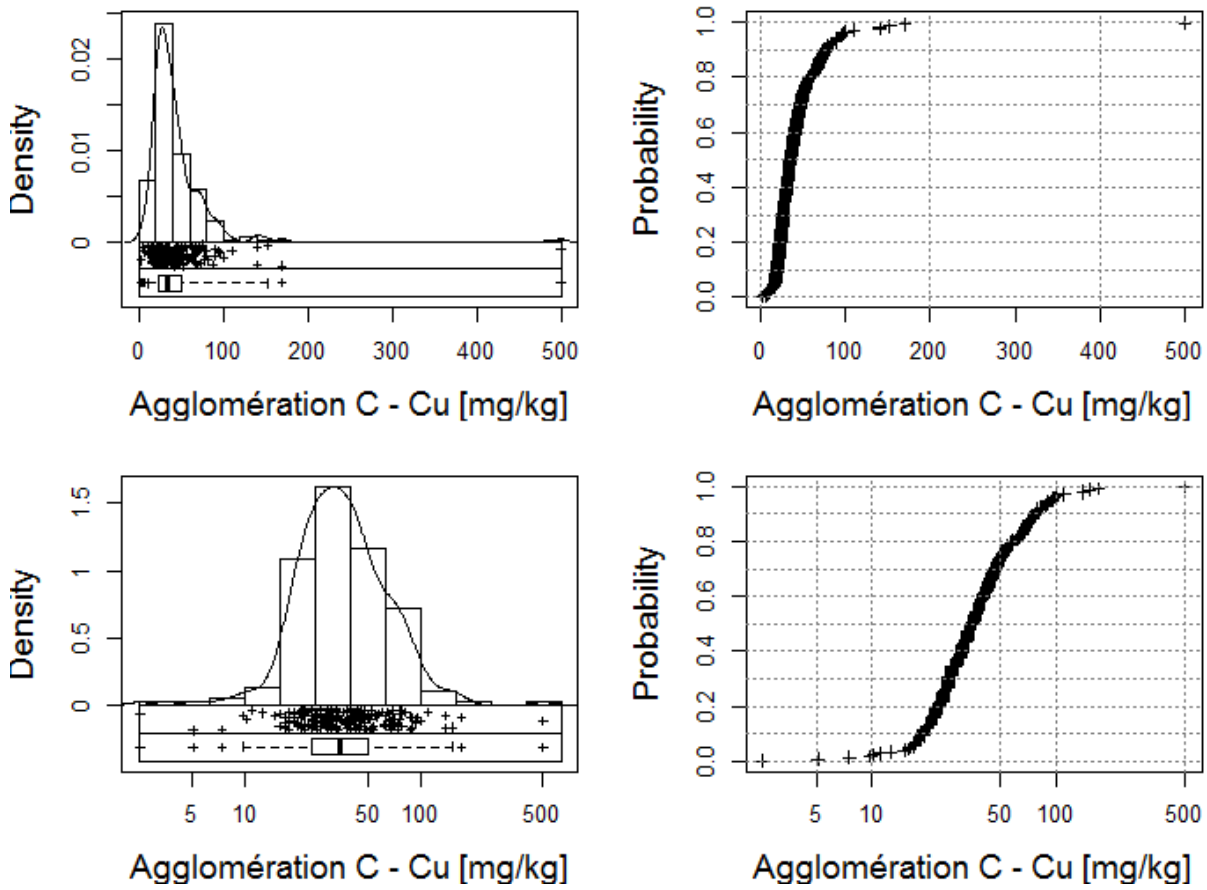


Figure 21 : Représentations graphiques brutes (haut) et logarithmique (bas) de la population cuivre (mg/kg) pour l'agglomération C.

La Figure 21 correspond à un exemple facile à interpréter grâce à un effectif relativement élevé et une population globalement homogène. L'exemple de l'arsenic pour l'agglomération B, Figure 22, est moins simple et montre un histogramme et une courbe densité peu interprétable (graphique supérieurs). La transformation logarithmique permet de mettre en évidence une distribution asymétrique négative (graphiques inférieurs). On peut également remarquer plusieurs ensembles de points alignés verticalement sur les courbes de l'ECDF (graphiques de droite) à cause de plusieurs résultats d'analyses donnant le même résultat et de la censure à la même valeur de limite de quantification 1 mg/kg pour les trois valeurs les plus faibles. Cet effet diminue la lisibilité de l'ECDF qui était déjà réduit par le faible effectif.

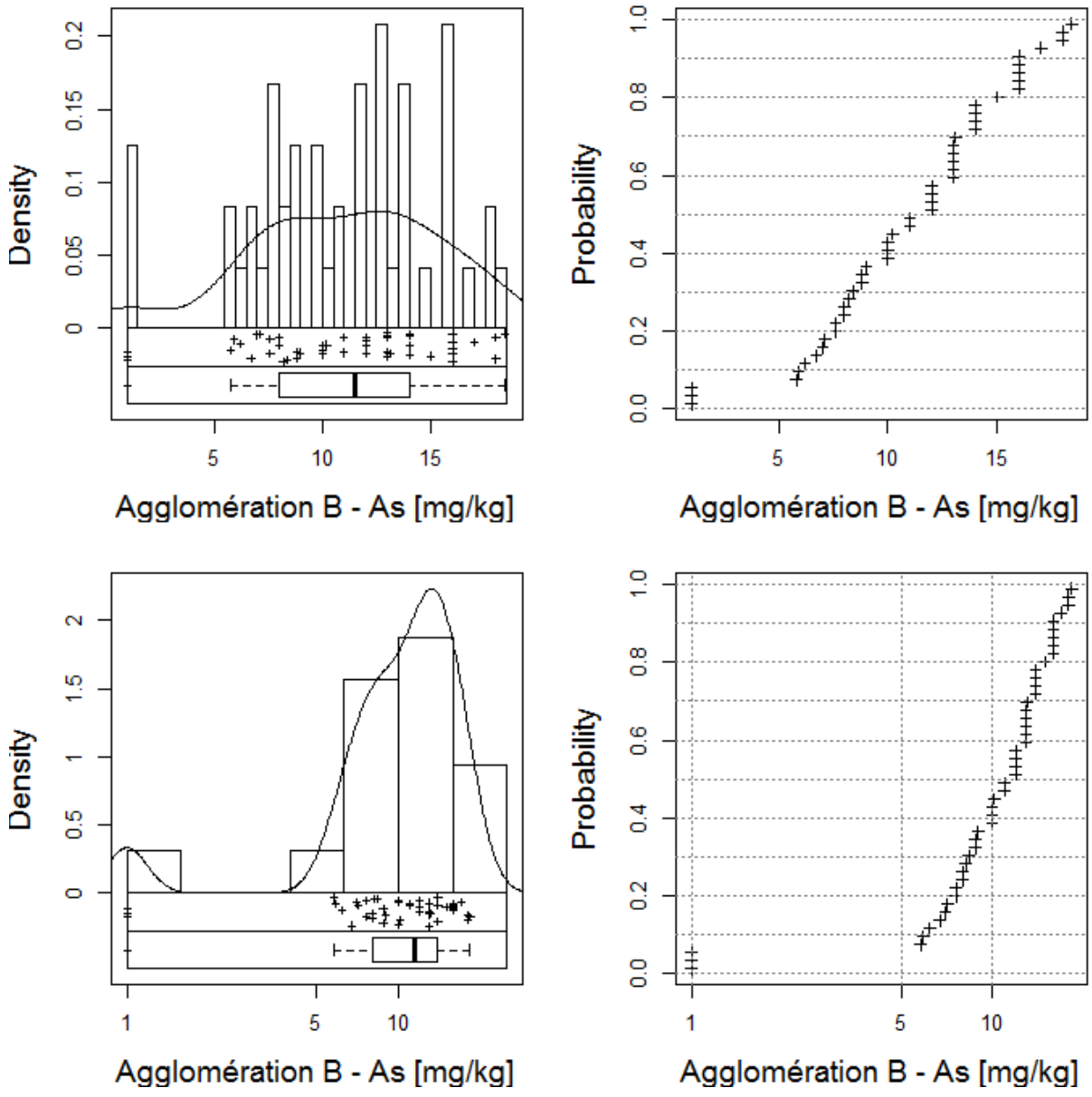


Figure 22 : Représentations graphiques brutes (haut) et logarithmique (bas) de la population arsenic (mg/kg) pour l'agglomération B.

4.7 ÉTAPE 7 - ÉTABLISSEMENT DU FOND PÉDO-GÉOCHIMIQUE ANTHROPISÉ

Étape 7

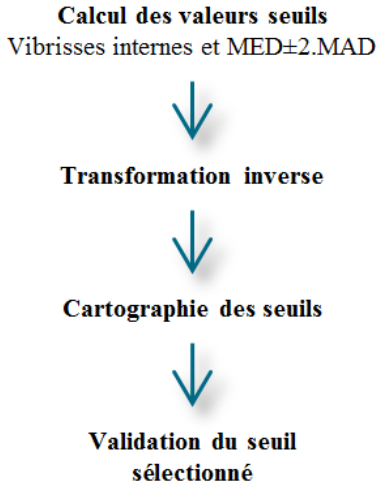


Figure 23 : Étape 7 de l'arbre de décision du traitement statistique.

Après cet ensemble d'étapes, l'établissement d'un fond pédo-géochimique anthropisé peut être envisagé (Figure 23).

Il convient premièrement d'écarter du jeu de données toutes les substances ayant un effectif inférieur à 30 individus[32] et ce en prenant en compte uniquement les valeurs non censurées. De ce fait un jeu de données comportant 34 échantillons dont 10 étant des mesures en dessous de la limite de quantification ne sera pas sélectionné. Dans notre cas, les populations AsA, CdA, CdB, PyrA et PyrB (voir Tableau 11) doivent être écartés, nous les garderons à titre de comparaison pour l'étude.

Les données censurées sont remplacées par la valeur de la limite de quantification (substitution à 100 %). Le traitement préalable (transformation) et la robustesse des tests utilisés permettent de réduire l'effet causé par la censure.

Les jeux de données transformés (parce qu'identifiés comme présentant une distribution non normale) doivent être utilisés pour cette étape. Ici, par souci de comparaison, nous effectuerons également les calculs pour les données brutes. Une fois le calcul effectué, la transformation inverse doit être appliquée afin de revenir à l'unité des données brutes.

Les résultats sont disponibles dans le Tablau 11 (Annexe 2).

Les cartes géochimiques (Annexe 3) peuvent maintenant être construites avec les seuils calculés.

[32] Dans le cas d'une analyse globale (sur la France entière par exemple) il ne faut en aucun cas les supprimer.

5. Conclusion

Les méthodes Kaplan-Meier, Maximum Likelihood Estimation et Regression on Order Statistics permettent de fournir des statistiques descriptives adaptées aux données à faible effectif, censurées et asymétriques. Il convient cependant de vérifier la normalité des distributions avant d'appliquer ces méthodes. La fiabilité des calculs proposés dans ce rapport repose sur l'hypothèse de normalité des distributions. Lorsque la population statistique étudiée n'est pas normale, une transformation Box-Cox peut être appliquée avant de poursuivre le traitement.

La représentation graphique des données doit être réalisée à l'aide de la combinaison d'un histogramme, d'une densité, d'une boîte à moustaches, d'un dispersogramme unidimensionnel et d'une fonction de répartition empirique. Leur utilisation individuelle peut conduire à des erreurs d'interprétation puisque chaque représentation met l'accent sur un aspect différent des distributions statistiques. En les combinant on s'assure de disposer du maximum d'informations que l'on peut tirer du jeu de données étudié.

Nous proposons d'approcher la détermination du FPGA à partir du calcul de la moyenne plus deux écart types calculé sur un échantillon dont la distribution est gaussienne (avec une transformation si besoin) et selon les méthodes calculatoires suivantes : « *MEAN* ± *2SD* » et les vibrisses internes de la boîte à moustaches. Il existe également des méthodes graphiques, très utilisées dans la littérature, mais leurs caractéristiques ne s'adaptent pas aux données disponibles principalement en cas d'effectif trop faible et d'absence de données représentatives d'une population caractéristique d'une contamination anthropique marquée. Les données des points SLE pourraient être utilisées pour employer des méthodes graphiques dans la mesure où une partie d'entre eux appartient à une distribution présentant des valeurs plus hautes que les prélèvements SLU. Leur prise en compte peut en effet faciliter l'apparition d'un point d'inflexion souvent difficile à repérer.

Une carte géochimique doit être réalisée pour visualiser la relation entre la répartition spatiale des valeurs mesurées avec les seuils du FPGA déterminés. La concordance de ces derniers avec des zones caractérisées par un ensemble de points dont la valeur est supérieure au FPGA pourrait permettre l'identification d'une source de pollution ou d'une anomalie naturelle. L'effectif réduit des données disponibles dans les trois agglomérations étudiées ne permet pas de mettre en évidence suffisamment de points pour justifier les seuils calculés. Là encore une solution pourrait consister en l'inclusion des données obtenues avec les prélèvements SLE. Une partie de ces données apparaîtrait comme supérieure aux seuils du FPGA calculés. Une autre, correspondant aux sols des établissements ne posant pas de problème pour la ou les substances considérées, pourrait être considérée comme représentative du FPGA.

La réalisation d'un programme informatique sous R© (www.r-project.org) connecté à la base de données PostgreSQL et permettant l'automatisation des tests et calculs proposés faciliterait grandement le traitement des données et permettrait de le rendre utilisable à un usager non statisticien. De plus, une étude géostatistique poussée est à envisager afin d'explorer les possibilités qu'offre la simulation de variables aléatoires dans l'objectif de produire des jeux de données simulés à effectif augmenté. Cette solution permettrait l'utilisation du *Concentration Area plot* pour déterminer un fond géochimique en considérant la répartition spatiale des points de mesures (grâce au protocole proposé ici, cela est possible en seulement deux étapes).

6. 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] - http://www.installationsclassees.developpement-durable.gouv.fr/Politique-de-gestion-des-sites-et.html [Accès le 07 04 2017].

[2] **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].

[3] **J.-F. Brunet, F. Guiet, C. Blanc, V. Laperche, P. Balon et N. Aubert,** « Établissement de fonds pédo-géolchimiques urbains et industriels en parallèle à l'Opération ETS du Ministère du Développement durable » rapport BRGM RP-64845-FR - (2015).

[4] **MEEM**, « Établissements 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 en Avril 2016] - http://www.installationsclassees.developpement-durable.gouv.fr/Demarche-Etablissements-Sensibles.html [Accès le 07 04 2017].

[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] **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.

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

[11] **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).

[12] **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).

[13] **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).

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

[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] **F. P. G. R. REIMAN C.**, «Background and threshold: critical comparison of method of determination,» vol. 346, n° %1Page 1-16, (2005).

[17] **C. LEPELTIER,** « A Smiplified Statistical Treatment of Geochemical data by graphical Representation » *Economic Geology,* vol. 64, n° % 1 Page 538-550, (1969).

[18] **L. CARY et A. LEYNET**, « Proposition de valeurs du fond géochimique naturel des sols à partir des données de l'Inventaire Minier National » BRGM, (2011).

[19] **JF. Brunet, F. Guiet, E. Taffoureau, B. Bourgine, C. Blanc et L. Sancho**, « Établissement de fonds pédogéochimiques urbains en parallèle à l'Opération ETS du Ministère de l'Écologie. Rapport intermédiaire » BRGM, Orléans, (2016).

[20] **J.-M. P. Mompelat**, « Unités cartographiques et évaluation de l'aléa mouvements de terrain en guadeloupe (Antilles françaises) » BIUS JUSSIEU PARIS, Paris, (1994).

Annexe 1

Fiches descriptives des méthodes statistiques utilisées et traitement sous R software

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 concentrations d'arsenic de l'agglomération A suit 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]

Shapiro Wilk avec R Software

Traitement classique pour un groupe

ENTREE :

VALEUR
0,1
0,5
0,9
1
…

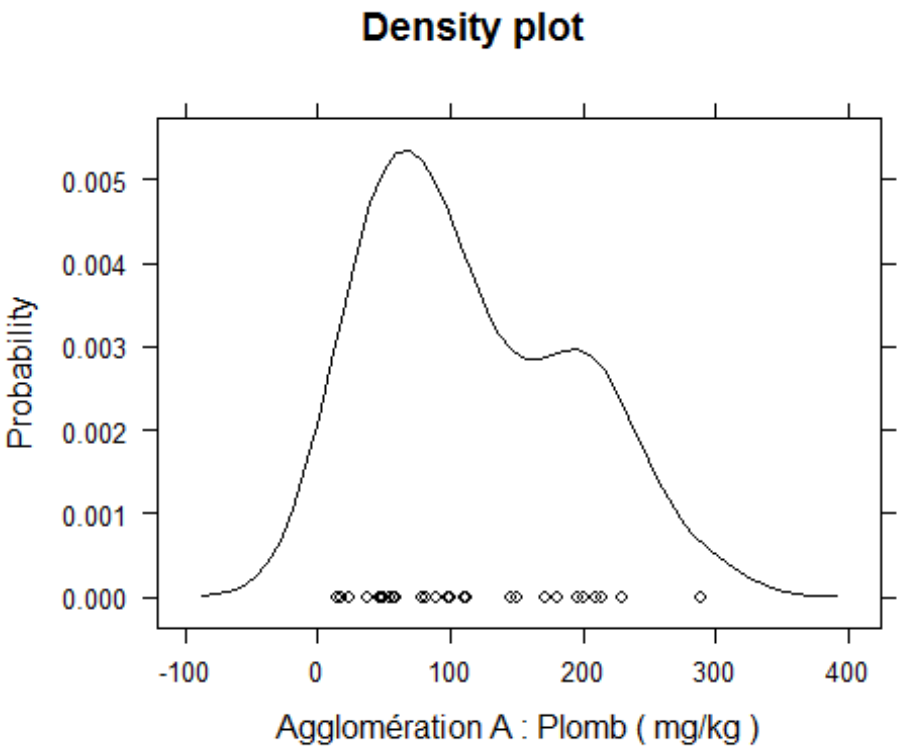
CODE :

```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("library")
# ********************************************************

######### Traitement classique pour un groupe #########
library("lattice")

TABLE <- read.csv("NomFichier.csv", header = T, sep = ";") #Nom du fichier
csv à modifier

#Test
res = shapiro.test(TABLE$VALEUR)

#Stockage du résultat
W = round(res$statistic, 3) #Valeur arrondie à 3 chiffres après la virgule
pvalue = round(res$p.value, 3)

#Affichage dans la console
matrice <- matrix(c(W, pvalue), nrow=1, ncol=2, byrow=T)
rownames(matrice) <- c("Plomb : ") #Nom à modifier
colnames(matrice) <- c("W","  p-value")
matrice

#Paramètres modifiables du graphique
title_plot = "Density plot"
label_x = "Agglomération A : Plomb"
label_y = "Probability"
unite = "mg/kg"
```

```
densityplot(TABLE$VALEUR, plot.points = TRUE,
            main = title_plot,
            xlab = paste(label_x,"(",unite,")"),
            ylab = label_y,
            #Mise en forme modifiable
            # xlim = c(-100,400), #Echelle de l'axe des abscisses
            # ylim = c(-0.0005,0.0055), #Echelle de l'axe des ordonnées
            col = "black", #Couleur
            lty = 1, #Style de trait
            cex = 0.7) #Taille des points
```

SORTIE :

```
##              W   p-value
## Plomb :  0.927     0.042
```



Traitement classique pour plusieurs groupes

ENTREE :

GROUPE	VALEUR
A	0,1
A	0,5
A	0,9
…	…
B	0,2
B	1,5
B	2,3
…	…
C	4 ,8
C	5
C	5,1

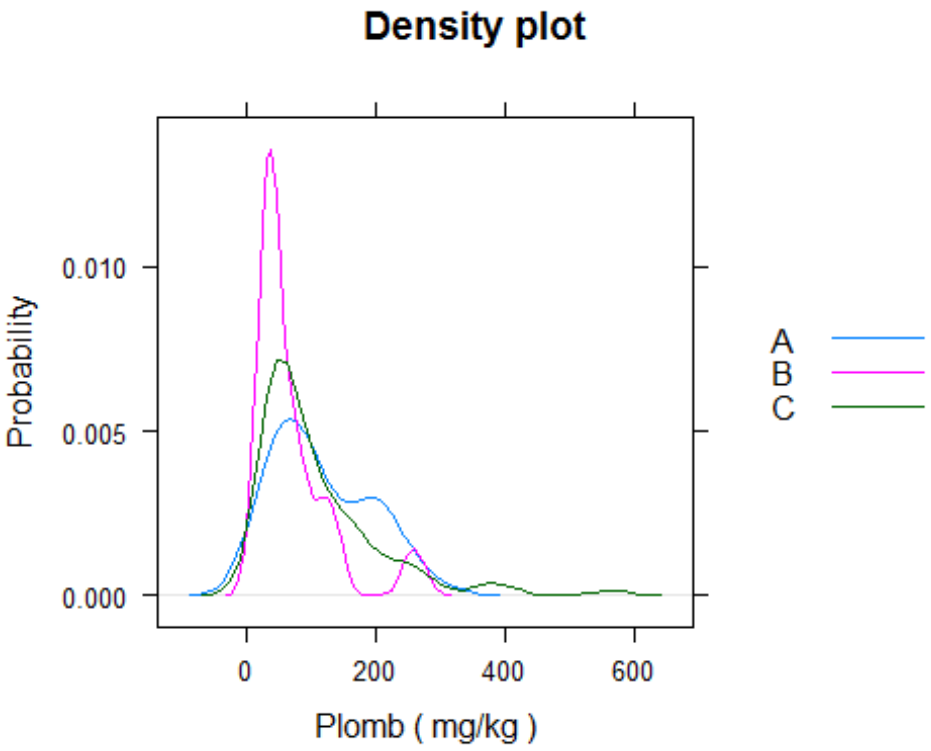
CODE :

```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("library")
# ********************************************************

######### Traitement classique pour plusieurs groupes #########

TABLE <- read.csv("NomFichier.csv", header = T, sep = ";") #Nom du fichier
csv à modifier

#Test
res = tapply(TABLE$VALEUR, TABLE$GROUPE, shapiro.test)

#Stockage du résultat (arrondi à 3 chiffres après la virgule)
GroupeA = c(round(res$A$statistic, 3), round(res$A$p.value, 3))
GroupeB = c(round(res$B$statistic, 3), round(res$B$p.value, 3))
GroupeC = c(round(res$C$statistic, 3), round(res$C$p.value, 3))

#Affichage dans la console
matrice <- matrix(c(GroupeA, GroupeB, GroupeC), nrow=3, ncol=2, byrow=T)
rownames(matrice) <- c("Agglomération A :", "Agglomération B :", "Agglomé
ration C :") #Nom à modifier
colnames(matrice) <- c("W","  p-value")
matrice

#Paramètres modifiables du graphique
title_plot = "Density plot"
label_x = "Plomb"
label_y = "Probability"
unite = "mg/kg"
```

```
#Tracé du graphique
densityplot(x = TABLE$VALEUR, groups = TABLE$GROUPE,
            main = title_plot,
            xlab = paste(label_x,"(",unite,")"),
            ylab = label_y,
            #Mise en forme modifiable
            # xlim = c(-100,700), #Echelle de l'axe des abscisses
            # ylim = c(-0.001,0.015), #Echelle de l'axe des ordonnées
            plot.points = TRUE,
            ref = TRUE,
            auto.key = list(space = "right"))
```

SORTIE :

```
##                         W   p-value
## Agglomération A : 0.927     0.042
## Agglomération B : 0.739     0.000
## Agglomération C : 0.789     0.000
```



Traitement classique pour plusieurs groupes

ENTREE :

GROUPE	VALEUR
A	0,1
A	0,5
A	0,9
…	…
B	0,2
B	1,5
B	2,3
…	…
C	4 ,8
C	5
C	5,1

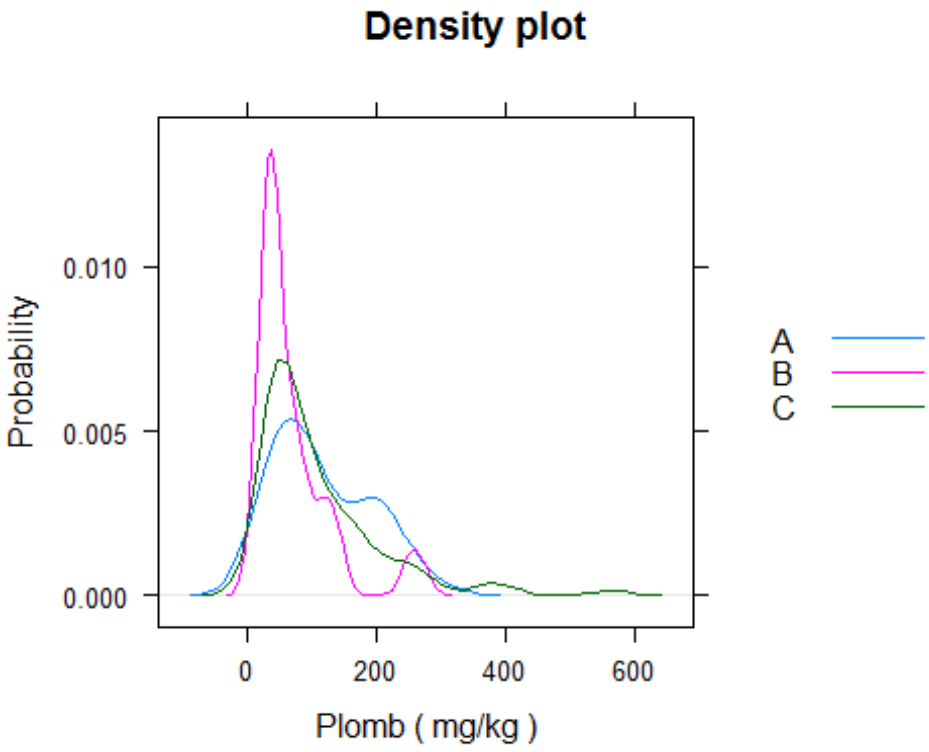
CODE :

```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("library")
# ********************************************************

######### Traitement classique pour plusieurs groupes #########

TABLE <- read.csv("NomFichier.csv", header = T, sep = ";") #Nom du fichier c
sv à modifier

#Test
res = tapply(TABLE$VALEUR, TABLE$GROUPE, shapiro.test)

#Stockage du résultat (arrondi à 3 chiffres après la virgule)
GroupeA = c(round(res$A$statistic, 3), round(res$A$p.value, 3))
GroupeB = c(round(res$B$statistic, 3), round(res$B$p.value, 3))
GroupeC = c(round(res$C$statistic, 3), round(res$C$p.value, 3))

#Affichage dans la console
matrice <- matrix(c(GroupeA, GroupeB, GroupeC), nrow=3, ncol=2, byrow=T)
rownames(matrice) <- c("Agglomération A :", "Agglomération B :", "Agglomera
tion C :") #Nom à modifier
colnames(matrice) <- c("W","  p-value")
matrice

#Paramètres modifiables du graphique
title_plot = "Density plot"
label_x = "Plomb"
label_y = "Probability"
unite = "mg/kg"
```

```
#Tracé du graphique
densityplot(x = TABLE$VALEUR, groups = TABLE$GROUPE,
            main = title_plot,
            xlab = paste(label_x,"(",unite,")"),
            ylab = label_y,
            #Mise en forme modifiable
            # xlim = c(-100,700), #Echelle de l'axe des abscisses
            # ylim = c(-0.001,0.015), #Echelle de l'axe des ordonnées
            plot.points = TRUE,
            ref = TRUE,
            auto.key = list(space = "right"))
```

SORTIE :

```
##                         W   p-value
## Agglomération A : 0.927     0.042
## Agglomération B : 0.739     0.000
## Agglomération C : 0.789     0.000
```



Fiche 2

Regression on Order Statistics (ROS) – Version robuste

Objectif	Calcul de statistiques descriptives d'une population censurée
Principe	À l'aide d'un PP-plot 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	[10]

Fiche 3

Analyse en Composantes Principales (ACP)

Objectif	Étude 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 de la population de données étudiée 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], [10]

ACP avec R Software

ACP des Données Brutes

ENTREE :

IDENTIFIANT	As	Cu	Cr	Pb	Zn	Ni	Cd	Hg	AGGLO
1	21	79	19	210	250	16	0.6	1.1	A
2	10	48	19	200	190	18	0.5	0.8	A
…	…	…	…	…	…	…	…	…	…
128	5.8	37	14	47	71	11	0.22	0.11	B
129	18	26	31	33	100	25	0.05	0.025	B
…	…	…	…	…	…	…	…	…	…
245	4.7	20	17	42	79	11	0.22	0.28	C
246	6.8	17	18	51	57	15	0.05	0.16	C
…	…	…	…	…	…	…	…	…	…

La valeur de la limite de quantification peut être utilisée pour remplacer les valeurs censurées.

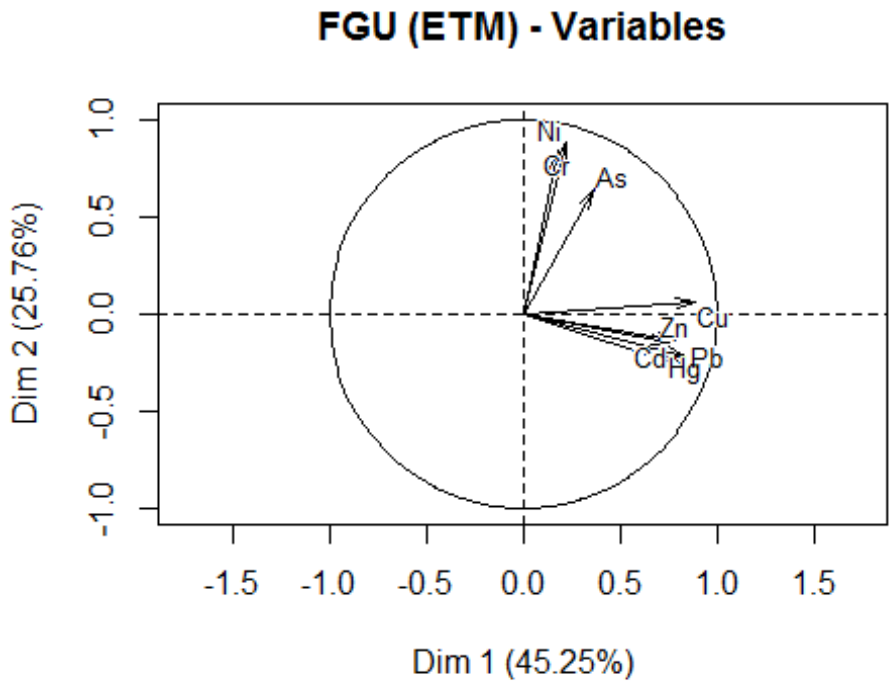
CODE :

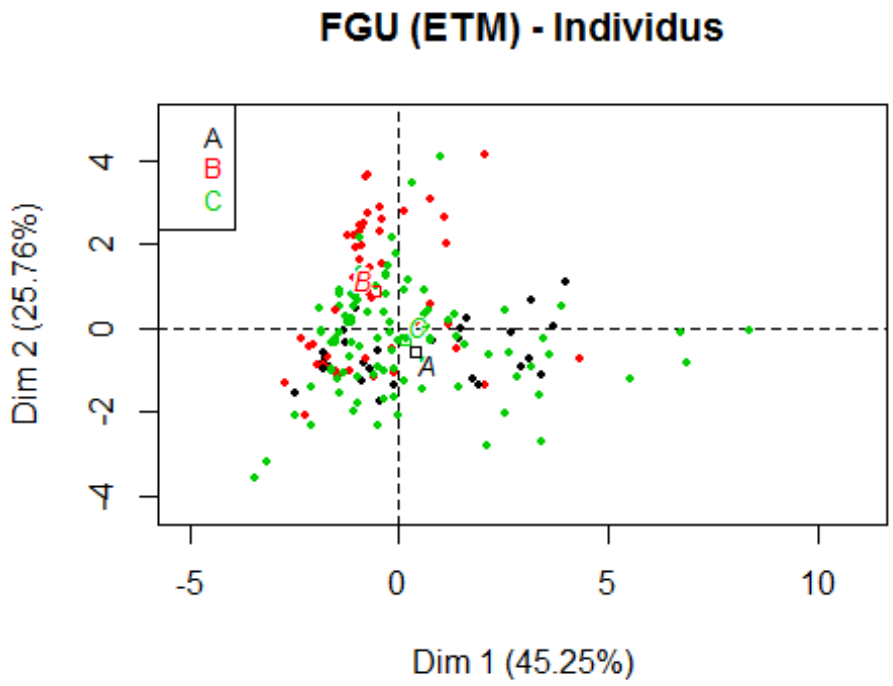
```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("FactoMineR")
# install.packages("zCompositions")
# install.packages("missMDA")
# ************************************************************

library("FactoMineR")
library("zCompositions")
library("missMDA")


ETM <- read.csv("fichier.csv",
                header=T, #Les colonnes du fichier contiennent une entête
                sep=";", #Le point-virgule est utilisé comme séparateur
                row.names = 1) #La colonne « 1 » contient les noms des
                               individus


########### ACP données brutes ###########
res.pca = PCA( ETM,
               quali.sup = 9, #Indique le numéro de colonne de la variable
                              qualitative
               ncp = 5, #Nombre de dimensions prises en compte
               graph = F) #Pas d'affichage du graphique construit par défaut

plot.PCA(res.pca,
         axes=c(1, 2),
```

```
choix="var", #Graphique des variables
         habillage=9, #Indique le numéro de colonne de la variable
                      qualitative
         shadow=T, #Paramètre d'affichage du texte pour éviter les
                   superpositions
         cex=0.8, #Taille du texte
         title = "FGU (ETM) - Variables")
```

```
plot.PCA(res.pca,
         axes=c(1, 2),
         choix="ind", #Graphique des individus
         habillage=9,
         select="cos2 0.999", #Sélection des points avec un cosinus carré
                              supérieur à 0.999 (utilisé ici pour éviter
                              l'affichage par défaut des identifiants à
                              côté des points)
         unselect = 0, #Gestion de la transparence des points
         shadow=T,
         cex=0.8,
         title = "FGU (ETM) - Individus")
```

SORTIE :




FGU (ETM) - Individus
A
B
C
Dim 2 (25.76%)
4
2
0
-2
-4
-5
0
5
10
Dim 1 (45.25%)

ACP des données log-transformées

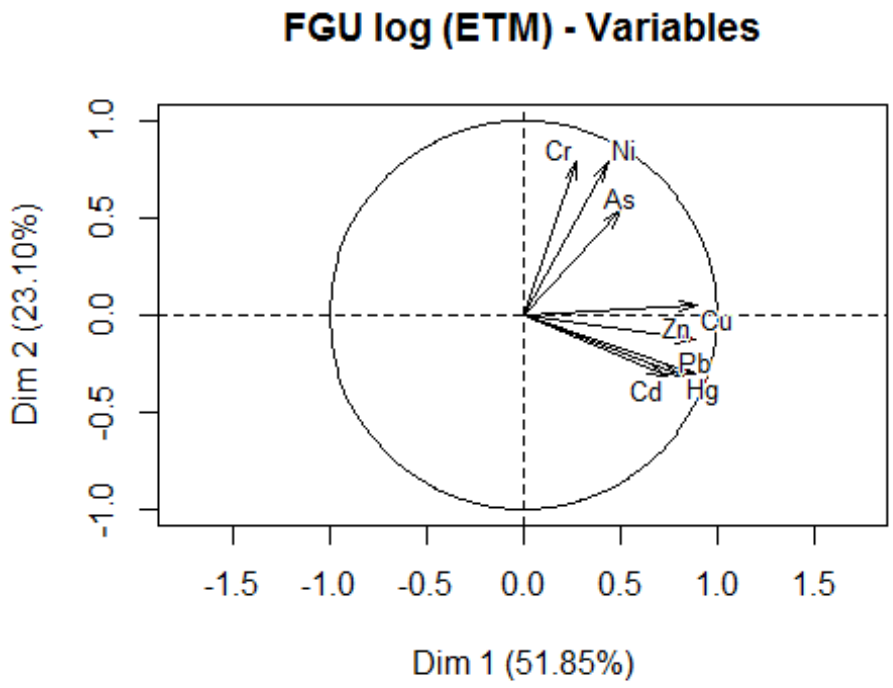
ENTREE :

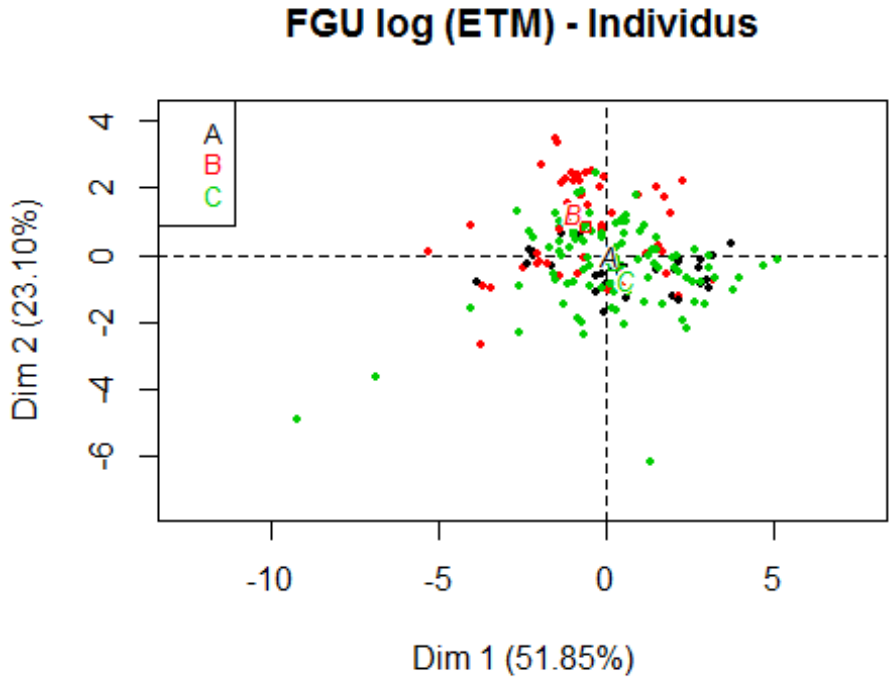
Voir « ENTREE » pour l'ACP des données brutes

CODE :

```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("FactoMineR")
# install.packages("zCompositions")
# install.packages("missMDA")
# ********************************************************

library("FactoMineR")
library("zCompositions")
library("missMDA")

ETM <- read.csv("fichier.csv",
                header=T, #Les colonnes du fichier contiennent une entête
                sep=";", #Le point-virgule est utilisé comme séparateur
                row.names = 1) #La colonne « 1 » contient les noms des
                               individus

############ ACP données log-transformées ############

##Application de la transformation logarithmique
ETM.log10 <- log(ETM[,1:8], base = 10)
ETM.log10["AGGLO"] <- ETM$AGGLO #Ajout colonne
colnames(ETM.log10)[which(names(ETM.log10) == "new.col")] <- "AGGLO" #Remp
lissage colonne ajoutée

res.pca = PCA( ETM.log10,
               quali.sup = 9,
               ncp = 5,
               graph = F)
plot.PCA(res.pca,
         axes=c(1, 2),
         choix="var",
         habillage=9,
         shadow=T,
         cex=0.8,
         title = "FGU log (ETM) - Variables")
plot.PCA(res.pca,
         axes=c(1, 2),
         choix="ind",
         habillage=9,
         select="cos2 0.999",
         unselect = 0,
         shadow=T, cex=0.8,
         title = "FGU log (ETM) - Individus")
```

SORTIE :


FGU log (ETM) - Variables
Cr
Ni
As
Zn
Cu
Pb
Cd
Hg
Dim 2 (23.10%)
Dim 1 (51.85%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
1.5


FGU log (ETM) - Individus
A
B
C
Dim 2 (23.10%)
Dim 1 (51.85%)
4
2
0
-2
-4
-6
-10
-5
5

Fiche 4

Histogramme

Objectif	Étude 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 graphique peut être réalisé à partir des données log-transformées.
Conditions d'application	Aucune condition particulière
Exemple	Figure 8 dans le corps du rapport
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]

Fiche 5

Densité

Objectif	Étude 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 graphique peut être réalisé à partir des données log-transformées.
Conditions d'application	Aucune condition particulière
Exemple	Figure 8 dans le corps du rapport
Avantages	- Plusieurs densités peuvent être superposées sur le même graphique afin de comparer la distribution de jeux 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]

Fiche 6

Dispersogramme unidimensionnel (ou *scatterplot*)

Objectif	Étude 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 8 dans le corps du rapport
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]

Fiche 7

Boxplot de Tukey (ou boîte à moustaches)

Objectif Étude 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[33] 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 :

$$Frontière\ supérieure = Q3 + 1{,}5 \times DI = 3{,}9$$

$$Frontière\ inférieure = Q1 - 1{,}5 \times DI = 0{,}7$$

[33] 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.

	Les frontières permettent de calculer les vibrisses (ou moustaches) de la boîte centrale : $$Vibrisse\ supérieure = \max(x[x \leq Frontière\ supérieure]) = 2{,}8$$ $$Moustache\ inférieure = \min(x[x \geq Frontière\ inférieure]) = 1{,}7$$ Les moustaches permettent d'apprécier la symétrie de la distribution. Voir Figure 8 dans le corps du rapport
Conditions d'application	Aucune condition particulière
Avantages	- Représentation graphique 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]

Boxplot avec R Software

Traitement pour un Groupe Unique

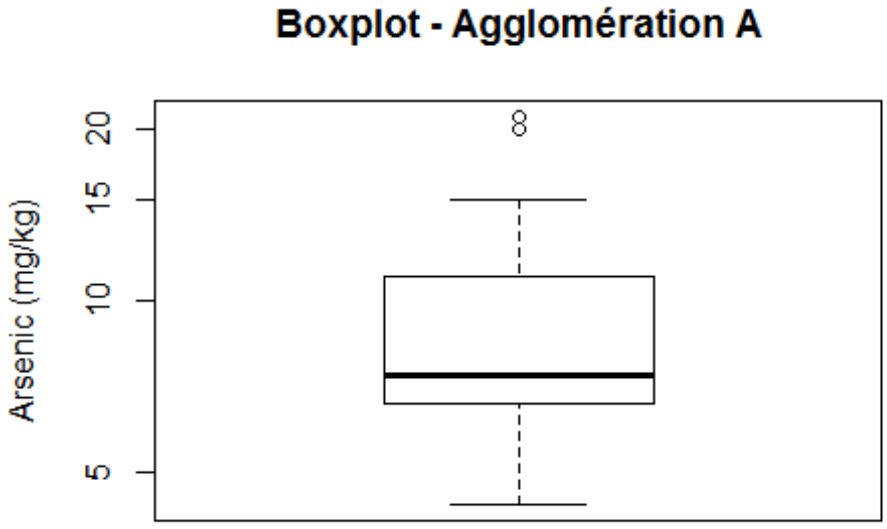
ENTREE :

IDENTIFIANT	VALEUR
1	0,1
2	0,5
3	0,9
4	1
…	…

CODE :

```
##Traitement pour un groupe unique
As <- read.csv("fichier.csv", header = T, sep=";")

boxplot(As$VALEUR,
        ylab = "Arsenic (mg/kg)",
        log = "y", ##Il faut comprendre ici l'axe « y »
        notch = FALSE,
        horizontal = FALSE)
title("Boxplot - Agglomération A")
```

SORTIE :



Traitement pour Plusieurs Groupes

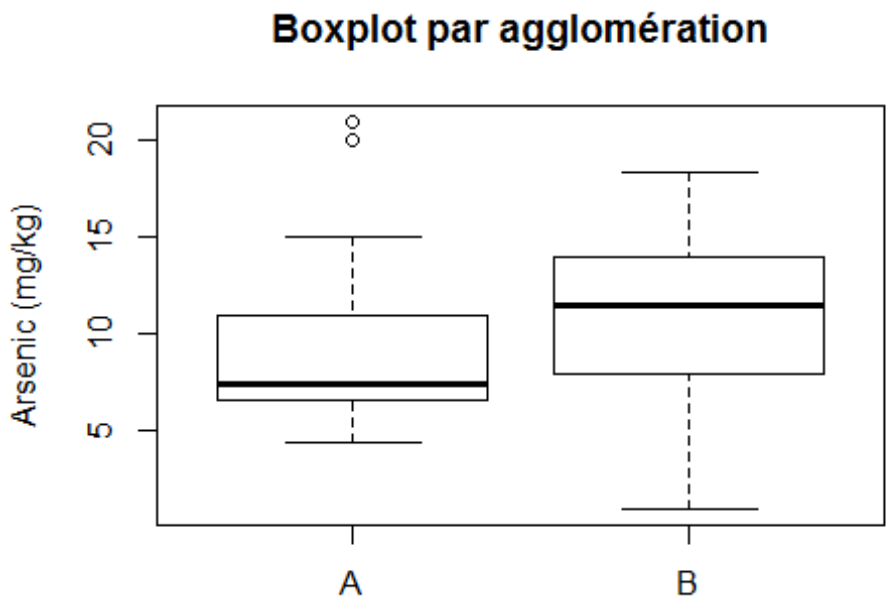
ENTREE :

IDENTIFIANT	VALEUR	GROUPE
1	0,1	A
2	0,5	A
3	0,9	A
4	1	A
…	…	
28	0,2	B
29	1,5	B
30	2,3	B
31	4,5	B
32	5,8	B
33	9	B

CODE :

```
##Traitement pour plusieurs groupes
AsGroup <- read.csv("fichier.csv", header = T, sep=";")

boxplot(VALEUR ~ AGGLO,
        data = AsGroup,
        ylab = "Arsenic (mg/kg)",
        log = "", ##Voir utilisation dans le cas d'un groupe unique
        notch = FALSE,
        horizontal = FALSE)
title("Boxplot par agglomération")
```

SORTIE :



Traitement pour Plusieurs Groupes avec Censure

ENTREE :

Fichier : *DataTbl.csv*

ID	COMP1	COMP4	COMP7	COMP12	COMP41	COMP58	COMP61	AGGLO
1	4.4	15.3	<0.1	<0.02	<0.002	<0.59	<0.5	A
2	5.1	17	<0.1	0.02	<0.002	<0.6	<0.53	A
3	5.3	24	<0.1	<0.02	<0.002	<0.87	<0.83	A
4	5.6	38	<0.1	<0.02	0.0031	0.97	<1.4	A
…	…	…	…	…	…	…	…	…
128	<1	12	<0.1	<0.01	<0.001	<0.34	<0.33	B
129	<1	17	<0.1	<0.01	<0.001	<0.45	<0.37	B
130	<1	20.5	<0.1	<0.01	<0.001	<0.49	<0.42	B
131	5.8	21	<0.1	<0.01	<0.001	<0.5	<0.45	B

' < ' : valeur censurée

Fichier : *Parameters.csv*

Parameter	Label	Groupby	Xlabel	Scale
COMP1	Arsenic	AGGLO	Agglo	Natural
COMP4	Plomb	AGGLO	Agglo	Natural
COMP7	Cadmium	AGGLO	Agglo	Natural
COMP12	Acenaph	AGGLO	Agglo	Natural
COMP41	PCB138	AGGLO	Agglo	Natural
COMP58	PCDD58	AGGLO	Agglo	Natural
COMP61	PCDF61	AGGLO	Agglo	Natural
COMP70	OCDF	AGGLO	Agglo	Natural

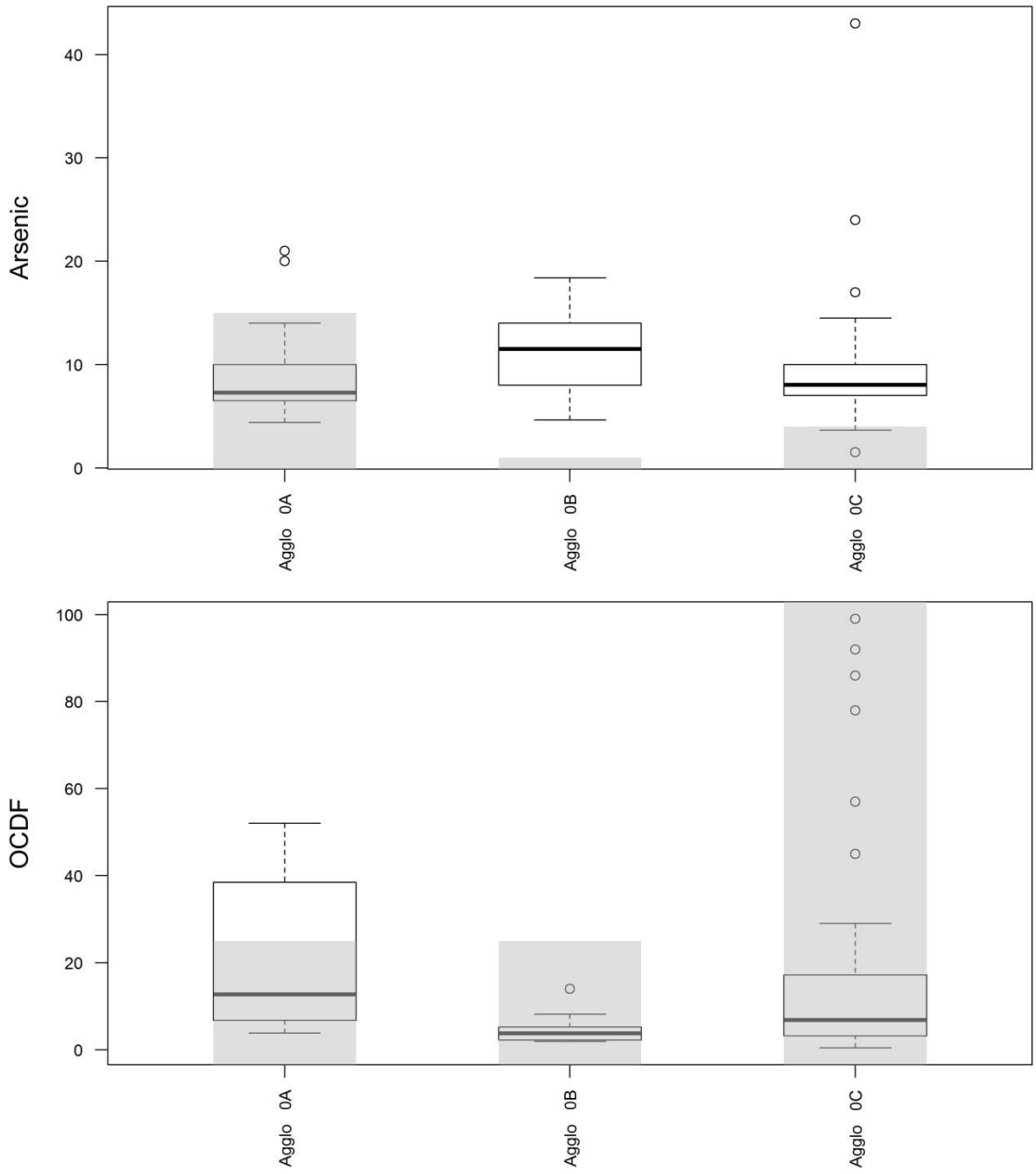
CODE :

(Entête du fichier *SummaryStatsROS-NEW.r* à remplir dans *…\30.Stage_LS\Données\Traitement_R\Boxplot\Boxplots 3 Agglo*)

```
#********* User updates for local use ****************************************
#
#  Entrer le chemin d'accès du dossier de travail (remplacer "\" par "/")
#
WD <- "//brgm.fr/eDOP/eDOP_PS14D3E027_D3E_POL_FGU_2_ADEME_14/e-DOP_DOCUMENTS_TRAVAIL/P2_DT_Realisation/A.RéalisationPpale/30.Stage_LS/Données/Traitement_R/Boxplots 3 Agglo/Selection"
#
#  Identifier le fichier des données et le fichier des paramètres
#
input_file <- "DataTbl.csv"
parm__file <- "Parameters.csv"
#
#  Entrer le nom de la colonne relative aux groupes de données (le site de prélèvement par exemple)
#
Groupby <- "AGGLO"
#
```

```
#  Entrer le nom de l'axe des abscisses qui sera utiliser pour les graphiques. Cel
ui-ci sera combiné avec le nom du groupe.
#  Exemple: "Agglo B"
#
Xlabel <- "Agglo"
#
#  Entrer le nombre de données non-censurées requises pour générer des estimations
par la méthode ROS.
#  Le minimum est 3, entrer un nombre supérieur si besoin.
#
nGreq  <- 3;
```

SORTIE : (exemples de l'arsenic et de l'OCDF)



Fiche 8

Fonction de répartition empirique – ECDF

Objectif	Étude 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 graphique peut être tracé avec une échelle logarithmique en abscisse.
Conditions d'application	Aucune condition particulière
Exemple	Figure 8 dans le corps du rapport
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]

ECDF avec R Software

Traitement pour un Groupe Unique sans Censure

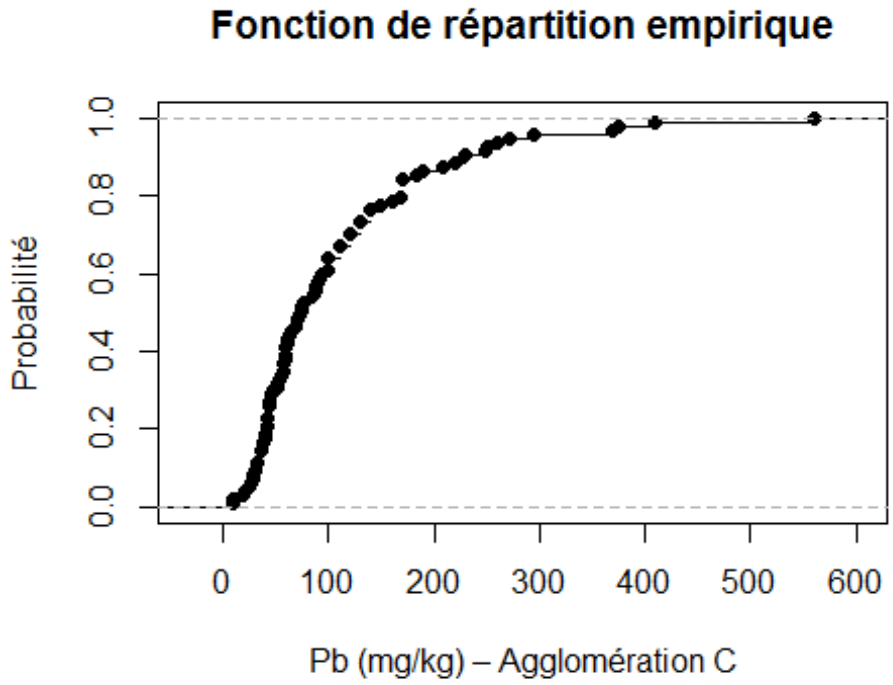
ENTREE :

IDENTIFIANT	VALEUR
1	0,1
2	0,5
3	0,9
4	1
…	…

CODE :

```
# ***** A ne compiler que lors de la première utilisation ****
# install.packages("NADA")
# install.packages("lattice")
# install.packages("latticeExtra")
# ********************************************************

############ Traitement classique ############
TABLE <- read.csv("PbCour.csv", header = T, sep = ";") #Nom du fichier csv
à modifier

res <- ecdf(TABLE$VALEUR) #"res" pour "résultat"

#Paramètres modifiables du graphique
title_plot = "ECDF"
label_x = "Agglomération C : Plomb"
label_y = "Probability"
unite = "mg/kg"

#Tracé du graphique
plot(res,
     main = title_plot,
     xlab = paste(label_x," (",unite,")"),
     ylab = label_y,
     #Mise en forme modifiable
     col = "black", #Couleur
     lty = 1, #Style de trait
     verticals = TRUE, #Traits verticaux
     do.points = TRUE, #Affichage des points
     cex = 0.7) #Taille
```

SORTIE :


Fonction de répartition empirique
Probabilité
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600
Pb (mg/kg) – Agglomération C

Traitement pour un Groupe Unique avec Censure (<50%)

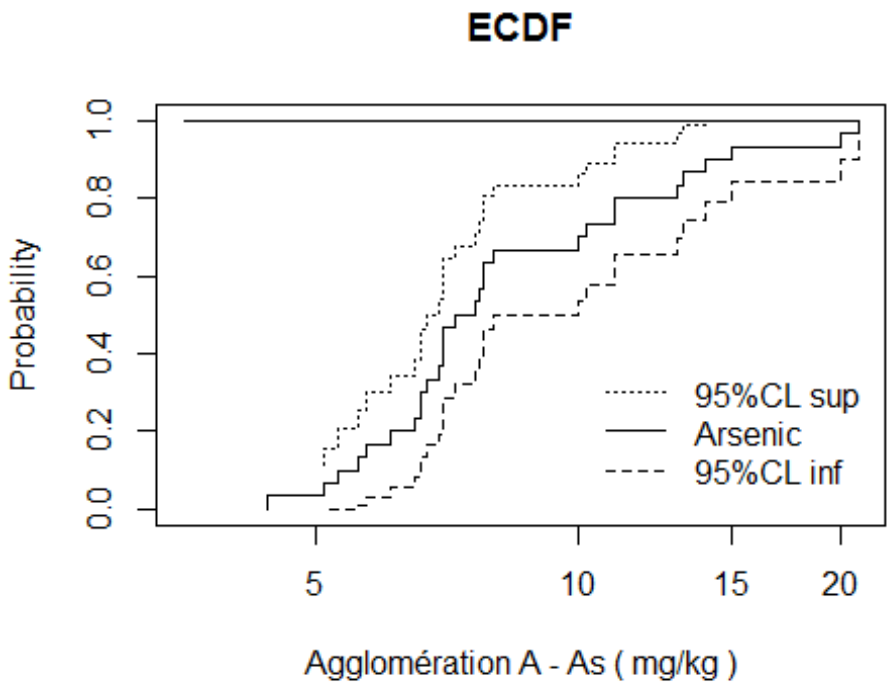
ENTREE :

IDENTIFIANT	VALEUR	CODE_LQ
1	0,1	TRUE
2	0,5	TRUE
3	0,9	FALSE
4	1	FALSE
…	…	…

CODE :

```
########### Traitement avec censure <50% ###########
library("NADA")

TABLE <- read.csv("AsMar.csv", header = T, sep = ";") #Nom du fichier csv à m
odifier

res = cenfit(TABLE$VALEUR, TABLE$CODE_LQ) #"res" pour "résultat"

#Paramètres modifiables du graphique
title_plot = "ECDF"
label_x = "Agglomération A : As"
label_y = "Probability"
unite = "mg/kg"

#Tracé du graphique
plot(res,
     main = title_plot,
     xlab = paste(label_x,"(",unite,")"),
     ylab = label_y)

#Paramètres modifiables de légende
text_leg <- c("95%CL sup","Arsenic","95%CL inf")
x_leg <- 10 #Position de la légende sur l'axe x
y_leg <- 0.4 #Position de la légende sur l'axe y
col_leg <- c("black","black","black") #Couleur
lty_leg <- c(3, 1, 2) #Style de trait

#Tracé de la légende
legend(x=x_leg, y=y_leg, legend=text_leg, col=col_leg,lty=lty_leg,
       #Mise en forme modifiable
       bty="n", #Encadré ("o" pour afficher)
       cex=1) #Taille
```

SORTIE :


ECDF
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
95%CL sup
Arsenic
95%CL inf
Agglomération A - As (mg/kg)

Traitement pour Plusieurs Groupes sans Censure

ENTREE :

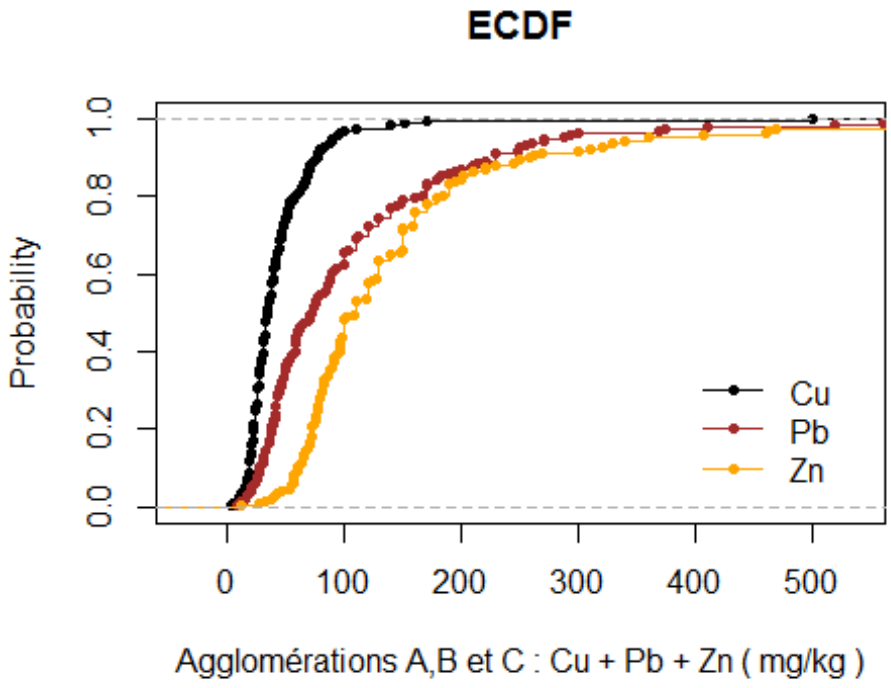
IDENTIFIANT	Cu	Pb	Zn
1	0,1	0,5	0,2
2	0,5	2,4	0,6
3	0,9	3	1,6
4	1	3,1	2,9
…	…	…	…

CODE :

```
########### Traitement classique pour plusieurs groupes ###########
library(lattice)
library(latticeExtra)

TABLE <- read.csv("CuPbZnCour.csv", header = T, sep = ";") #Nom du fichier c
sv à modifier

res1 <- ecdf(TABLE$Cu) #Nom des colonnes à modifier
res2 <- ecdf(TABLE$Pb) #Nom des colonnes à modifier
res3 <- ecdf(TABLE$Zn) #Nom des colonnes à modifier
#res4 <- ecdf(TABLE$ColumnName) ...à modifier pour plus/moins de groupes

#Paramètres modifiables du graphique
title_plot = "ECDF"
label_x = "Agglomérations A,B et C : Cu + Pb + Zn" #Ajouter les substances su
pplémentaires ici
label_y = "Probability"
unite = "mg/kg"

#Tracé du graphique
plot(res1,
     main = title_plot,
     xlab = paste(label_x,"(",unite,")"),
     ylab = label_y,
     #Mise en forme modifiable
     col = "black", #Couleur
     verticals = TRUE, #Traits verticaux
     do.points = TRUE, #Affichage des points
     cex = 0.7) #Taille
plot(res2, add = TRUE, col = "brown", verticals = TRUE, do.points = TRUE, c
ex = 0.7)
plot(res3, add = TRUE, col = "orange", verticals = TRUE, do.points = TRUE,
cex = 0.7)
#plot(res4, add = TRUE, col = "blue", verticals = TRUE, do.points = TRUE, cex = 0.
7)
```

```
#Paramètres modifiables de légende
text_leg <- c("Cu", "Pb", "Zn") #Ajouter les substances supplémentaires ici
col_leg <- c("black", "brown", "orange") #Ajouter les couleurs supplémentaires
ici
x_leg <- 400 #Position de la légende sur l'axe x
y_leg <- 0.4 #Position de la légende sur l'axe y
pch_leg <- c(20, 20, 20) #Ajouter les symboles supplémentaires ici
lty_leg <- c(1,1,1) #Ajouter les styles de traits supplémentaires ici

#Tracé de la légende
legend(x=x_leg, y=y_leg, legend=text_leg, col=col_leg, pch=pch_leg, lty=lty
_leg,
       #Mise en forme modifiable
       bty="n", #Encadré ("o" pour afficher)
       cex=1) #Taille
```

SORTIE :



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]

Fiche 10

MEAN ± 2.SD

Objectif

Détermination d'une valeur seuil pour caractériser le fond géochimique d'une population statistique

Principe

Les seuils sont déterminés par le calcul suivant :

$$MOYENNE \pm 2 \times ECART\ TYPE$$

Ainsi toute valeur inférieure à la $MOY - 2 \times ET$ ou supérieure à la $MOY + 2 \times ET$ est considérée comme n'appartenant pas au fond géochimique. Pour une distribution normale, 4,6 % des valeurs sont considérées comme valeurs extrêmes (2,3 % parmi les valeurs faibles/hautes). Les 95,4 % restant représentent la gamme de variation du fond géochimique du jeu de données.

Cette méthode est fortement basée sur la symétrie de la distribution des données. Cependant, en géochimie environnementale, il est fréquent que les distributions soient asymétriques positives. Une transformation logarithmique est donc recommandée avant de procéder à la détermination d'un fond géochimique par cette méthode. Une transformation inverse sera alors nécessaire afin d'interpréter le résultat dans l'unité d'origine.

$On\ note\ x_i\ le\ ième\ individu\ d'une\ population\ d'effectif\ n,$

$$Moyenne\ arithmétique = \bar{x} = \sum \frac{x_i}{n}$$

$$Ecart\ type = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Exemple

Cuivre – Agglomération B

Effectif	MEAN	SD	MEAN - 2.SD		MEAN + 2.SD	
			brut	*log*	*brut*	*log*
48	36,2	18,7	-1.2	11	74	89

Conditions d'application

- Distribution normale
- Effectif ≥ 30

Avantages

Simplicité et rapidité

Inconvénients	Précision proportionelle à la zone d'étude et à la taille du jeu de données Définition géochimique Les outliers sont considérés comme pouvant apparaître uniquement aux extrémités de la distribution Transformation nécessaire
Bibliographie	[5], [16]

Fiche 11

Centiles

Objectif

Détermination de valeurs seuils pour caractériser le fond géochimique d'une population statistique.

Principe

Le principe de la méthode des centiles repose sur l'hypothèse qu'il existe toujours, au sein du jeu de données, un certain pourcentage d'outliers.

Après avoir trié le jeu de données dans l'ordre croissant, les 2 %, 2,5 % ou 5 % supérieurs des données sont définis arbitrairement par le statisticien comme les outliers de la population étudiée. Il en résulte une valeur de fond géochimique, respectivement le 98ème, le 97,5^{ème} ou le 95ème centile.

Pour obtenir une gamme de valeurs, la même opération est effectuée pour les valeurs faibles. Par exemple, pour un pourcentage de 2%, on obtient le 2ème et le 98ème centile comme valeurs basse et haute du fond géochimique.

Cuivre - Agglomération B

Effectif	2 %	98 %	5 %	95 %
48	10	88	14	74

Conditions d'application

- Effectif ≥ 30

Avantages

Simplicité et rapidité

Aucune transformation nécessaire

Aucune distribution hypothétique nécessaire

Inconvénients

Le pourcentage de valeurs aberrantes peut varier en fonction des caractéristiques du jeu de données (taille de la zone d'étude, effectif,…)

Hypothèse sous-jacente : il existe toujours des outliers dans un jeu de données

Bibliographie

[5] ; [3]

Fiche 12

MED ± 2.MAD

Objectif

Détermination de valeurs seuils pour caractériser le fond géochimique d'une population statistique.

Principe

Cette méthode est une version robuste de la méthode « MEAN ± 2.SD ». Les estimateurs classiques de tendance centrale, moyenne arithmétique et écart-type, sont remplacés par leurs équivalents robustes, respectivement la médiane et l'écart absolu moyen.

$On\ note\ x_i\ le\ ième\ individu\ d'une\ population\ d'effectif\ n\ triée\ dans\ l'ordre\ croissant,$

$$Médiane = M = \begin{cases} si\ n\ est\ impair\ alors\ n = 2p+1, & x_{p+1} \\ si\ n\ est\ pair\ alors\ n = 2p, & \frac{1}{2}(x_p + x_{p+1}) \end{cases}$$

$$Ecart\ absolu\ moyen = MAD = 1{,}4826 \times Médiane(|x_i - M|)$$

La constante « 1,4826 » est un facteur d'échelle dépendant de la distribution de la population étudiée. Ici, il correspond au facteur d'échelle employé pour une distribution normale.

Cette méthode est fortement basée sur la symétrie de la distribution des données. Cependant, en géochimie environnementale, il est fréquent que les distributions soient asymétriques positives. Une transformation logarithmique est donc recommandée avant de procéder à la détermination d'un fond géochimique par cette méthode. Une transformation inverse sera alors nécessaire afin d'interpréter le résultat dans l'unité d'origine.

Exemple

Cuivre – Agglomération B

Effectif	MEDIAN	MAD	MEDIAN - 2.MAD		MEDIAN + 2.MAD	
			brut	*log*	*brut*	*log*
48	33	13.34	6.3	14	60	80

Conditions d'application

- Distribution normale
- Effectif ≥ 30
- Pourcentage de valeurs extrêmes < 50 %

Avantages

Simplicité et rapidité

Inconvénients	Précision proportionelle à la zone d'étude et à la taille du jeu de données Définition géochimique Les outliers sont considérés comme pouvant apparaître uniquement aux extrémités de la distribution Transformation nécessaire
Bibliographie	[5], [16]

Fiche 13

Vibrisse interne supérieure du boxplot de Tukey

Objectif	Détermination d'une valeur seuil pour caractériser le fond géochimique d'une population statistique.
Principe	La vibrisse interne supérieure du boxplot de Tukey (voir fiche 7) est définie pour identifier automatiquement les valeurs extrêmes hautes. $$Vibrisse\ supérieure = \max(x[x \leq Q3 + 1{,}5 \times DI])$$ Cette méthode est fortement basée sur la symétrie de la distribution des données. Cependant, en géochimie environnementale, il est fréquent que les distributions soient asymétriques positives. Une transformation logarithmique est donc recommandée avant de procéder à la détermination d'un fond géochimique par cette méthode. Une transformation inverse sera alors nécessaire afin d'interpréter le résultat dans l'unité d'origine.
Conditions d'application	Distribution normale - Effectif ≥ 30
Avantages	Méthode robuste Par construction, il est impossible d'obtenir des valeurs négatives (à la différence des autres méthodes).
Inconvénients	Transformation nécessaire pour approcher une distribution symétrique.
Bibliographie	[3] ; [5]

Fiche 14

Tangente à l'ECDF

Objectif	Détermination d'une valeur seuil pour caractériser le fond géochimique d'une population statistique.
Principe	La fonction de répartition empirique (voir Fiche 8) ou *ECDF* est un outil puissant lorsqu'il s'agit de visualiser les déviations d'une distribution par rapport à un modèle supposé. La répartition des données est clairement visible et les valeurs extrêmes sont détectables facilement sous forme de points isolés. En combinant les informations visibles sur l'*ECDF*, il devient possible de tracer une droite épousant au mieux la partie linéaire de la courbe normalement sigmoïde. La valeur seuil supérieure (respectivement inférieure) correspond à l'abscisse de la première valeur s'écartant de la droite tracée.
Conditions d'application	Distribution normale - Effectif ≥ 150
Avantages	Visibilité de chaque point de mesure (en particuliers des outliers)
Inconvénients	Effectif élevé requis
Bibliographie	[3]; [5]

Fiche 15

Concentration Area Plot (CA-plot)

Objectif	Détermination d'une valeur seuil pour caractériser le fond géochimique d'une population statistique.
Principe	L'intérêt du CA-plot réside dans son approche « fractale » de la structure des données. Après une interpolation triangulaire, on trace le pourcentage de valeurs interpolées supérieures à une valeur donnée. On peut donc observer la relation entre le pourcentage de la zone d'étude présentant une valeur donnée et cette même valeur. Par définition, les valeurs appartenant au fond géochimique seront les plus fréquentes et représenteront un pourcentage élevé de la zone d'investigation. À l'inverse, les zones contenant des valeurs extrêmes (hautes ou basses) représenteront un pourcentage faible sur le CA-plot. L'étude de la courbe ainsi réalisée permet d'identifier les différentes distributions fractales présentes au sein du jeu de données.
Conditions d'application	Distribution normale Distribution spatiale des points relativement homogène - Effectif ≥ 30
Avantages	Prise en compte de la répartition spatiale des données Augmentation de l'effectif par interpolation
Inconvénients	Méthode de détermination graphique (précision limitée) Transformation nécessaire pour approcher une distribution symétrique Méthode d'interpolation influente sur le résultat
Bibliographie	[5]

Annexe 2

Tableaux récapitulatifs des paramètres statistiques utilisés et résultats du FPGA

	Effectif	%LQ	Min	Q1	Med	Moy	Q3	Max	s	MAD	DIQ	p-value
As	175	4%	1	7	8.72	9.5	11	43	4.5	1.92	4	< 0,0001
AsA	30	10%	4.4	6.6	7.4	9.0	11	21	4.2	1.5	4.4	0.0001
AsB	48	6%	1	8	11.5	11.0	14	18.4	4.4	3.2	6	0.1262
AsC	97	1%	1.5	7	8.04	8.8	10	43	4.5	1.24	3	< 0,0001
Cu	175	0%	2.6	24	34	40.7	49	170	24.6	11	25	< 0,0001
CuA	30	0%	10.2	25	37.2	42.7	61	90	20.4	13.2	36.5	0.0747
CuB	48	0%	7.5	24	33	36.2	42.5	95	18.7	9	18.5	0.0007
CuC	97	0%	2.6	24	33	42.4	51	170	28.1	12	27	< 0,0001
Cd	175	49%	0.05	0.21	0.5	0.46	0.5	2.9	0.32	0.1	0.29	< 0,0001
CdA	30	30%	0.1	0.19	0.47	0.46	0.67	1.37	0.33	0.27	0.48	0.0066
CdB	48	52%	0.1	0.19	0.24	0.31	0.4	0.9	0.18	0.14	0.21	0.0002
CdC	97	54%	0.05	0.5	0.5	0.54	0.52	2.9	0.34	0	0.02	< 0,0001
Pyr	175	17%	0.02	0.05	0.15	0.33	0.3	8.9	0.78	0.1	0.25	< 0,0001
PyrA	30	10%	0.02	0.09	0.21	0.32	0.326	1.81	0.39	0.12	0.235	< 0,0001
PyrB	48	48%	0.02	0.05	0.05	0.11	0.125	0.54	0.11	0.02	0.075	< 0,0001
PyrC	97	4%	0.02	0.09	0.18	0.44	0.43	8.9	1.01	0.12	0.34	< 0,0001

Minimum (Min), 1er Quartile (Q1), Médiane (Med), Moyenne arithmétique (Moy), 3ème Quartile (Q3), Maximum (Max), Ecart-type (s), Median Absolute Deviation (MAD), Distance Inter Quartile (DIQ), p-value (Shapiro-Wilk)

Tableau 10 : Tableau récapitulatif des paramètres statistiques descriptifs des prétraitement recommandés. Réalisé avec l'arsenic (As), le cuivre (Cu), le cadmium (Cd) et le Pyrène (Pyr) en mg/kg dans les agglomérations A, B, C et la combinaison des jeux de données des trois agglomérations.

	Vibrisse interne inférieure		Vibrisse interne supérieure		MED - 2.MAD		MED + 2.MAD	
	brut	log	brut	log	brut	log	brut	log
Cu	2.6	9.8	82	110	1.4	12	67	96
CuA	10.2	10.2	90	90	-1.9	10	76	132
CuB	7.5	12.6	63	95	6.3	14	60	80
CuC	2.6	11	89	152	-2.6	11	69	100
As	1.0	3.7	17	21	3.0	4.4	14	17
AsA	4.4	4.4	15	21	3.0	3.8	12	14
AsB	1.0	5.8	18	18	2.0	4.9	21	27
AsC	3.7	4.3	15	17	4.4	4.9	12	13
Cd	0.05	0.10	0.90	1.37	0.20	0.26	0.80	1.0
CdA	0.10	0.10	1.37	1.37	-0.32	0.12	1.26	1.8
CdB	0.10	0.10	0.70	0.9	-0.16	0.05	0.64	1.1
CdC	0.48	0.48	0.52	0.52	0.50	0.50	0.50	0.5
Pyr	0.020	0.020	0.640	3.900	-0.147	0.014	0.447	1.637
PyrA	0.020	0.020	0.499	1.811	-0.138	0.046	0.558	0.954
PyrB	0.020	0.020	0.220	0.310	-0.009	0.018	0.109	0.141
PyrC	0.020	0.020	0.850	3.900	-0.164	0.018	0.524	1.782

Brut : valeur calculée à partir des données brutes ; Log : valeur calculée à partir des données log-transformées puis ramenée à l'échelle originelle

Tableau 11 : Résultats des calculs du FPGA avec l'arsenic (As), le cuivre (Cu), le cadmium (Cd) et le Pyrène (Pyr) en mg/kg dans les agglomérations A, B, C et la combinaison des jeux de données des trois agglomérations

Annexe 3

Projections spatiales des résultats obtenus avec différentes méthodes de détermination du FPGA

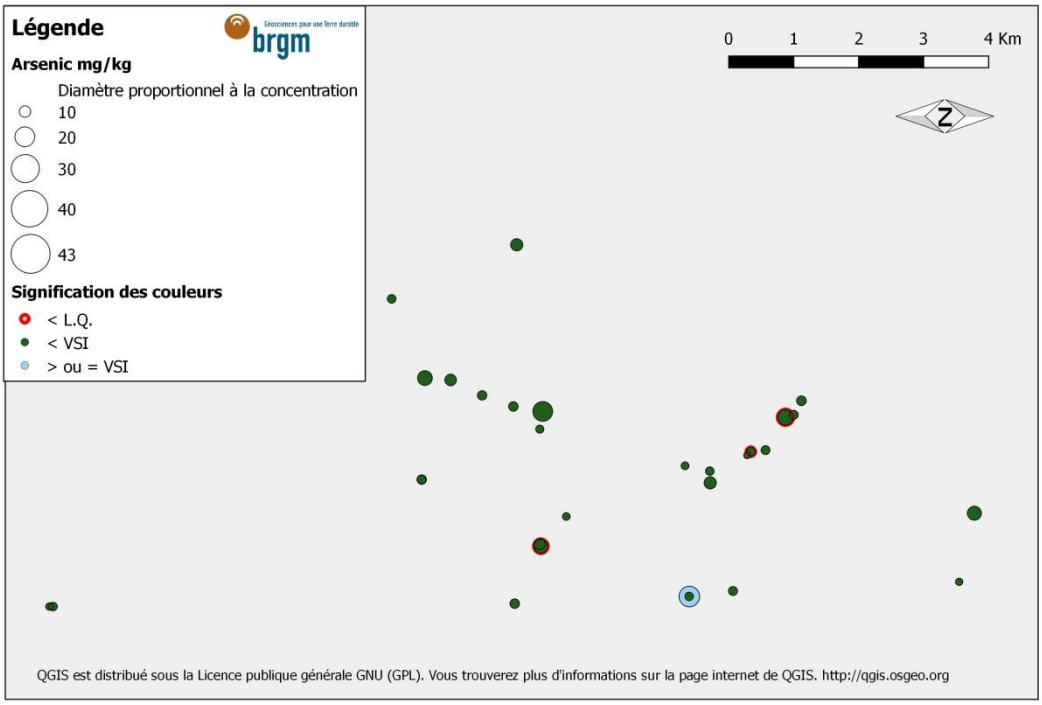


Figure 1 : Répartition spatiale des échantillons d'arsenic de l'agglomération A en fonction du seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

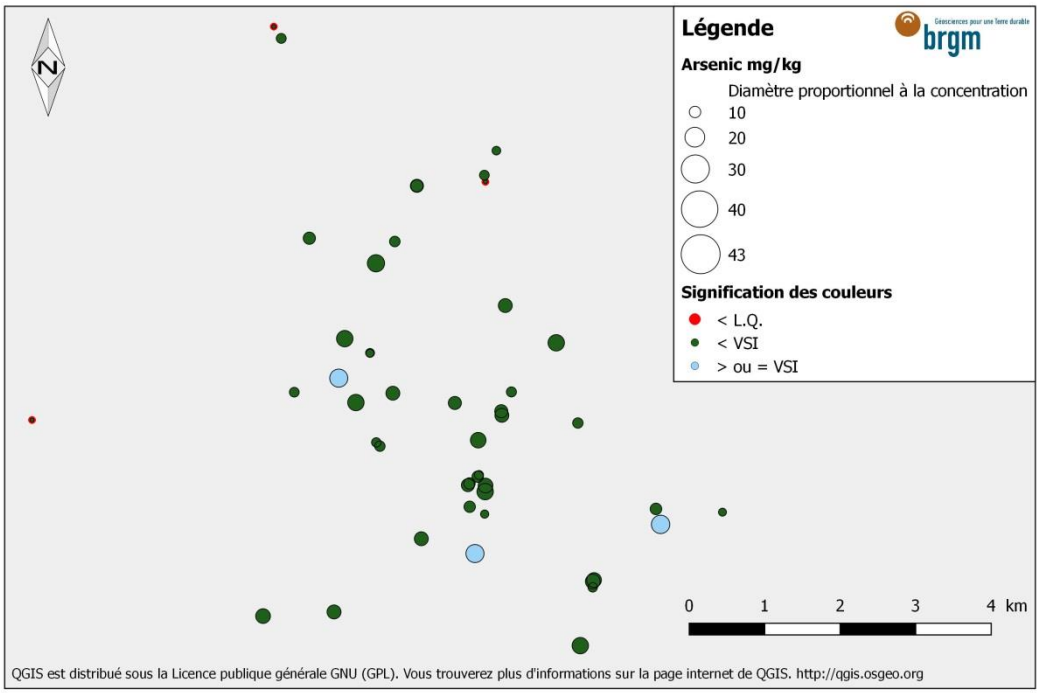


Figure 2 : Répartition spatiale des échantillons d'arsenic de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

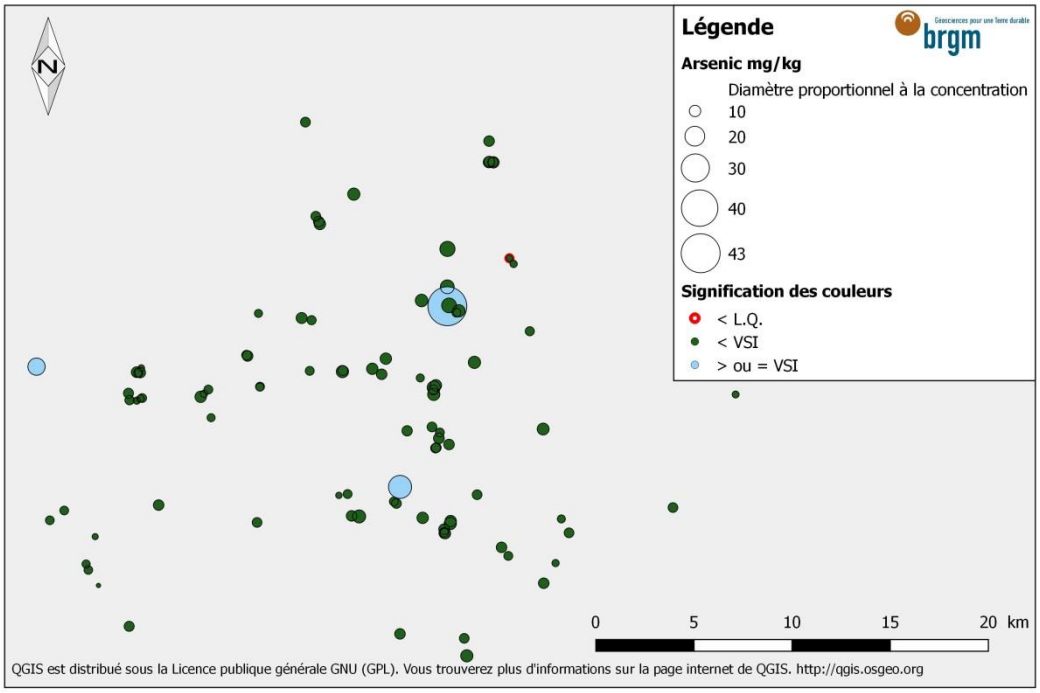


Figure 3 : Répartition spatiale des échantillons d'arsenic de l'agglomération C en fonction du seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

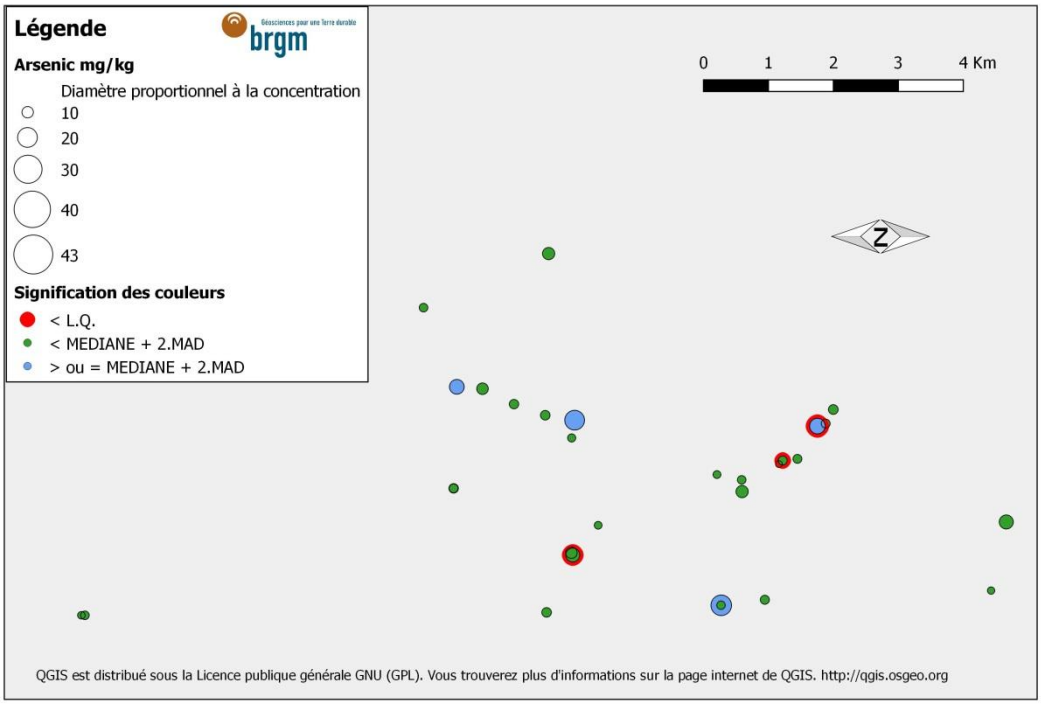


Figure 4 : Répartition spatiale des échantillons d'arsenic de l'agglomération A en fonction du seuil du FPGA calculé avec la méthode MED +2 MAD.

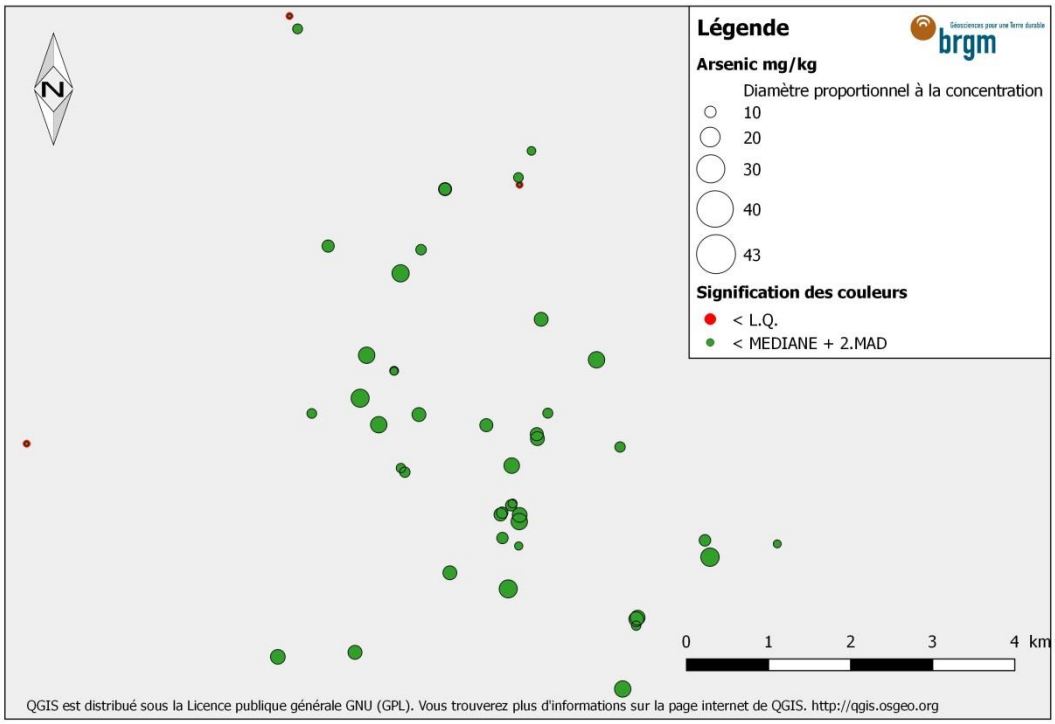


Figure 5 : Répartition spatiale des échantillons d'arsenic de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

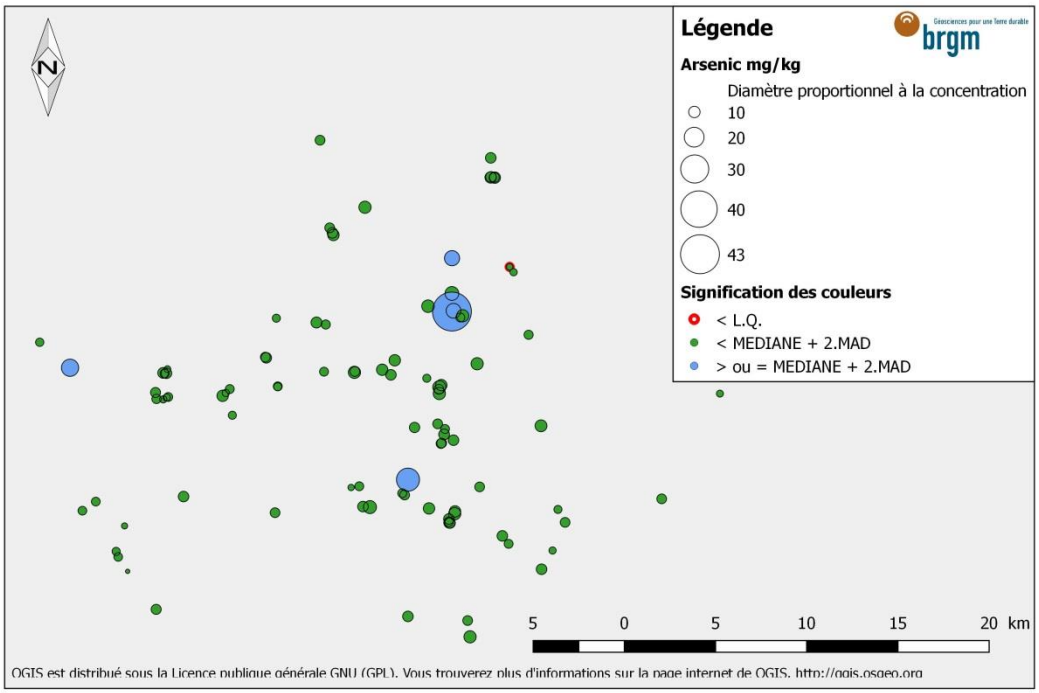


Figure 6 : Répartition spatiale des échantillons d'arsenic de l'agglomération C en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

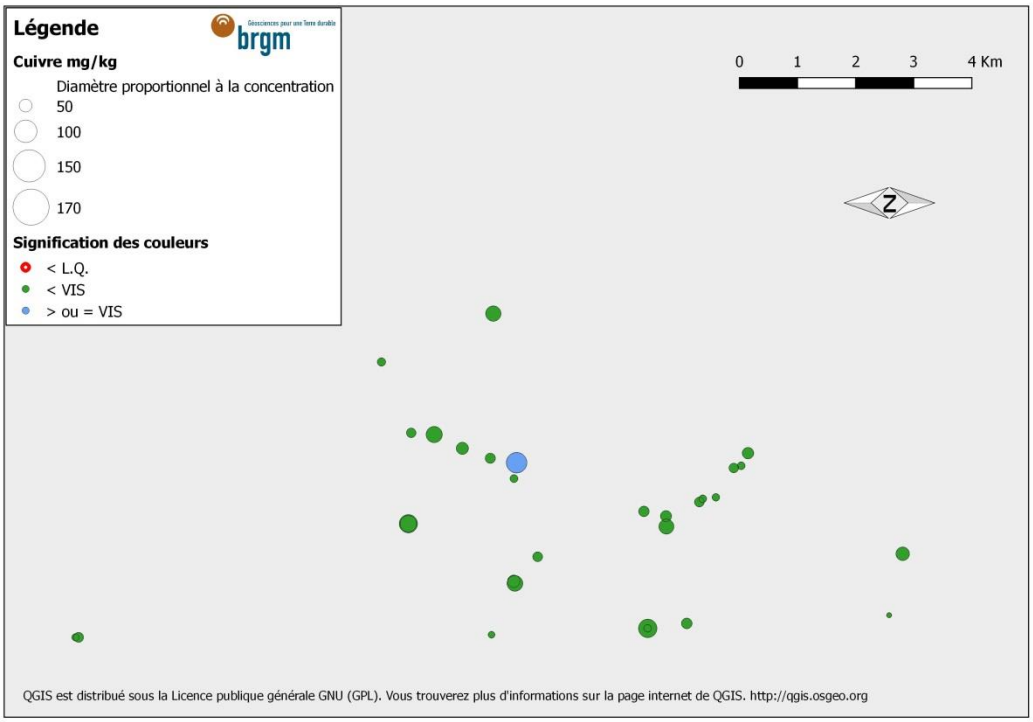


Figure 7 : Répartition spatiale des échantillons de cuivre de l'agglomération A en fonction du seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

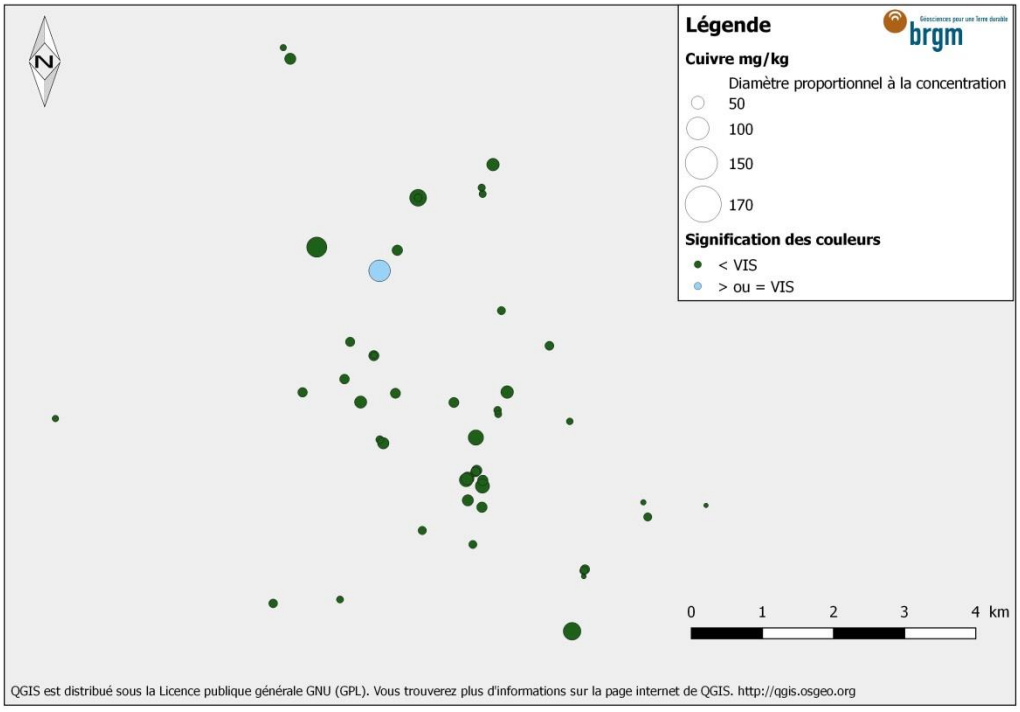


Figure 8 : Répartition spatiale des échantillons de cuivre de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

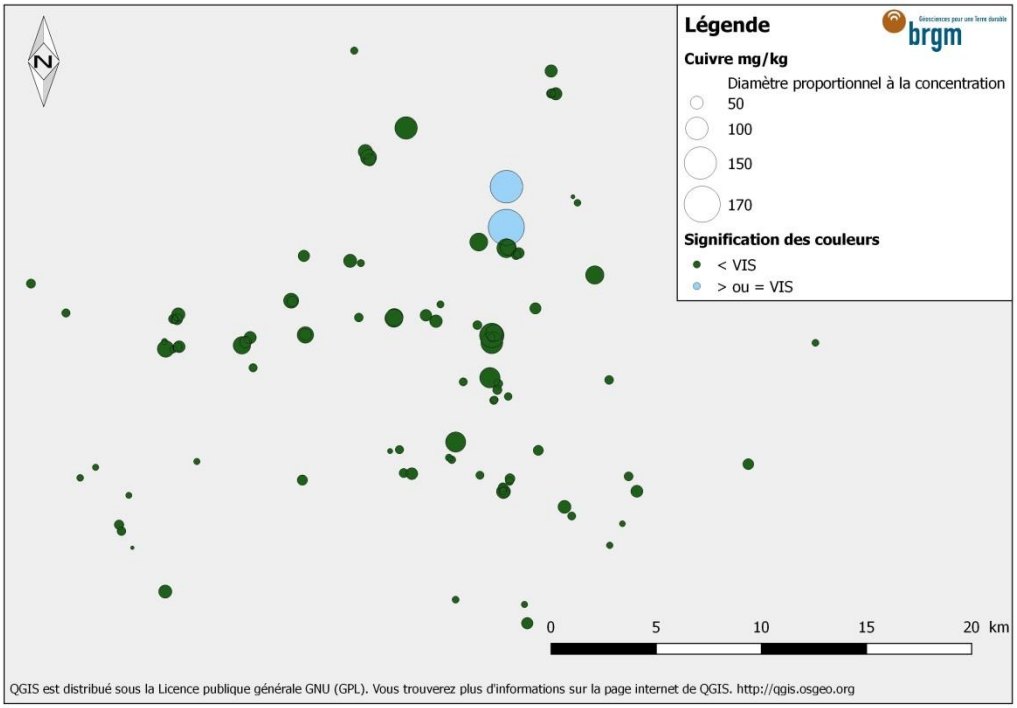


Figure 9 : Répartition spatiale des échantillons de cuivre de l'agglomération C en fonctiondu seuil du FPGA calculé avec la méthode de la Vibrisse Supérieure Interne (V.S.I.).

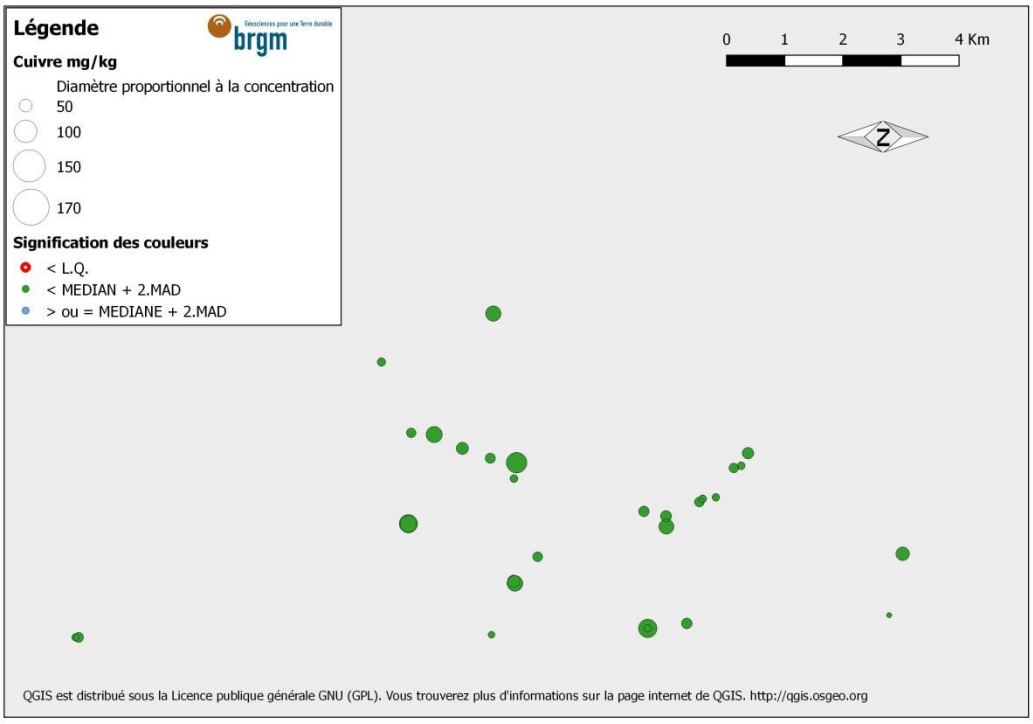


Figure 10 : Répartition spatiale des échantillons de cuivre de l'agglomération A en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

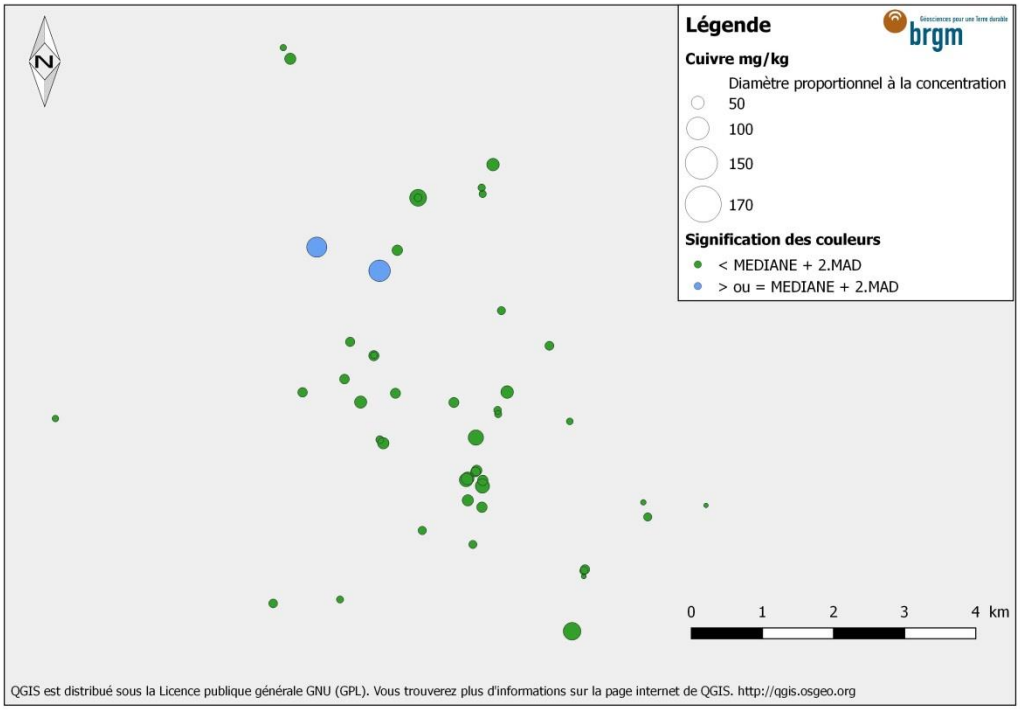


Figure 11 : Répartition spatiale des échantillons de cuivre de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

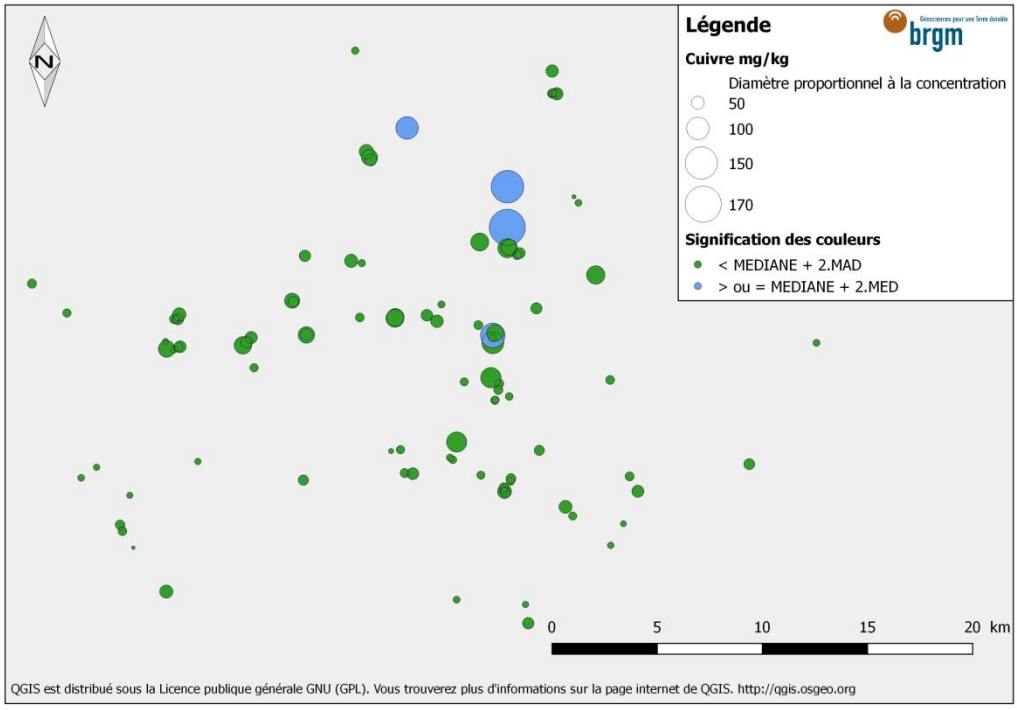


Figure 12 : Répartition spatiale des échantillons de cuivre de l'agglomération C en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

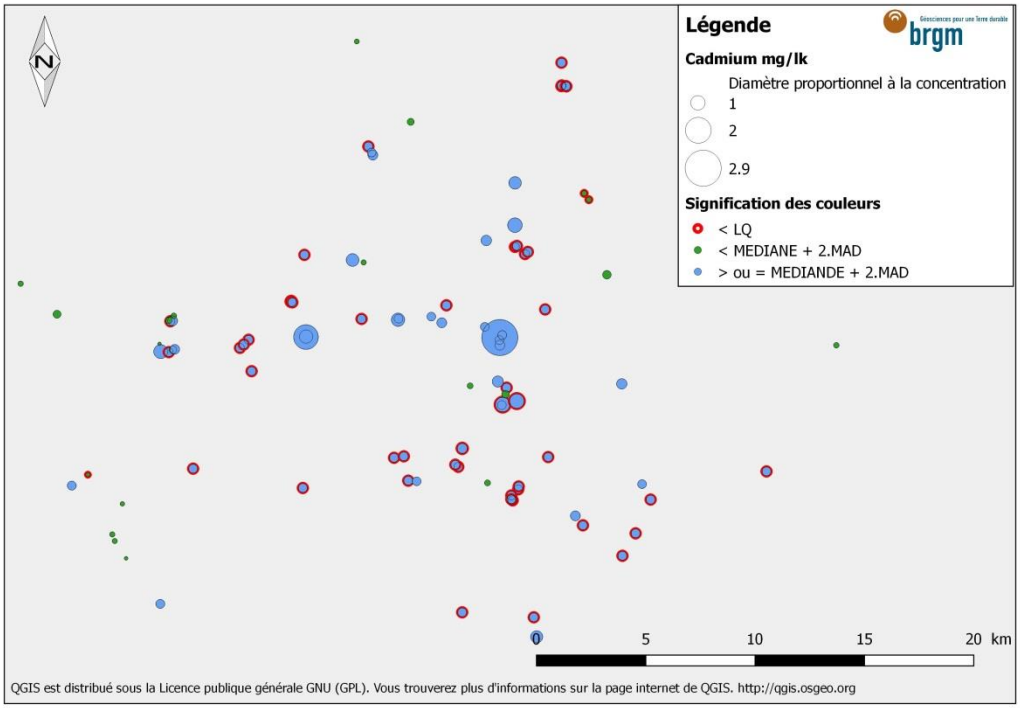


Figure 13 : Répartition spatiale des échantillons de cadmium de l'agglomération C en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.

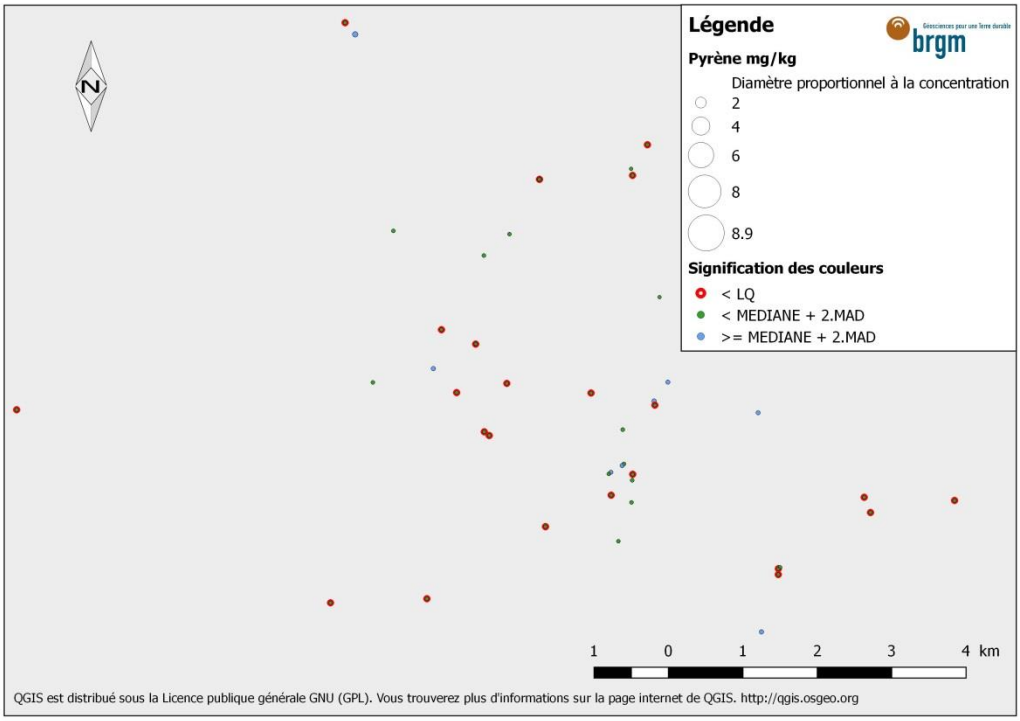


Figure 14 : Répartition spatiale des échantillons de pyrène de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.
(classes de concentrations déterminées à partir des données de l'ensemble des agglomérations A B et C).

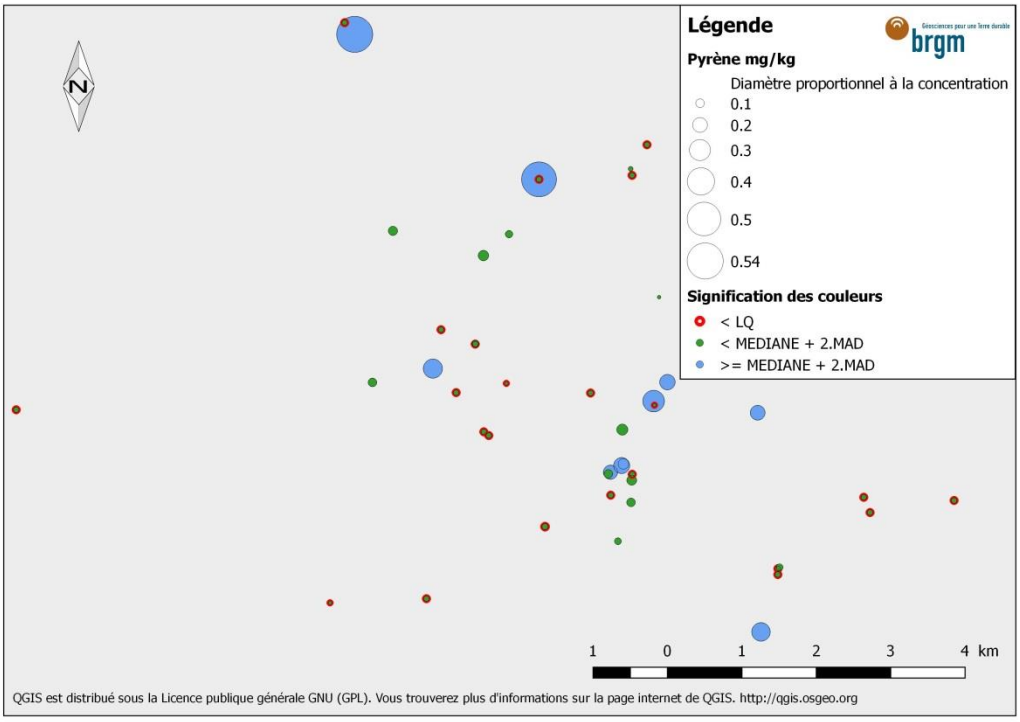


Figure 15 : Répartition spatiale des échantillons de pyrène de l'agglomération B en fonction du seuil du FPGA calculé avec la méthode MED + 2MAD.
(classes de concentrations déterminées à partir des données de l'agglomération B seule : échelle locale).

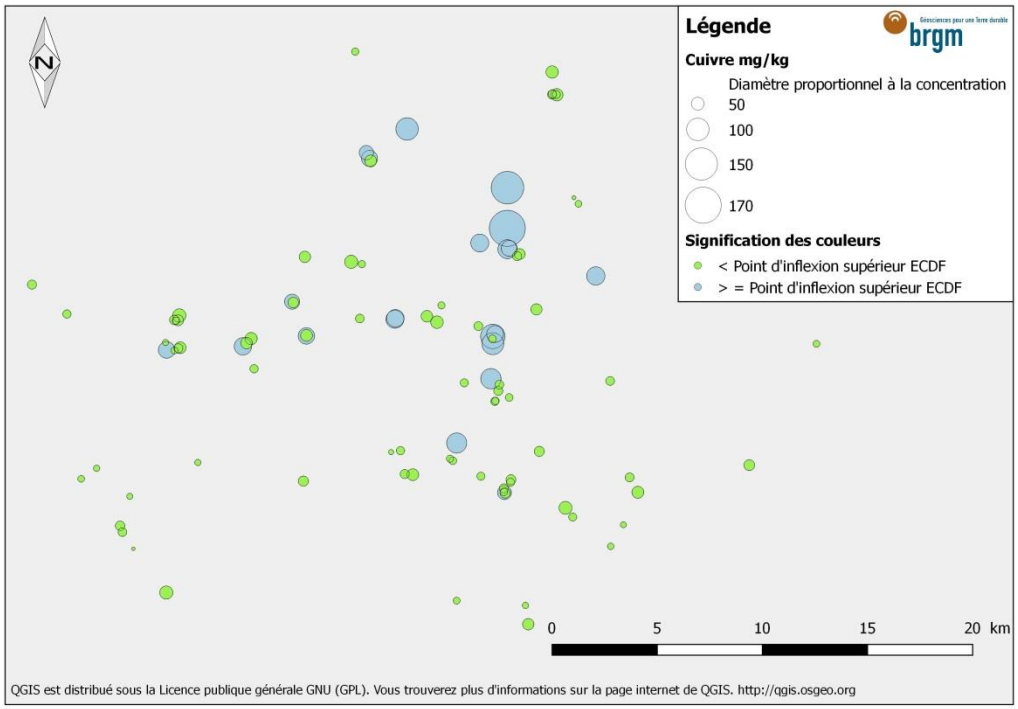


Figure 16 : Répartition spatiale des échantillons de cuivre de l'agglomération C en fonction du seuil du FPGA calculé avec la méthode la tangente à l'ECDF.

Géosciences pour une Terre durable
brgm

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