Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Méthodes mathématiques robustes

pour la détermination de seuils de pollution

Rapport n°2

adressé au

BRGM

(à l'attention de M. Stéphane Belbeze)

par la

Société de Calcul Mathématique SA

SCM SA, Janvier 2019

Rédaction par Gaëlle Tournié, Alisson Stocchetti, Léonore Rousselle

En référence au bon de commande du 27 septembre 2018 n°engagement 211416

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel : 01 42 89 10 89. Fax : 01 42 89 10 69. www.scmsa.eu Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

Résumé Opérationnel

Contexte de l'étude

Le BRGM travaille actuellement sur le projet "Établissement d'un fond géochimique urbain et industriel" (dit FGU). Pour ce projet, plusieurs prélèvements des sols d'une agglomération ont été effectués à différents endroits et à différentes profondeurs. Chaque prélèvement a été analysé du point de vue des teneurs en As, Cu, Cr, Pb, Zn, Ni, Cd, Hg, Ba, Co, Se, Mo, Sb, Cyanures totaux, Naphtalène, HAP, Phénols, benzène, BTEX, Hydrocarbures, dioxines. Les données sont regroupées dans la base "Fusion_Ech_150119_SO2" fournie par le BRGM.

L'objet du rapport est de valider une méthode de calcul des seuils de pollution parmi les quatre qui sont proposées :

– Centile 90% : c'est une valeur qui partage la série ordonnée en deux sous-ensembles tels que 90% des données se trouvent en dessous de cette valeur et 10% au-dessus ;

– Centile 95% : c'est une valeur qui partage la série ordonnée en deux sous-ensembles tels que 95% des données se trouvent en dessous de cette valeur et 10% au-dessus ;

– MED ± 2.MAD : "médiane ± 2 x Median Absolute Deviation (médiane des écarts absolus à la médiane)". Cette méthode est la version robuste de "moyenne ± 2 x écart-type", pour s'affranchir de l'influence des valeurs extrêmes (qui sont nombreuses dans notre jeu de données). En effet, cette méthode fait intervenir la MAD, une mesure robuste qui détermine la dispersion des données. Le résultat de cette méthode donne une valeur seuil ;

– Vibrisse interne supérieure du boxplot de Tukey : le boxplot de Tukey est une représentation graphique des statistiques de base (minimum, maximum, quantiles, médiane, valeurs extrêmes). Les vibrisses correspondent aux extrémités du boxplot. La valeur de la vibrisse interne supérieure, valeur utilisée pour déterminer un seuil, est définie par la formule suivante : $Vibrisse(X) = max(X[X \leq Q75\% + 1{,}5 \times (Q75\% - Q25\%)])$.

Le BRGM souhaite une comparaison critique des résultats obtenus avec chaque méthode, accompagnée d'éventuels commentaires.

Nos conclusions :

1. **Les méthodes MED ± 2.MAD et "vibrisse interne supérieure" sont à exclure.**

En effet, elles requièrent, pour pouvoir être utilisées, une hypothèse de normalité qui n'est pas satisfaite en général. En outre, les résultats donnés par MED ± 2.MAD sont très éloignés des résultats donnés par les trois autres méthodes (cf. II.B. Seuils obtenus pour les quatre méthodes) ;

Pour satisfaire l'hypothèse de normalité, le BRGM voudrait transformer les données, en utilisant par exemple une transformation logarithmique (ou Box-Cox). Mais, même en utilisant

une telle transformation, les données ne suivent pas nécessairement une loi normale (cf. II.A.4. Robustesse des méthodes). Les données log-transformées n'ont rien à voir avec les données originales : il est difficile de transposer les résultats obtenus aux données originales. Concernant la "vibrisse interne supérieure", il faut que la distribution des données soit symétrique, ce qui n'est pas le cas (cf. II.A.4. Robustesse des méthodes).

2. **Les méthodes Q90% et Q95% sont robustes et peuvent être utilisées pour déterminer un seuil de pollution**

Nous préconisons donc l'emploi de ces méthodes. Voyons maintenant laquelle choisir. Nous avons comparé les seuils de pollution obtenus à partir des centiles 90% et 95% avec les concentrations naturelles des polluants du type "éléments à traces métalliques". Les seuils Q90% et Q95% obtenus pour :

– Mo, Ba sont inférieurs à la variabilité naturelle ;
– As, Sb, Co, Cr, Ni et Zn sont inclus dans la variabilité naturelle ;
– Cd, Cu, Hg et Pb sont supérieurs à la variabilité naturelle.

Notre conclusion est que le centile 90% sera préféré au centile 95% si le polluant est nocif et inversement si le polluant n'est pas nocif.

En effet, le centile Q90% donnera un seuil de pollution plus restrictif que Q95%.

Les polluants comme l'arsenic, le cadmium et le plomb sont parmi les plus nocifs : on utilisera le centile 90%. Dans les prélèvements, le cadmium et le plomb ont des concentrations généralement plus élevées que celles observées dans la Nature.

Inversement, le centile 95% sera préféré quand la concentration d'un polluant non nocif est incluse dans la variabilité naturelle : c'est le cas pour le cuivre, le nickel et le zinc.

Pour répondre à une question spécifiquement posée par le BRGM, nos analyses montrent que les variations de concentration sont les mêmes en sols urbains ou ruraux. Au sein même des sols urbains, il n'y a pas de point particulier de pollution.

Nos recommandations

Pour les polluants nocifs, il faut utiliser la méthode Q90%. On compare le seuil obtenu avec la variabilité naturelle. Si on est au dessus de la variabilité naturelle, danger : il faut dépolluer ou éviter de construire.

Pour les polluants non-nocifs, on utilise la méthode Q95% et on compare avec la variabilité naturelle. La décision de dépollution est alors politique.

Analyses de données

Le BRGM, en suite à la réception de notre premier rapport (décembre 2018) a posé la question des données inférieures à la limite de quantification (LQ) : comment les traiter ?

Plus de la moitié des polluants ont plus de 50% de leurs concentrations inférieures à la LQ. Il y a des incohérences au sein des données : dans le fichier "Fusion_Ech_150119_SO2_LD" remis par le BRGM, la concentration d'un polluant apparaît à la fois comme inférieure strictement et égale à sa LQ.

Le BRGM préconise de substituer les valeurs des concentrations de polluant inférieures à la LQ par 50% de leurs LQ respectives. Nous ne possédons pas les valeurs des LQ : nous supprimons les concentrations inférieures à la LQ.

Sommaire

Résumé Opérationnel 2
I. Données 6
A. Présentation des données 6
B. Analyse des données 7
1. Identification des outliers 7
2. Données inférieures à la limite de quantification (LQ) 12
3. Le cadre de référence des prélèvements 14
4. La concentration des polluants au sein des prélèvements 23
II. Exploitation des résultats 40
A. Présentation des méthodes 40
1. Méthodes 1 et 2 : Centile 90% et 95% 40
2. Méthode 3 : MED + 2MAD 41
3. Méthode 4 : Vibrisse interne supérieure du boxplot de Tukey 42
4. Robustesse des méthodes 43
5. Nombre d'échantillons 44
B. Seuils obtenus pour les quatre méthodes 45
C. Proportion des prélèvements supérieurs aux seuils 46
D. Comparaison avec la variabilité naturelle 47
III. Influence et recommandation des sites 48
A. Influence des sites sur les seuils 50
B. Détermination des prélèvements pollués 57
IV. Classement des prélèvements 58
V. Conclusion 62
Annexe : Concentration en fonction de la profondeur 63
Bibliographie 67

I. Données

A. Présentation des données

Le BRGM nous a fourni un certain nombre de données, dont la base de données des prélèvements "Fusion_Ech_150119_SO2".

Cette base de données (notée BDD) contient 37 paramètres, présentés dans le tableau ci-dessous :

Nom du paramètre	Unité	Exemple	Description
N	Aucune	1	Index
NOM	Aucune	Z91P6	Nom du sondage
XL93	m	567096.32	Coordonnées Lambert 93
YL93	m	6273022.49	Coordonnées Lambert 93
TO	m	2	Fin d'échantillonnage
FROM	m	0.9	Début d'échantillonnage
As	mg/kg	9	Concentration d'Arsenic (traces métalliques)
Sb	mg/kg	7	Concentration d'Antimoine
Ba	mg/kg	130	Concentration de Baryum
Cd	mg/kg	36	Concentration de Cadmium (traces métalliques)
Co	mg/kg	9	Concentration de Cobalt
Cr	mg/kg	40	Concentration de Chrome (traces métalliques)
Cu	mg/kg	38	Concentration de Cuivre (traces métalliques)
Hg	mg/kg	6	Concentration de Mercure (traces métalliques)
Ni	mg/kg	33	Concentration de Nickel (traces métalliques)
Pb	mg/kg	51	Concentration de Plomb (traces métalliques)
Se	mg/kg	5	Concentration de Sélénium
Mo	mg/kg	20	Concentration de Molybdène
Zn	mg/kg	139	Concentration de Zinc (traces métalliques)
HAPs	mg/kg	2	Concentration en Hydrocarbures Aromatiques Polycycliques (somme des 16) (substances persistantes organiques)
InPHE	mg/kg	11	Indice de Phénol
NAP	mg/kg	1	Concentration de Naphtalène
CNT	mg/kg	2	Cyanure totaux (substances persistantes organiques)
PCBs	mg/kg	0,02	Concentration de Polychlorobiphényles (substances persistantes organiques)
B	mg/kg	0,05	Concentration de Benzène
BTEX	mg/kg	0,25	Concentration de Benzène, Toluène, Ethyl benzène et Xylène
TEX	mg/kg	0,2	Concentration de Toluène, Ethyl benzène et Xylène
C5C10	mg/kg	30	Hydrocarbure
C10C40	mg/kg	20	Hydrocarbure

Nom du paramètre	Unité	Exemple	Description
DIOX	mg/kg	4,91	Concentration de Dioxines (substances persistantes organiques)
OUT	Aucune	1	Outliers : 9 si oui, 1 si non
ECH	Aucune	S2	Libellé profondeur d'échantillonnage
EP	m	0,05	Epaisseur de l'échantillon
GEOBAP	Aucune	Eau	Codage information usage
CLC	Aucune	Aéroports	Codage information usage corine land cover
CEN	m	0,025	Centre d'échantillonnage
GEOLE	Aucune	BP	Code geole

B. Analyse des données

La base de données BDD contient 1 679 prélèvements.

1. Identification des outliers

Afin d'identifier les outliers, nous utilisons une approche différente de celle utilisée par le BGRM. En effet, le BGRM procède de façon visuelle à partir d'une poursuite de projection multidimensionnelle.

a. La méthode du boxplot

En utilisant la méthode du boxplot sur les données brutes, nous trouvons 846 outliers, contre 323 outliers trouvés par le BRGM.

Pour déterminer les outliers, le BRGM transforme les données brutes via le log-transformé, puis applique la méthode du boxplot :

> *"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* [BRGM, p.42]."

En règle générale, nous évitons de travailler sur des données transformées : nous travaillons sur des données brutes. Néanmoins, nous avons testé la méthode utilisée par le BRGM (données transformées puis boxplot) pour pouvoir comparer les résultats : nous trouvons 658 outliers, contre 323 outliers détectés par le BRGM.

Quelle que soit la méthode utilisée ("données brutes et boxplot" ou "données transformées et boxplot"), le nombre d'outliers observé reste largement supérieur à celui observé par le BRGM. Nous ne connaissons pas l'entièreté des règles utilisées par le BRGM pour identifier ses outliers. Aussi, nous resterons sur la méthode qui ne suscite aucune transformation des données : dans la plupart des cas, la log-transformation n'aide pas à rendre les données moins variables ou plus normales et peut, au contraire, les rendre plus variables et plus faussées [Log].

Explication de la méthode :

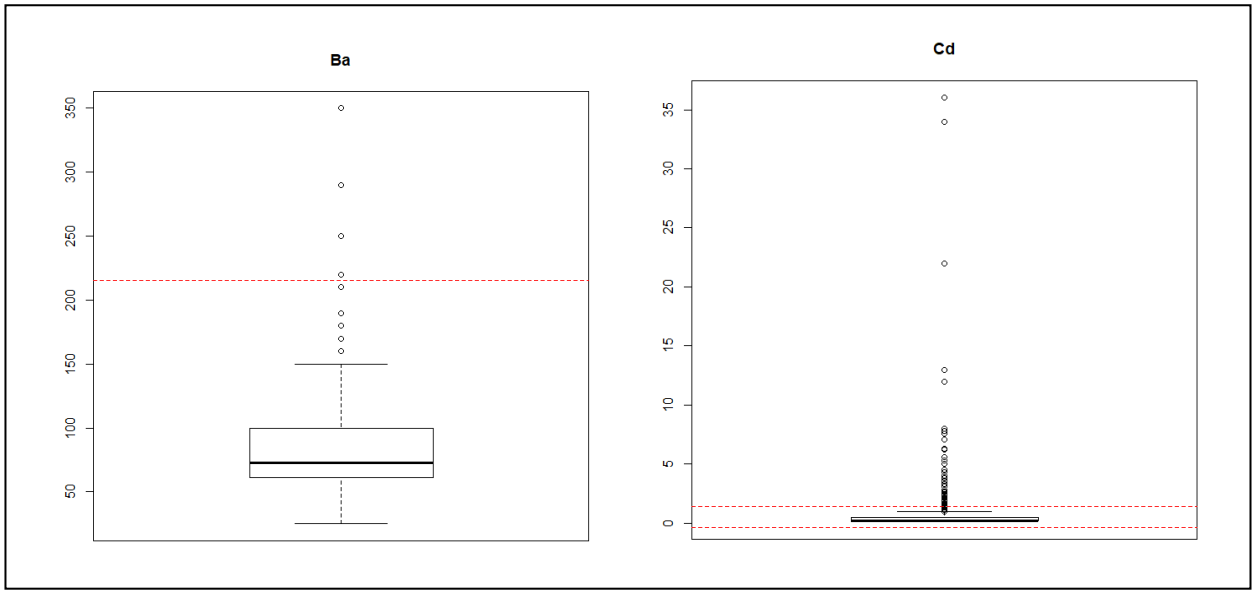
Nous commençons par tracer le boxplot afin de visualiser la répartition des concentrations pour le polluant donné.

Selon la littérature, les valeurs situées au-delà des vibrisses internes sont considérées comme des anomalies et toutes celles situées au-delà des vibrisses externes comme des anomalies « extrêmes » [Vibrisse]. La définition des vibrisses est la suivante :

$$Vibrisse\ inférieure\ (X) = Q25\% - \alpha \times (Q75\% - Q25\%)$$
$$Vibrisse\ supérieure\ (X) = Q75\% + \alpha \times (Q75\% - Q25\%)$$

Avec $\alpha = 1{,}5$ il s'agit de vibrisses internes, avec $\alpha = 3$ de vibrisses externes. Pour identifier les outliers nous utilisons les vibrisses externes. Nous calculons ces vibrisses et nous les faisons apparaitre sur les boxplots à l'aide de lignes pointillées rouges pour une meilleure visualisation des outliers.

À titre d'exemple (cf. figure ci-dessous) : à gauche, tous les points du polluant Ba situés au-dessus de la ligne pointillée rouge sont des outliers. Le polluant Ba a moins d'outliers que le polluant Cd.



Enfin, nous récupérons l'ensemble des prélèvements dont la concentration pour le polluant donné est supérieure à la vibrisse externe supérieure (resp. inférieure à la vibrisse externe inférieure).

Ce travail est effectué pour la totalité des polluants, ce qui signifie qu'un même prélèvement peut ressortir comme outlier pour plusieurs polluants. Au total, 846 prélèvements sont identifiés comme outliers (i.e. la concentration d'au moins un polluant est en dehors de l'intervalle de concentration formé par les vibrisses externes). Sur ces 846 prélèvements, seulement 295 coïncident avec les outliers identifiés par le BGRM.

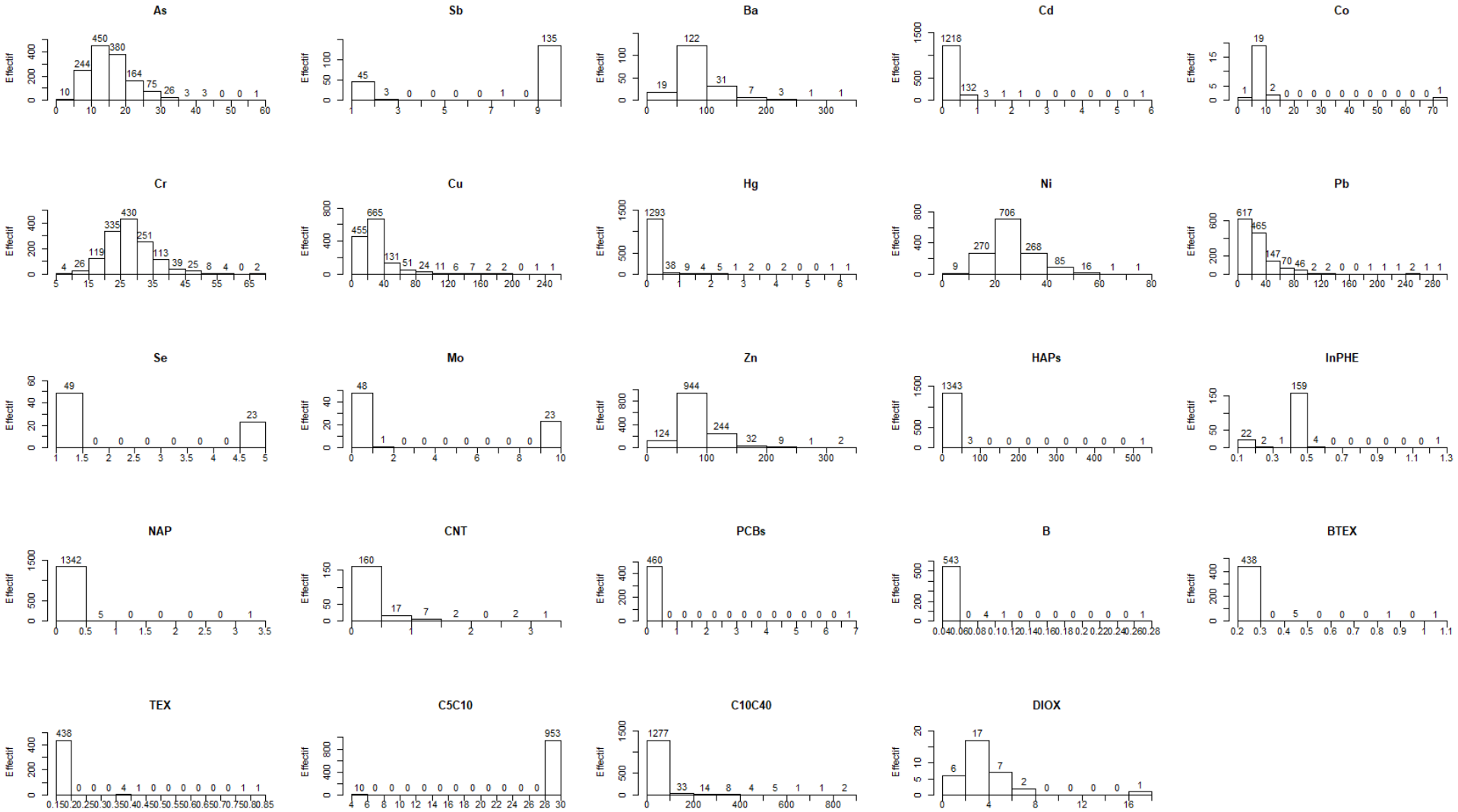
b. Comparaison entre les données sans outliers du BRGM et SCM

Pour rappel, le BRGM identifie avec ses méthodes 323 outliers, tandis que la SCM identifie 846 outliers en appliquant la méthode du boxplot.

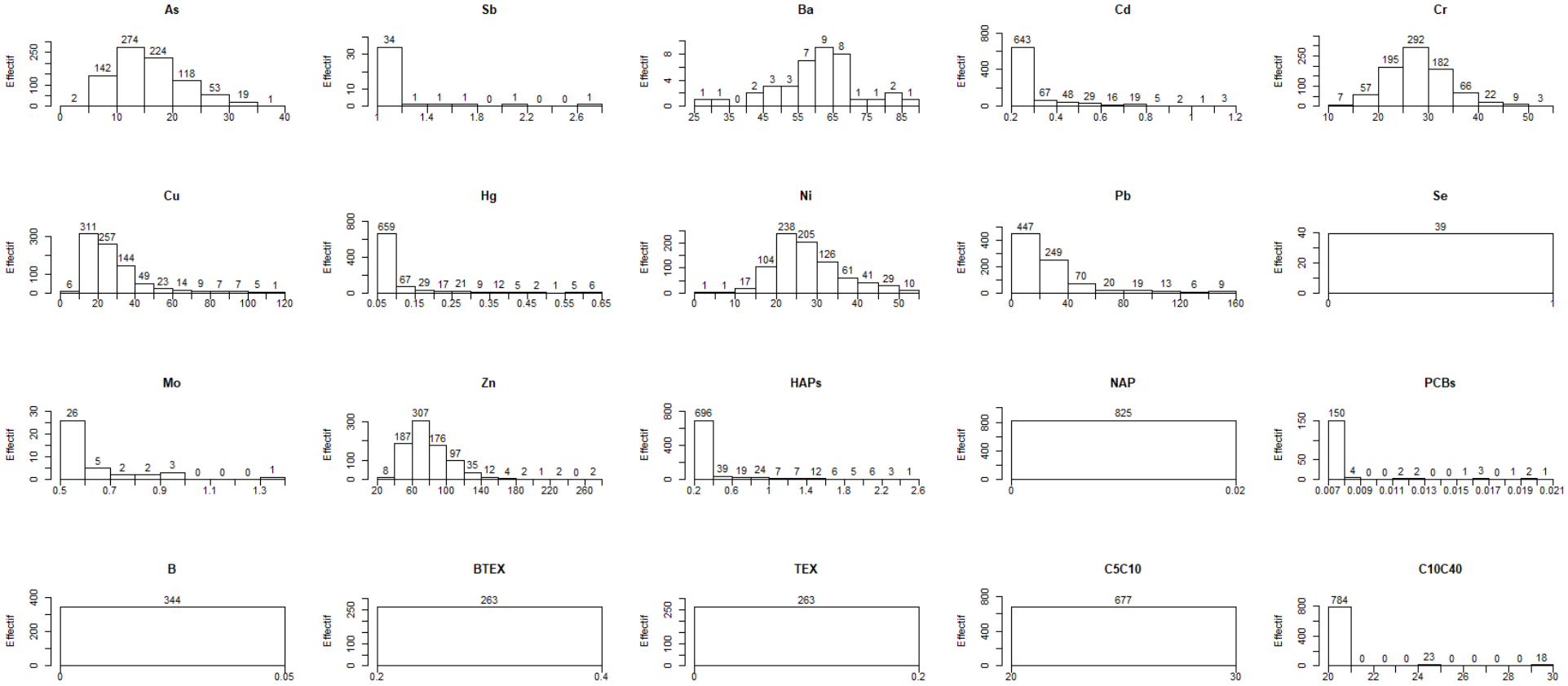
La méthode du boxplot est plus générale et plus adaptable à n'importe quelles données, mais elle est peut-être trop restrictive. En effet, plus de la moitié des prélèvements sont qualifiés d'outliers via cette méthode.

Nous voulons déterminer l'importance de l'influence du nombre d'outliers sur les distributions des concentrations des polluants. Pour ce faire, nous comparons les histogrammes des polluants sans les outliers détectés par le BRGM, avec les histogrammes des polluants sans les outliers obtenus par la SCM et vérifions si ces histogrammes diffèrent de manière significative (cf. graphiques ci-dessous).

Ci-dessous, les histogrammes des concentrations (en mg/kg) pour les polluants des prélèvements OUT = 1. Autrement dit, sans les outliers identifiés par le BGRM.



Ci-dessous, les histogrammes des concentrations (en mg/kg) pour les polluants des prélèvements sans les outliers que nous avons identifiés. Autrement dit, nous supprimons les données de tous les prélèvements dont au moins un polluant a une concentration considérée comme extrême par notre méthode boxplot.



Certains polluants (comme le Co, le InPHE, le CNT ou le DIOX) n'ont plus de valeurs suite à notre méthode d'identification des outliers.

La distribution des concentrations des polluants reste grossièrement la même (cf. As, Ba, Cd, Cr, Cu, Hg, Ni, Pb, Zn, HAPs, C10C40) avec parfois une étendue plus petite. Par exemple, pour le mercure, avec la méthode du BRGM les concentrations s'étendent de 0 à 7 mg/kg, alors qu'avec notre méthode boxplot elles s'étendent de 0,05 à 0,65 mg/kg. L'intervalle des valeurs peut être plus restreint, mais l'allure des histogrammes reste grossièrement la même : les pics de concentration restent dans les mêmes ordres de grandeur, que ce soit avec la méthode du BRGM ou la nôtre.

Suite à l'application de notre méthode, certains polluants (comme le NAP, le B, le BTEX, le TEX, ou le C5C10) se retrouvent avec toutes leurs valeurs de concentrations comprises dans un unique intervalle. Cet unique intervalle correspond toutefois au pic présent sur les histogrammes de la méthode BRGM, et ce pour l'ensemble des polluants cités.

Ayant des différences d'abscisse trop importantes pour quelques polluants, les seuils obtenus diffèrent entre les données sans les outliers du BRGM et sans les nôtres. Néanmoins, les proportions de prélèvements supérieurs au seuil sont grossièrement les mêmes.

Pour conclure, l'influence des outliers trouvés par le BRGM et la SCM sur les distributions des concentrations des polluants reste globalement la même : en effet, les histogrammes des données restantes pour le BRGM et la SCM ont grossièrement la même allure pour la majorité des polluants. Le choix de la méthode pour détecter les outliers n'est donc pas déterminant dans notre analyse.

Pour la suite de notre étude, nous garderons les outliers identifiés par le BRGM et travaillerons avec les données de la BDD sans ces outliers (i.e. nous conservons les données identifiées par "OUT = 1"). Nous avons donc un total de 1 356 prélèvements exploitables.

2. Données inférieures à la limite de quantification (LQ)

a. Analyse des données sur la LQ

Pour cette partie, nous utilisons le deuxième fichier fourni par le BRGM, "Fusion_Ech_150119_SO2_LD". Ce fichier contient l'information par rapport à la limite de quantification pour l'ensemble des polluants.

Pour rappel, la limite de quantification (LQ) correspond à la concentration la plus faible pouvant être mesurée de manière représentative.

Le tableau ci-dessous renseigne le nombre de données pour chaque polluant, ainsi que le nombre et la proportion de données inférieures à la limite de quantification.

Polluant	Nombre de prélèvements	Nombre de prélèvements < LQ	% de prélèvements < LQ
As	1356	7	0,52
Sb	184	176	95,65
Ba	184	0	0,00
Cd	1356	871	64,23
Co	23	0	0,00
Cr	1356	1	0,07
Cu	1356	1	0,07
Hg	1356	821	60,55
Ni	1356	1	0,07
Pb	1356	51	3,76
Se	72	72	100,00
Mo	72	49	68,06
Zn	1356	1	0,07
HAPs	1348	923	68,47
InPHE	189	179	94,71
NAP	1348	1227	91,02
CNT	189	125	66,14
PCBs	461	420	91,11
B	549	547	99,64
BTEX	445	442	99,33
TEX	445	442	99,33
C5C10	963	963	100,00
C10C40	1345	1041	77,40
DIOX	33	14	42,42

Lorsque le pourcentage de données inférieures à la LQ est élevé, l'utilisation de l'ensemble des valeurs brutes introduit un biais lors du traitement statistique, et les résultats deviennent ininterprétables.

> "*Si ces analysent comportent des résultats inférieurs aux limites de quantification analytiques (LQ), les ignorer, les remplacer par 0 ou par 100% de la LQ peut entraîner un biais préjudiciable à l'interprétation des résultats. La substitution par 50% de la LQ est un compromis rapide et facile qui permet de tenir compte de ces valeurs* [BRGM1, p14]."

> "*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* [BRGM, p17]."

Nous avons plus de la moitié des polluants avec plus de 50% de valeurs inférieures à la LQ (en rouge). Il s'agit surtout des hydrocarbures. Ne connaissant pas les valeurs des LQ respectives, nous ne pouvons pas substituer ces valeurs par 50% de leur LQ. Nous décidons de supprimer les valeurs de concentration inférieures à la LQ comme le suggère le BRGM.

Normalement, la suppression de ces valeurs ne pose pas de problème, car les paramètres que nous calculons ne sont pas sensibles. Or, ici, plus de la moitié des données de certains polluants sont supprimées, donc il est possible que les résultats obtenus par la suite soient influencés par cette modification. Cependant, n'ayant aucune information supplémentaire, c'est la seule modification qu'il nous ait possible d'effectuer.

Le tableau ci-dessous présente le nombre de prélèvements en fonction du nombre de polluants ayant une concentration inférieure à la LQ. Par exemple, 60 prélèvements présentent 3 concentrations de polluants inférieures à leur LQ respective.

Nombre de polluants < à la LQ	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Nombre de prélèvements	1	9	60	98	165	258	238	114	131	177	52	27	21	5

Seulement un prélèvement n'a aucun polluant ayant une concentration inférieure à la LQ. 26 prélèvements (soit 2% des prélèvements) ont la moitié (ou plus) des polluants qui ont une concentration inférieure à leur LQ.

b. Incohérences dans le fichier de données sur la LQ

Nous notons quelques incohérences au sein des données. En effet, une même valeur de concentration pour un polluant donné de la BDD est à la fois catégorisée comme « < » et comme « = » dans le fichier "Fusion_Ech_150119_SO2_LD". Nous explicitons deux exemples de ce cas de figure, mais il en existe plusieurs.

Le premier exemple concerne le chrome (Cr). La valeur de concentration 10 mg/kg est répertoriée comme inférieure à la LQ (ligne du fichier où le paramètre "N" vaut 1235), mais aussi comme égale à la LQ ("N = 2175" dans le fichier). De plus, une valeur de 9,4 mg/kg est aussi répertoriée comme égale à la LQ. Dans ce cas, comme se fait-il qu'une concentration de 10 mg/kg puisse être inférieure à la LQ si 9,4 mg/kg ne l'est pas ?

Le second exemple concerne le cadmium (Cd). Les valeurs de concentration 0,1 mg/kg, 0,2 mg/kg, 0,4 mg/kg, 0,5 mg/kg ou même 0,6 mg/kg sont indiquées comme étant inférieures à la LQ (e.g. "N = 2321" pour 0,6 mg/kg), alors que ces mêmes valeurs sont aussi répertoriées comme étant égales à la LQ (e.g. "N = 847" pour 0,6 mg/kg).

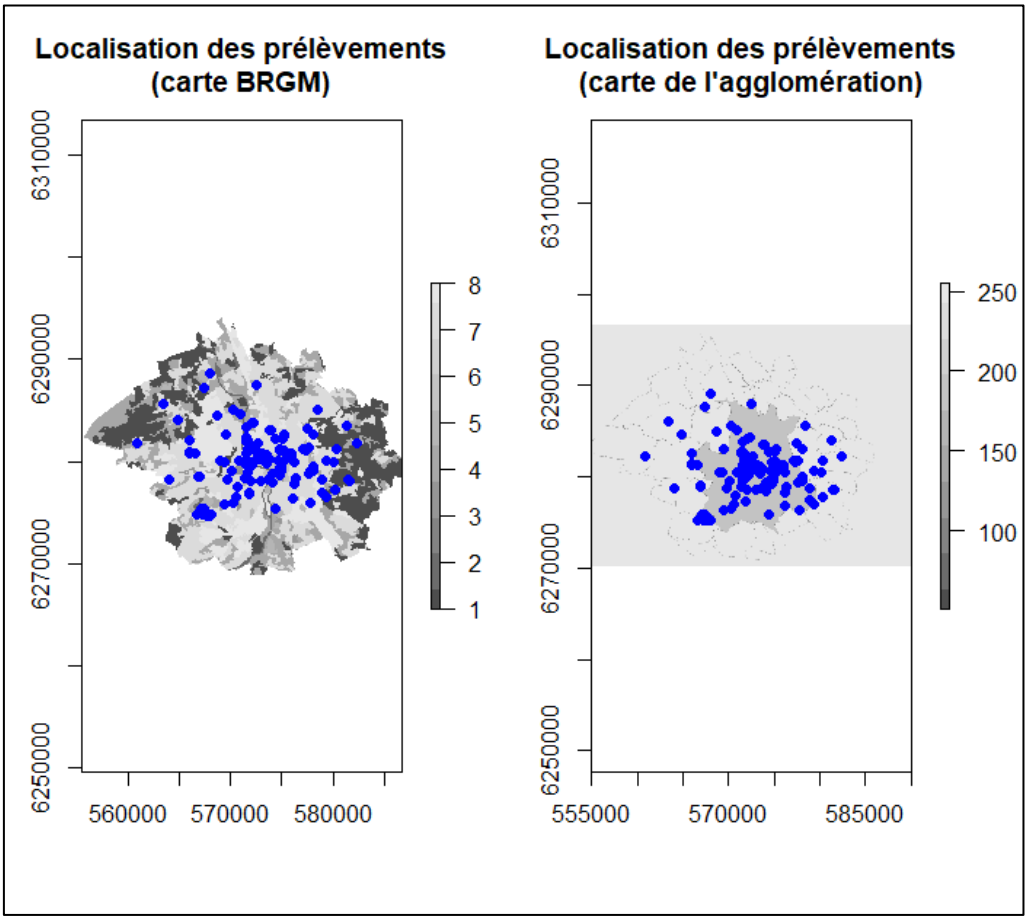
Nous décidons de travailler avec les données dont les concentrations inférieures à la LQ sont supprimées. Néanmoins, les résultats sont à interpréter avec vigilance suite aux incohérences mises en évidence sur les catégorisations des valeurs de concentration du fichier "Fusion_Ech_150119_SO2_LD".

3. Le cadre de référence des prélèvements

Dans cette partie, nous nous intéressons aux paramètres décrivant les conditions de prélèvements comme le lieu (paramètres XL93, YL93, GEOBAP, CLC, GEOLE) et la profondeur (paramètres FROM, TO, ECH, EP, CEN).

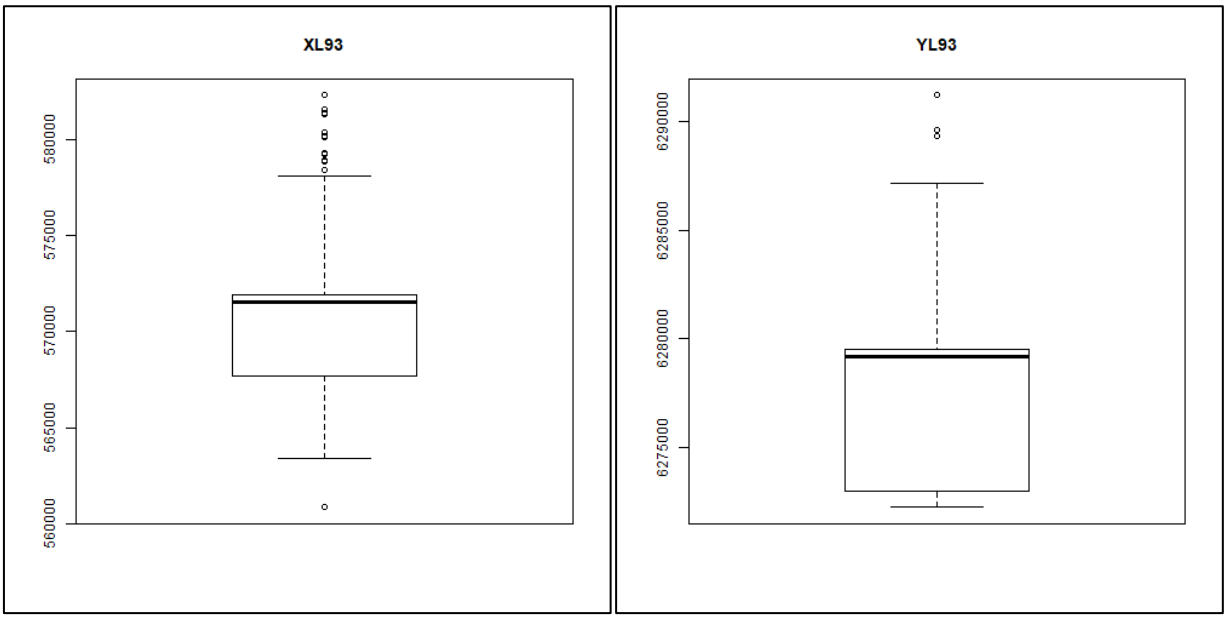
Les données utilisées sont les données décrites précédemment, c'est-à-dire les 1 356 prélèvements de la BDD du 15 janvier 2019 sans les outliers identifiés par le BRGM (OUT = 1) et avec les concentrations inférieures à la LQ supprimées. Les histogrammes / barplots obtenus sont similaires à ceux sans outliers du rapport d'avancement (cf. le rapport d'avancement "SCM_BRGM_R1_2018_12_05").

Chaque prélèvement a été effectué dans un lieu précis désigné par des coordonnées Lambert qui sont des données cartographiques (projection officielle pour les cartes de France métropolitaine). Les abscisses et ordonnées des coordonnées permettent de représenter les lieux de prélèvement sur une carte 2D (cf. figure ci-dessous).



- Paramètres "XL93" et "YL93" : les coordonnées des prélèvements

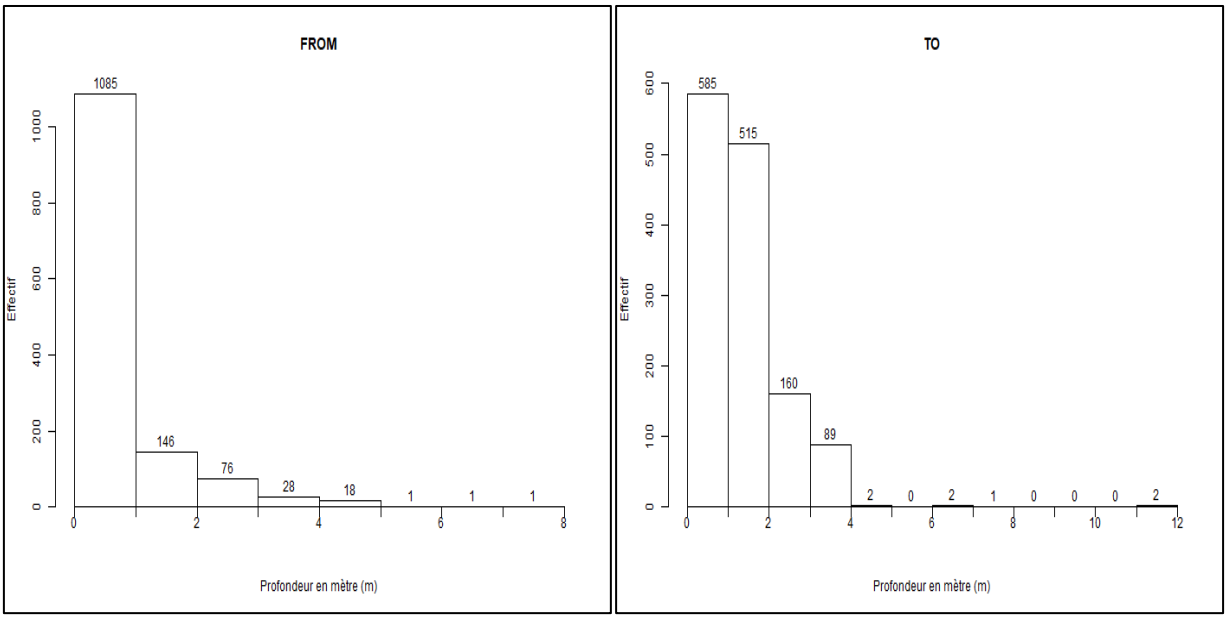
L'abscisse des coordonnées est donnée par le paramètre XL93 et l'ordonnée par le paramètre YL93. Les graphiques ci-dessous donnent la répartition des abscisses et des ordonnées des prélèvements.



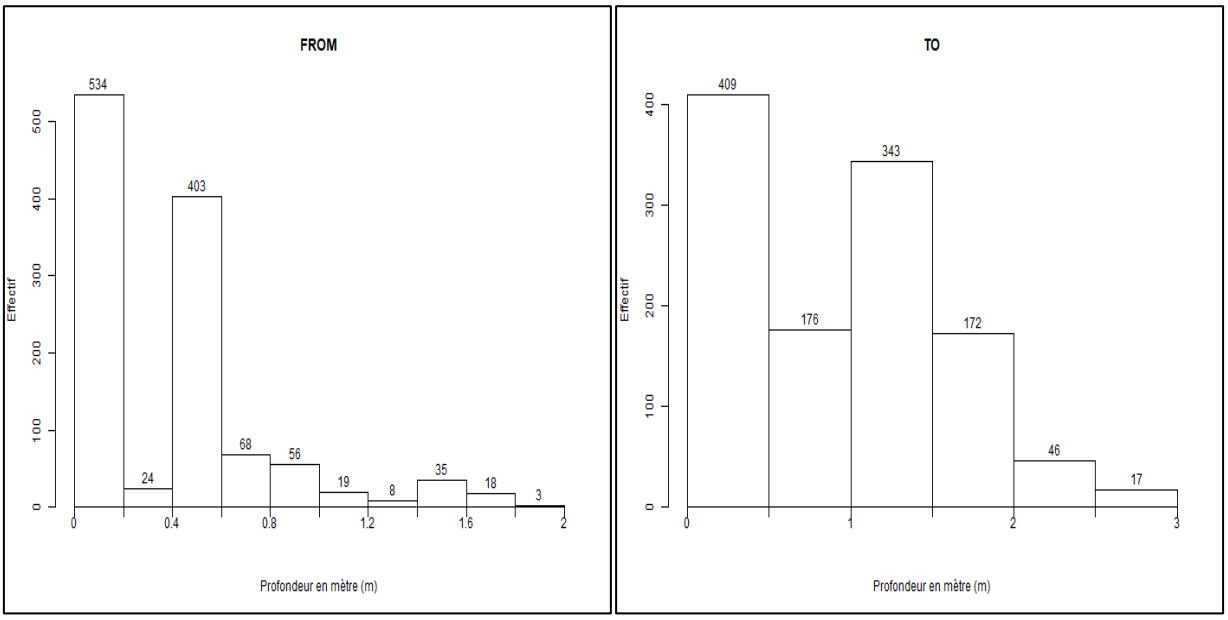
- Paramètres "FROM" et "TO" : la profondeur des prélèvements

Chaque prélèvement a une profondeur dite de « début » et une profondeur de « fin », ce qui permet de déterminer l'épaisseur totale du prélèvement.

Le début du prélèvement est renseigné par le paramètre FROM. En général, les prélèvements sont assez proches de la surface : ils commencent entre 0 et 1 mètre de profondeur. La fin du prélèvement est renseignée par le paramètre TO.



Nous décidons de faire un "zoom" sur les histogrammes. Avec les graphiques ci-dessous, nous avons représenté 90% des valeurs. Autrement dit, nous avons sélectionné toutes les valeurs inférieures au centile Q90%.



Avec le "zoom" sur l'histogramme, il est plus facile de se rendre compte que 90% des prélèvements débutent à la surface (entre 0 et 50 centimètres), alors que très peu débutent après 1 mètre de profondeur.

90% des valeurs de TO sont comprises entre 0 et 3 mètres de profondeur, ce qui signifie que les deux valeurs à 12 mètres sont des outliers non identifiés par le tri de base.

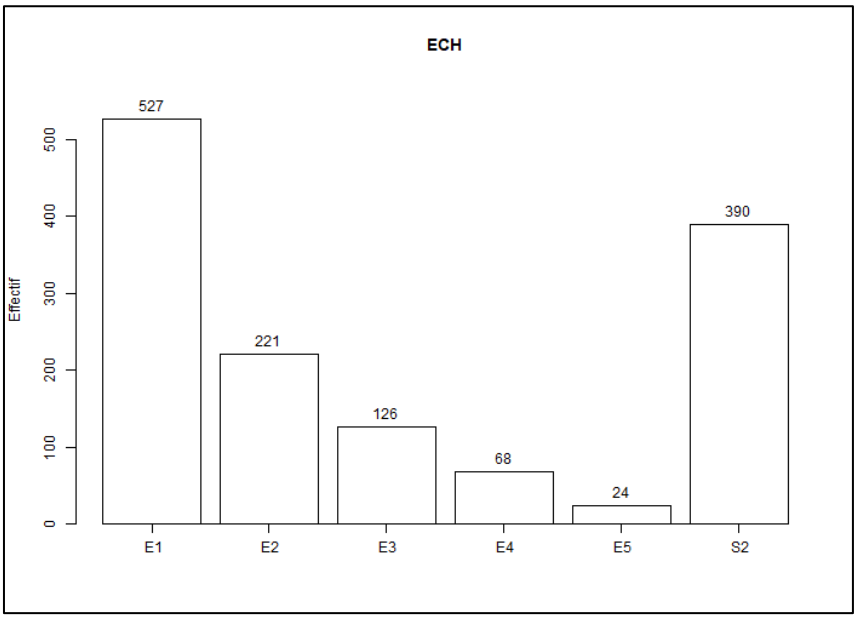
Beaucoup de prélèvements se terminent entre 0 et 1,5 mètre, ce qui signifie que la majorité des données correspondent à des prélèvements en surface ou peu profonds. Le paramètre suivant permet de labelliser les profondeurs.

- Paramètre "ECH" : catégorisation des profondeurs

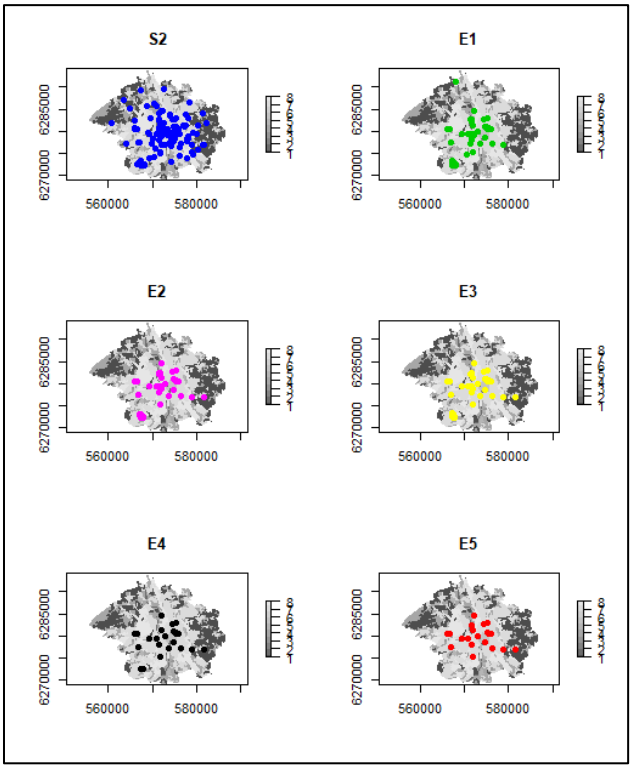
ECH est un paramètre qualitatif permettant de catégoriser les différentes profondeurs d'échantillonnage. Il y a au total 6 catégories :

— S2 : entre 0 et 55 centimètres ;
— E1 : entre 0 et 2 mètres ;
— E2 : entre 0 et 3 mètres ;
— E3 : entre 1 et 4 mètres ;
— E4 : entre 2,5 et 4,5 mètres ;
— E5 : entre 4 et 12 mètres.

Certaines catégories se chevauchent. Néanmoins, comme nous le constatons avec les paramètres précédents, la majeure partie des prélèvements (73%) se font en surface (entre 0 et 55 centimètres) ou peu profondément (cf. graphique ci-dessous).



Ci-dessous, la localisation des prélèvements en fonction de la profondeur. Il y a une carte par profondeur, car quelques prélèvements ont été effectués au même endroit, mais à différentes profondeurs donc les points se superposent (e.g. E2 et E3) et il devient difficile de les distinguer.



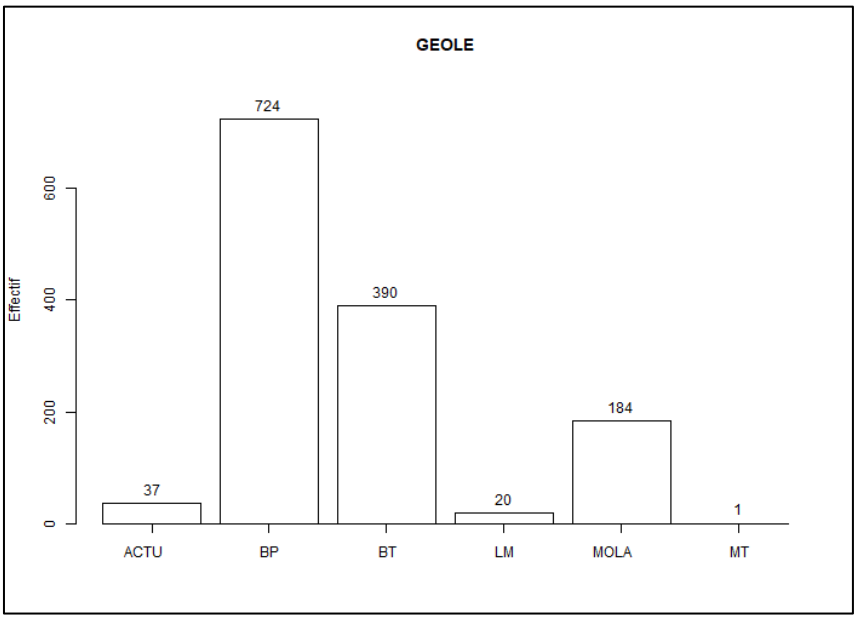
La répartition des prélèvements en surface (S2) est assez uniforme, mais ce n'est pas le cas pour les autres profondeurs. La distribution des profondeurs de prélèvements n'est pas aléatoire : les prélèvements sont faits là où une pollution est soupçonnée et non de manière aléatoire et uniforme sur l'ensemble de la zone étudiée.

- Paramètre "GEOLE" : couche géologique du prélèvement

Ce paramètre est également qualitatif. Il correspond à un code des couches géologiques au niveau des prélèvements. Il y a 6 catégories :

— ACTU : actuel ;
— BP : basse plaine ;
— BT : basse terrasse ;
— LM : lit majeur de la Garonne ;
— MOLA : formations molassiques ;
— MT : moyenne terrasse.

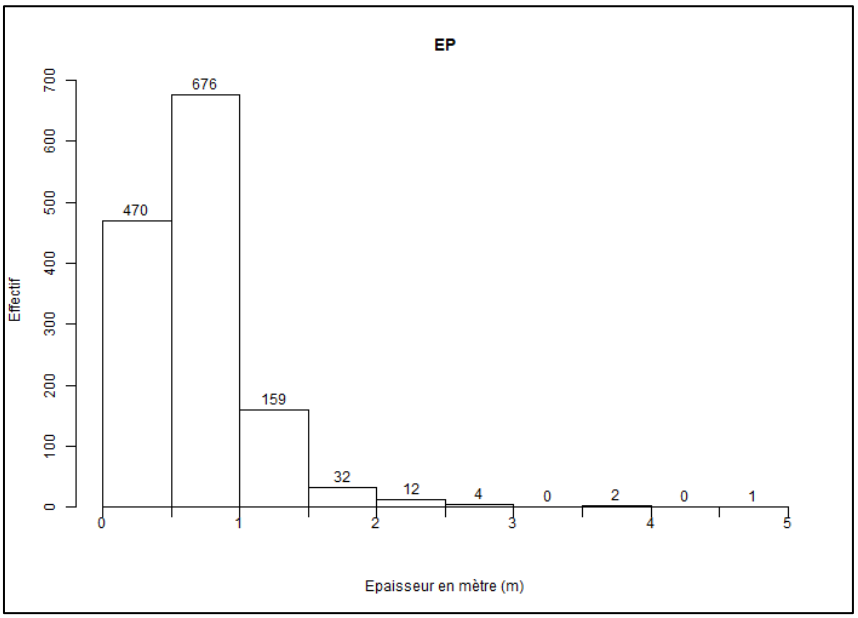
Selon les graphiques précédents et ci-dessous, les prélèvements en surface et les peu profonds sont effectués au sein de la basse plaine et éventuellement de la basse terrasse.



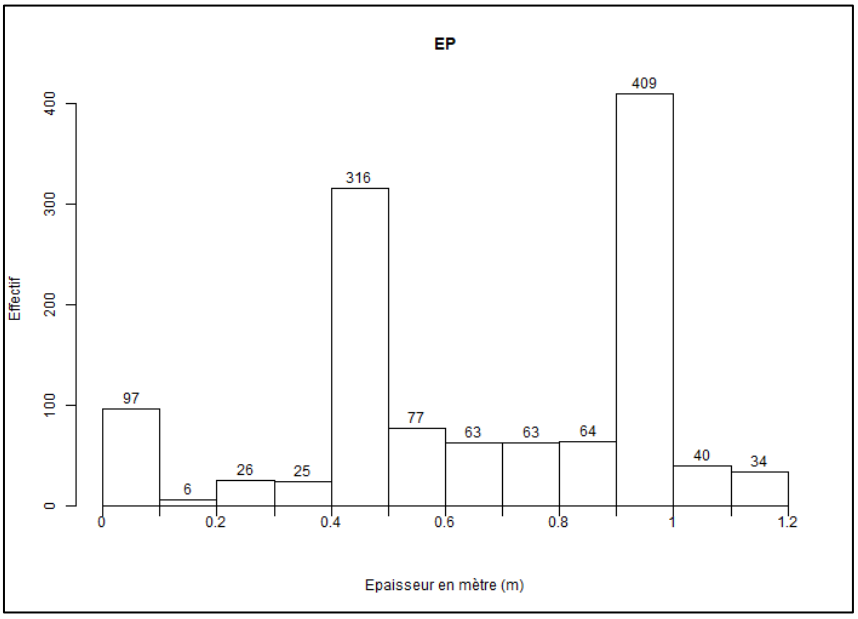
- Paramètre "EP" : épaisseur du prélèvement

Avec les paramètres FROM et TO, il est possible de récupérer l'épaisseur totale du prélèvement. C'est ce que donne le paramètre EP.

$$Epaisseur_{totale} = Profondeur_{fin} - Profondeur_{début} = TO - FROM = EP$$



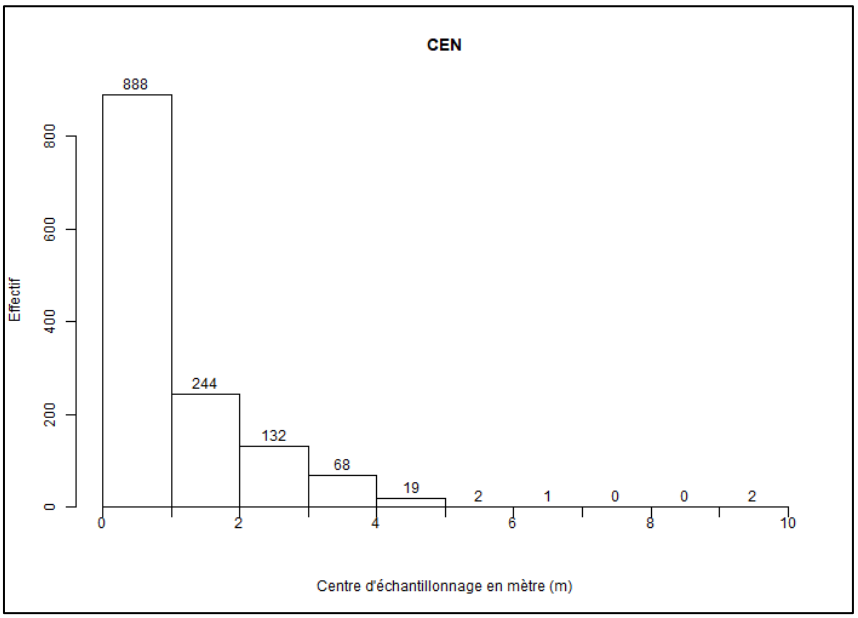
5 mètres de matières sont prélevés au maximum. Nous zoomons aussi l'histogramme afin d'obtenir un découpage plus précis des épaisseurs prélevées entre 0 et 1 mètre (i.e. en surface ou peu profondément). Le résultat est le suivant :



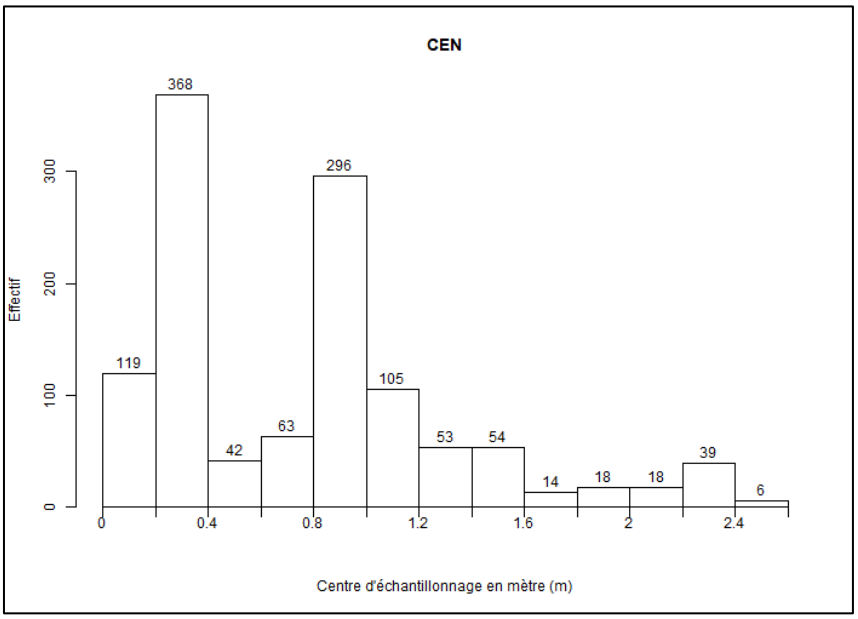
Généralement, ce sont des échantillons de 40 – 50 centimètres ou de 90 cm – 1 mètre qui sont prélevés. Dans 90% des cas, il y a moins de 1.3 mètre de prélevé.

- Paramètre "CEN" : milieu du prélèvement

CEN, paramètre quantitatif, renseigne le milieu de chaque échantillon prélevé.



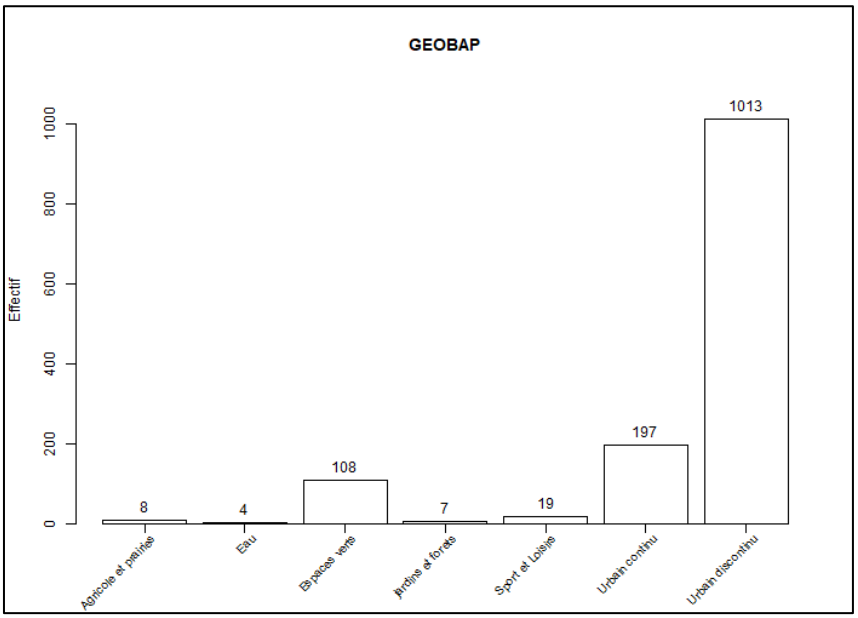
La conclusion reste la même : les prélèvements se font peu profondément.



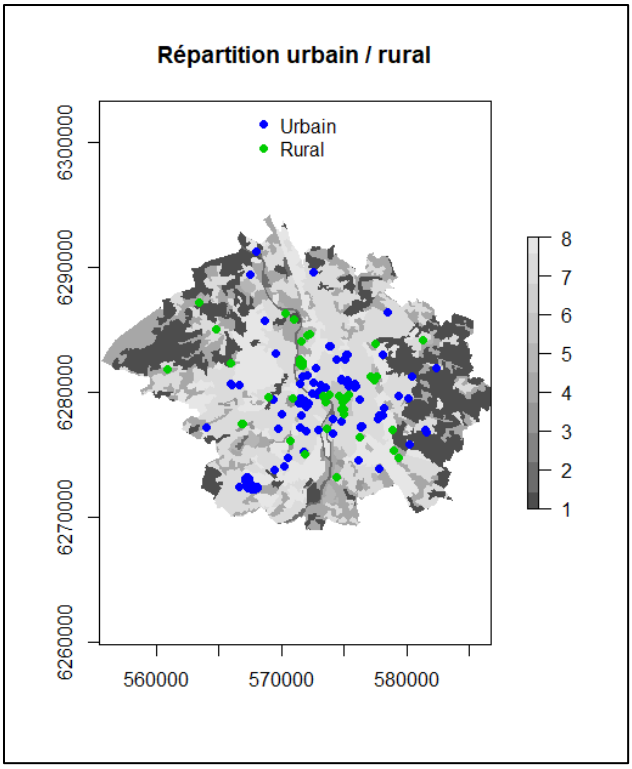
Avec le pic de centres entre 20 et 40 centimètres, nous retrouvons bien le pic des prélèvements de 40 à 50 centimètres d'épaisseur. De même, avec un centre se situant entre 80 et 90 centimètres, les prélèvements font moins de 2 mètres et se trouvent toujours dans l'une des catégories majoritaires de ECH qui se trouve être E1 (0 – 2 mètre).

- Paramètre "GEOBAP" : zone des lieux des prélèvements

Nous avons deux catégories de lieux de prélèvements : la zone rurale et la zone urbaine. La majorité des prélèvements se fait en zone urbaine (1 210 prélèvements contre 146 prélèvements en zone rurale).

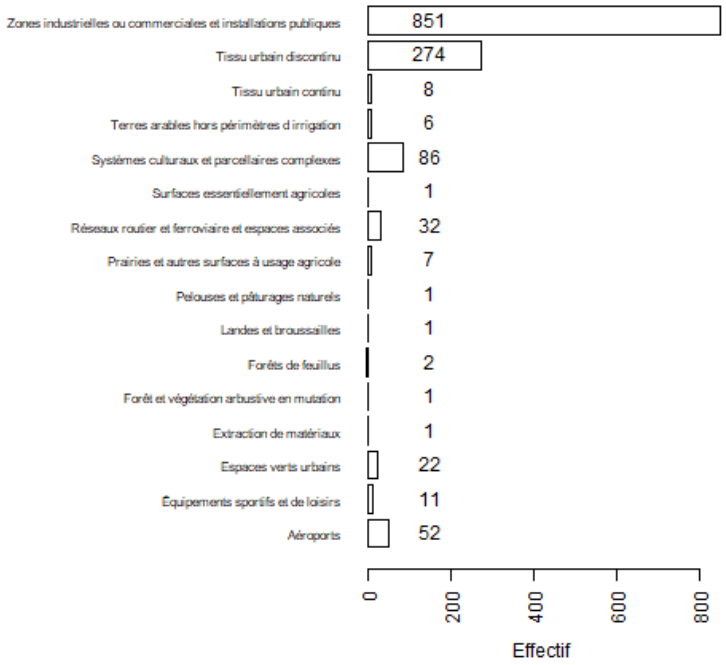


Le graphique ci-dessous permet de repérer géographiquement les prélèvements effectués en zone rurale et urbaine.



- Paramètre "CLC" : Corine Land Cover des prélèvements

Corine Land Cover est une base de données européenne d'occupation biophysique des sols. Elle est destinée à cartographier l'ensemble du territoire de l'Union européenne avec de nombreuses applications dans le domaine de l'environnement.



Les zones urbaines ressortent aussi majoritaires avec ce paramètre, ce qui reste cohérent avec le fait qu'il y ait plus de prélèvements en milieu urbain qu'en milieu rural.

- Conclusion sur le cadre de référence des prélèvements

Les prélèvements ont principalement lieu entre 0 et 2 mètres, c'est-à-dire en surface ou peu profondément. Ils sont de petites épaisseurs (moins de 1 mètre) et localisés en zones urbaines.

4. La concentration des polluants au sein des prélèvements

De même que pour la partie précédente, les données utilisées sont les données de la BDD du 15 janvier 2019 sans les outliers identifiés par le BRGM (OUT = 1) et avec les concentrations inférieures à la LQ supprimées.

Les nouveaux effectifs de chaque polluant (i.e. sans NA, sans outliers et sans données inférieures à la LQ) sont les suivants.

Polluant	Effectif
As	1349
Sb	8
Ba	184
Cd	485
Co	23
Cr	1355
Cu	1355
Hg	535

Polluant	Effectif
Zn	1355
HAPs	425
InPHE	10
NAP	121
CNT	64
PCBs	41
B	2
BTEX	3

Polluant	Effectif
Ni	1355
Pb	1305
Se	0
Mo	23

Polluant	Effectif
TEX	3
C5C10	0
C10C40	304
DIOX	19

Nous présentons les histogrammes de chaque polluant en fonction de leur concentration dans les prélèvements ainsi que la localisation des prélèvements sur la carte de l'agglomération en fonction de leur classe de concentration. Les histogrammes correspondent au code couleur de référence pour les classes de concentration.

Les histogrammes obtenus sont également similaires à ceux sans outliers du rapport d'avancement (cf. le rapport d'avancement "SCM_BRGM_R1_2018_12_05").

- As : arsenic

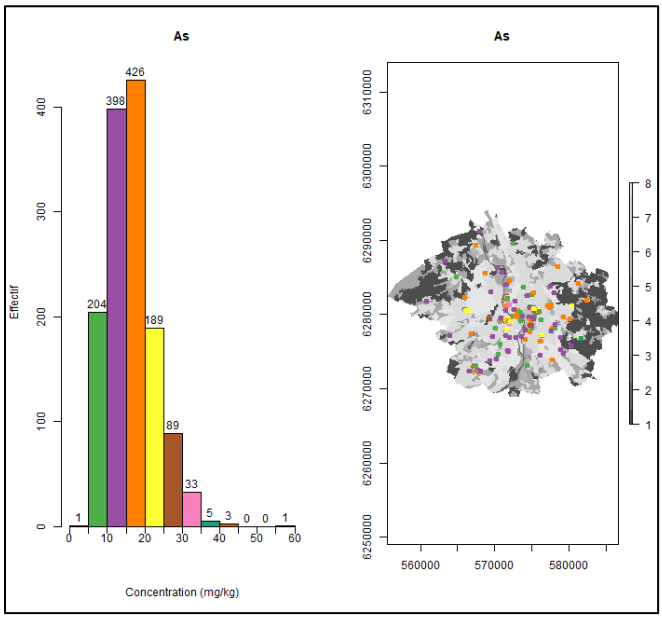
L'arsenic est un élément naturel. Dans la partie superficielle de l'écorce terrestre, sa concentration moyenne est évaluée à 2 mg/kg, mais localement le fond géochimique peut atteindre des valeurs beaucoup plus élevées (200 mg/kg dans des dépôts calcaires, phosphatés ou dans des schistes) [INERIS].

Dans les sols, la concentration en arsenic est généralement comprise entre 0,1 et 40 mg/kg d'après Bowen (1979) et entre 1 et 50 mg/kg d'après Backer et Chesnin (1975) [BRGMAs].

Il est aussi principalement émis par les industries produisant du cuivre, du plomb ou du zinc ainsi que dans l'agriculture.

L'arsenic est un élément chimique plutôt mobile, ce qui signifie qu'il est peu probable d'avoir des concentrations importantes dans un lieu spécifique. De ce fait, la pollution à l'arsenic est un vaste problème, car il se diffuse facilement.

Sur le graphique ci-dessous, les points violets, par exemple, représentent des classes de concentration en arsenic de [10 ; 15[mg/kg ; ils se répartissent sur la carte de manière uniforme.

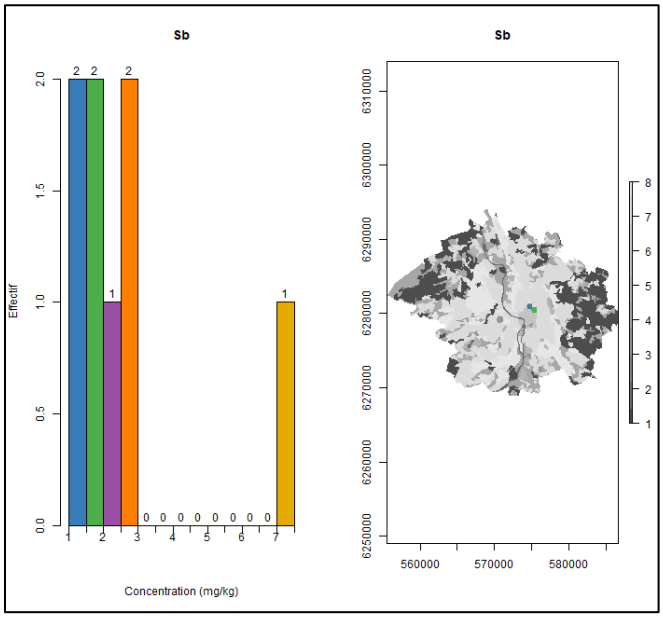


Les prélèvements présents dans une même classe de concentration ne sont pas regroupés en une zone sur la carte, mais sont au contraire dispersés sur l'ensemble de la carte (cf. les points violets ou orange).

- Sb : antimoine

L'antimoine est présent naturellement dans l'environnement, mais il y est aussi introduit par l'activité humaine. Il pollue principalement les sols et peut traverser de grandes distances dans les eaux souterraines.

Nous avons 8 prélèvements exploitables (i.e. sans NA) dans la BDD. La concentration d'antimoine est principalement comprise entre 1 et 3 mg/kg. Ainsi, quasiment la totalité des relevés de concentration de ce polluant est ininterprétable comme ils sont supprimés du fait d'être plus faibles que la concentration la plus faible pouvant être mesurée de manière représentative.

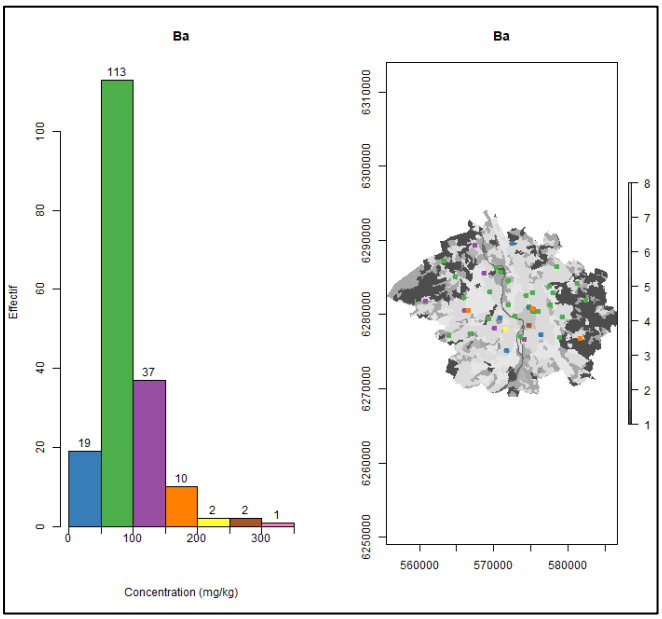


- Ba : baryum

Le baryum est un métal blanc argenté naturellement présent dans l'environnement en quantité très faible. Il y est présent en combinaison avec d'autres produits chimiques tels que le soufre, le carbone et l'oxygène.

On ne peut trouver de grandes quantités de baryum que dans les sols et quelques aliments à cause de son utilisation intensive dans les industries. Les composés de baryum qui sont persistants restent à la surface du sol, mais en général, les niveaux de baryum dans les sols sont faibles.

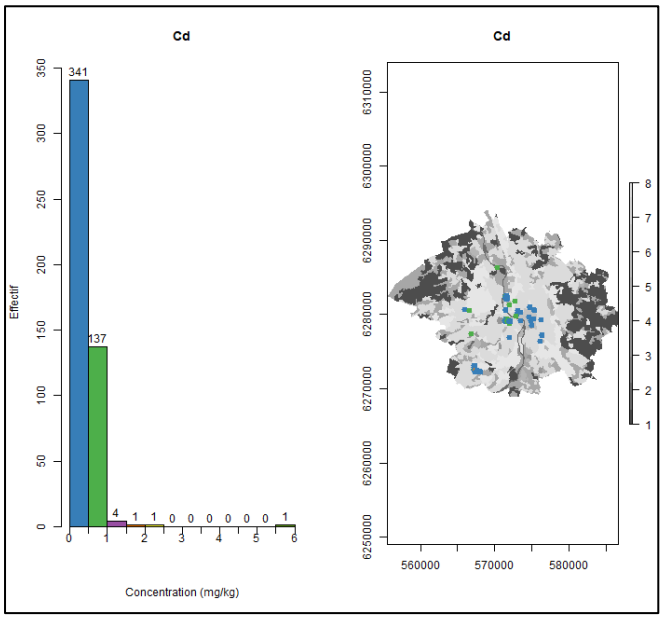
Nous avons 184 prélèvements présentant une concentration en baryum allant de 25 à 350 mg/kg. Aucun prélèvement de notre jeu de données n'a de concentration inférieure à 25 mg/kg. La plupart des prélèvements ont une concentration en baryum comprise entre 50 et 80 mg/kg.



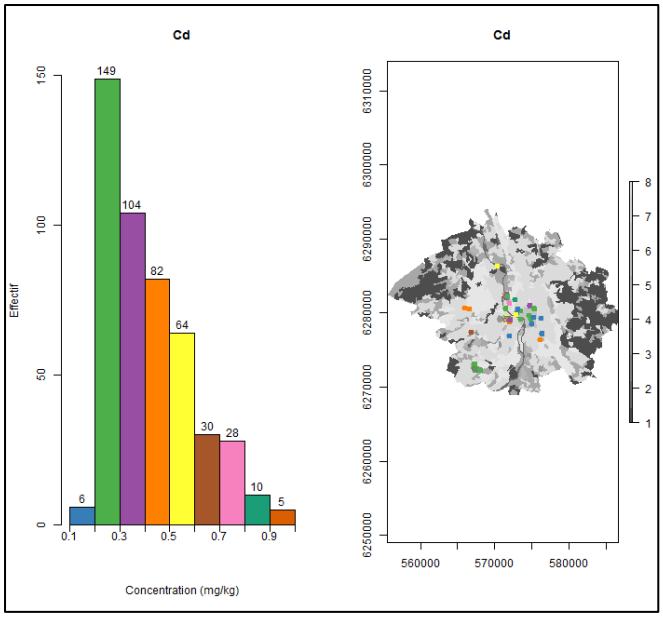
- Cd : cadmium

Le cadmium est présent en grande quantité de façon naturelle dans l'environnement. Par ailleurs, les industries émettent également des flux de déchets de cadmium, qui s'ajoutent aux quantités de cadmium présentes au naturel.

Suite à la suppression des valeurs inférieures à la LQ, il nous reste 485 prélèvements pour ce polluant. Sur ces 485 prélèvements, la concentration en cadmium est principalement comprise entre de 0 et 1 mg/kg.



Nous décidons de « zoomer » le pic de concentration afin d'obtenir une répartition plus précise des données entre 0 et 1 mg/kg.

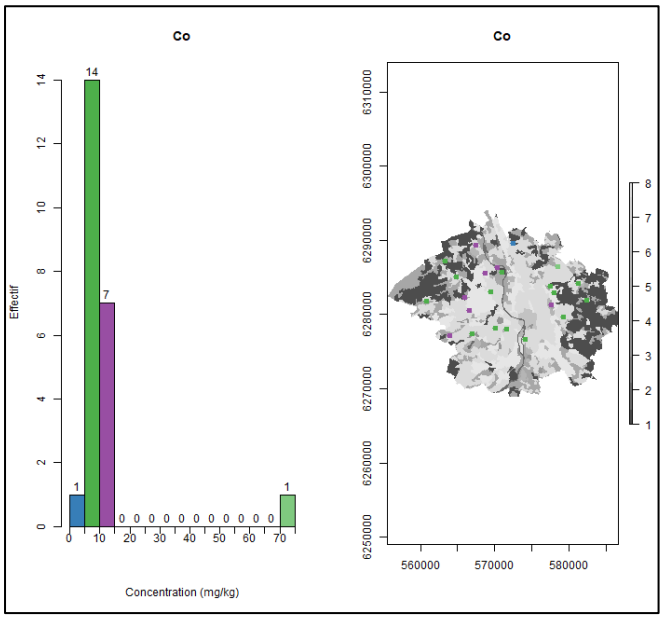


La concentration en cadmium est principalement comprise entre 0,2 et 0,6 mg/kg. Les prélèvements interprétables sont principalement situés centre de l'agglomération.

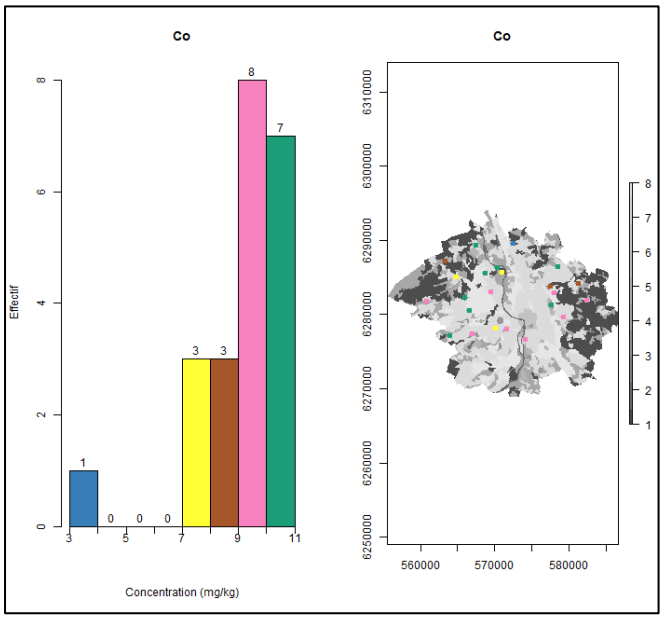
- Co : cobalt

Le cobalt est présent naturellement dans l'environnement.

Nous avons 23 prélèvements présentant une concentration allant de 0 à 75 mg/kg. Cependant, il n'est pas déraisonnable de considérer le prélèvement avec une concentration de plus de 70 mg/kg comme un potentiel outlier. Sur cette supposition, nous nous retrouvons avec des concentrations comprises entre 5 et 15 mg/kg.



Ci-dessous les résultats plus précis entre 0 et 15 mg/kg.

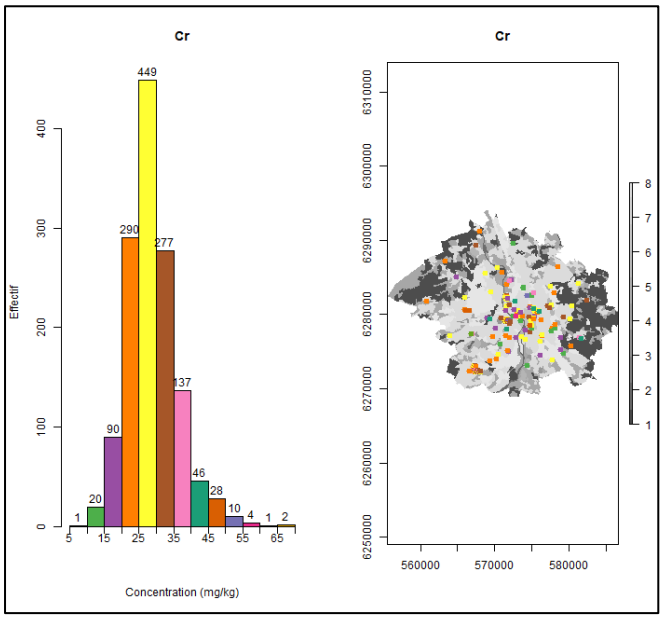


- Cr : chrome

Le chrome pénètre le sol lors des processus naturels et du fait de l'activité humaine. Dans le sol, le chrome se lie fortement aux particules du sol et ne se déplace pas jusqu'aux eaux souterraines.

La teneur moyenne du chrome dans la croûte terrestre serait comprise entre 100 et 200 mg/kg [INERIS].

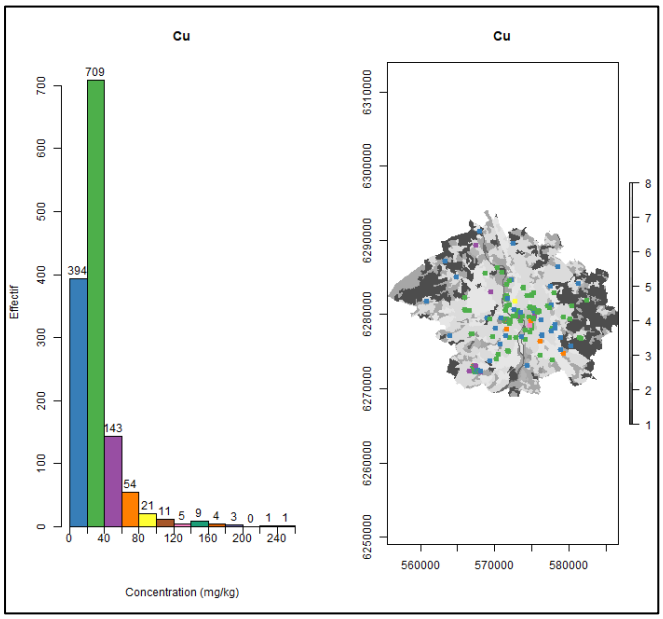
Pour le chrome, nous avons 1 355 prélèvements.



- Cu : cuivre

Tout comme le chrome, le cuivre se lie aux matières organiques et aux minéraux du sol et par conséquent ne se déplace pas jusqu'aux eaux souterraines. Le cuivre peut interrompre l'activité du sol, car il influence de façon négative l'activité des micro-organismes.

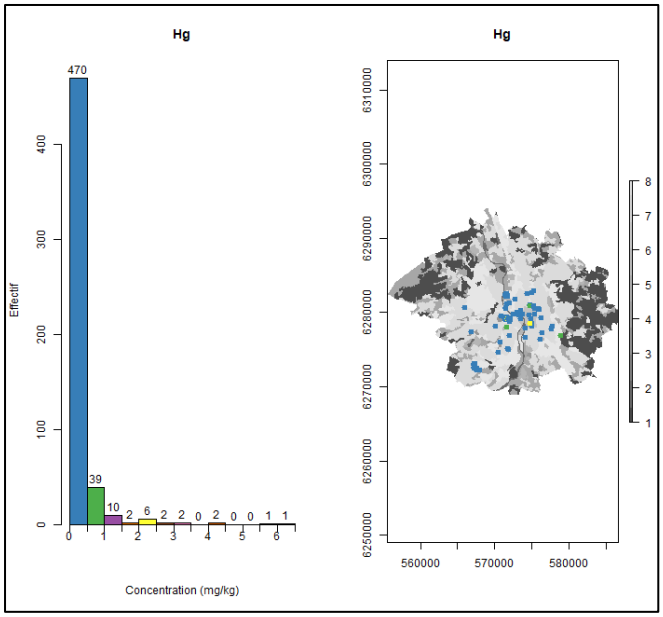
La composition moyenne de la croûte terrestre en cuivre serait comprise entre 45 et 70 mg/kg [INERIS]. Pour les prélèvements de notre jeu de données, les concentrations sont plutôt comprises entre 0 et 60 mg/kg (cf. graphiques ci-dessous).



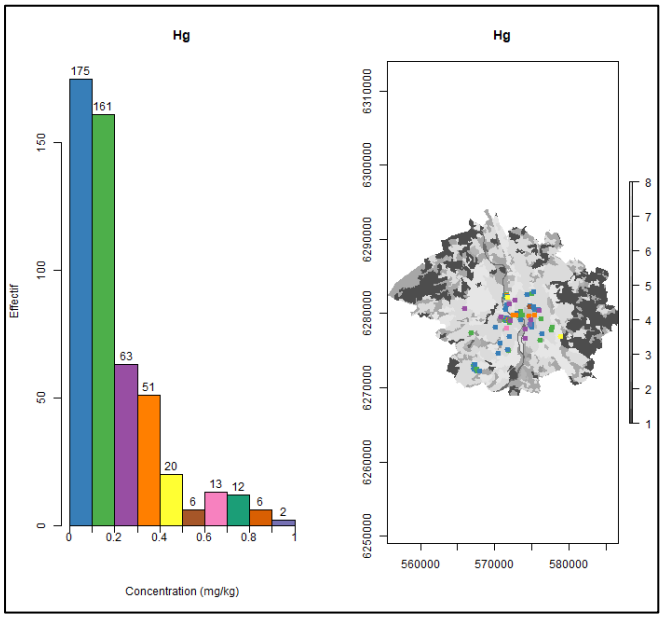
- Hg : mercure

Les roches et des sols contiennent du mercure : il est naturellement présent dans les sols entre 0,03 et 0,15 mg/kg. Dans les zones polluées, les concentrations peuvent varier de 40 à 500 mg/kg [BGRMHg].

Sur les 535 prélèvements, la concentration en mercure est principalement comprise entre de 0 et 1 mg/kg.



Nous décidons de « zoomer » le pic de concentration afin d'obtenir une répartition plus précise des données entre 0 et 1 mg/kg.

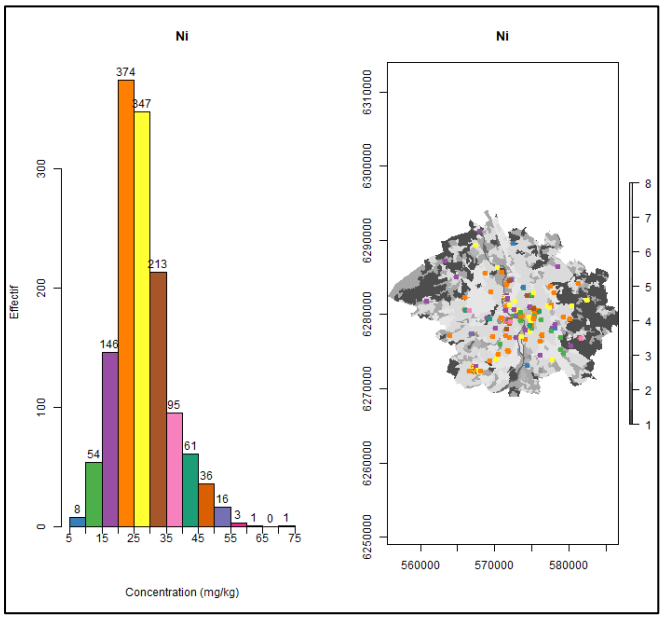


La concentration en mercure est principalement comprise entre 0 et 0,2 mg/kg. Les prélèvements interprétables sont principalement situés centre de l'agglomération.

- Ni : nickel

Le nickel est rencontré en quantités assez importantes dans la nature. La concentration en nickel des roches sédimentaires serait de l'ordre de 20 à 40 mg/kg, elle dépasserait 100 mg/kg dans les roches volcaniques [INERIS].

L'ordre de grandeur des concentrations en nickel pour notre BDD est de 15 à 40 mg/kg.

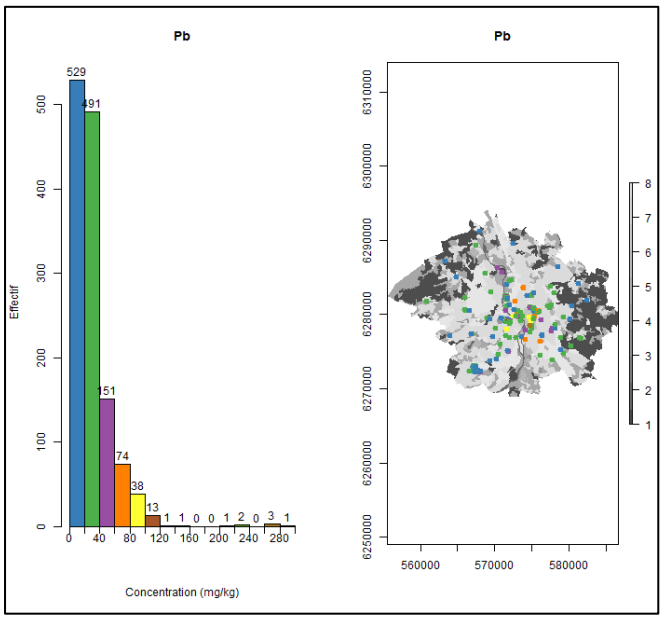


Les concentrations en nickel des prélèvements sont proches des concentrations naturellement présentes dans l'environnement. Ceci explique pourquoi ces intervalles de concentrations sont dispersés sur l'ensemble de la zone étudiée.

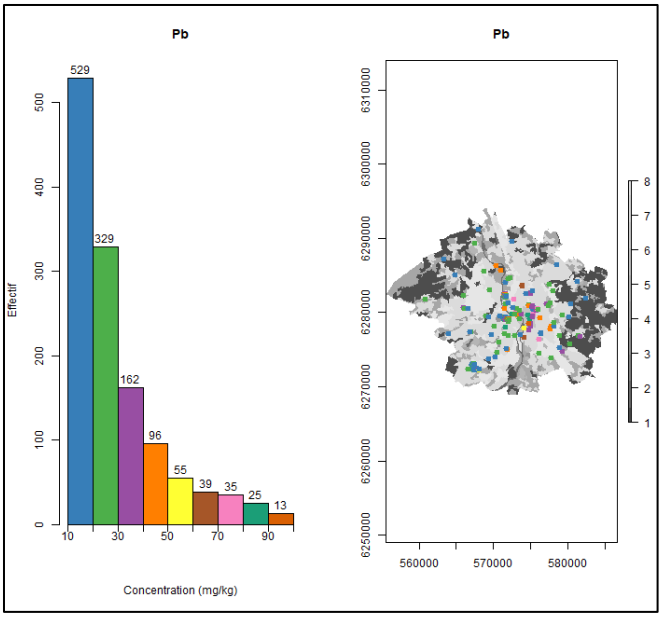
- Pb : plomb

La concentration moyenne en plomb de la croûte terrestre serait de l'ordre de 13 à 16 mg/kg [INERIS].

Sur les 1 305 prélèvements, la concentration en plomb est principalement comprise entre de 0 et 100 mg/kg.



Nous décidons de zoomer. Nous obtenons un pic de concentration en plomb entre 10 et 20 mg/kg, ce qui est de l'ordre de grandeur des concentrations naturelles.



- Se : sélénium

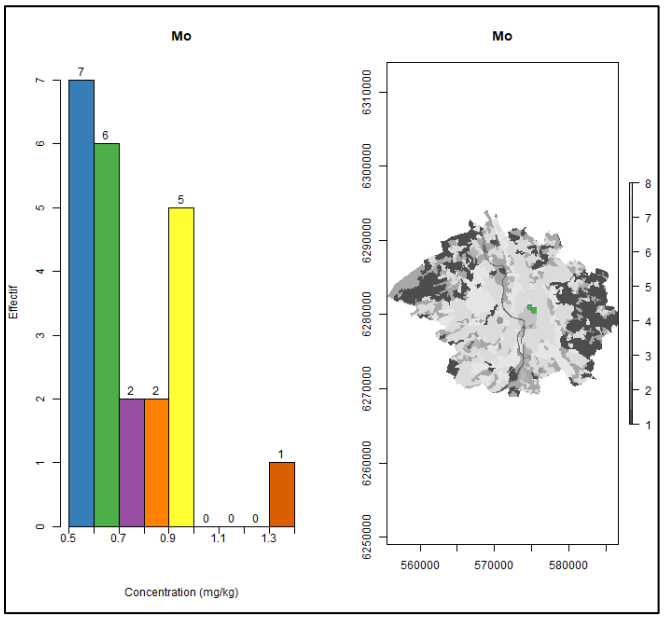
Les concentrations moyennes en sélénium dans les sols ordinaires, toutes granulométries confondues, sont de 0,10 à 0,70 mg/kg [INERISSe].

Nous n'avons aucun prélèvement exploitable (i.e. sans NA) dans la BDD. Toutes les mesures étant inférieures à la LQ selon le fichier "Fusion_Ech_150119_SO2_LD", elles sont à présent supprimées. Autrement dit, nous n'avons plus de données interprétables pour ce polluant.

- Mo : molybdène

Le molybdène est métal blanc, argenté et très dur.

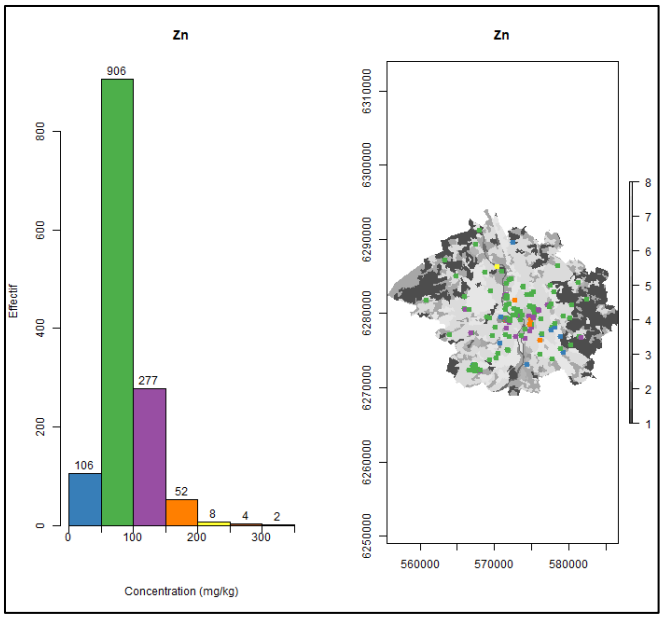
Les concentrations en molybdène restent assez faibles : elles sont comprises entre 0,5 et 1 mg/kg.



- Zn : zinc

Le zinc est un élément essentiel dans l'environnement. La teneur moyenne en zinc de la croûte terrestre serait comprise entre 70 et 132 mg/kg [INERIS].

La majorité des prélèvements a une concentration en zinc comprise entre 50 et 150 mg/kg (cf. graphique ci-dessous). Cet ordre de grandeur est proche de celui des concentrations naturelles en zinc.

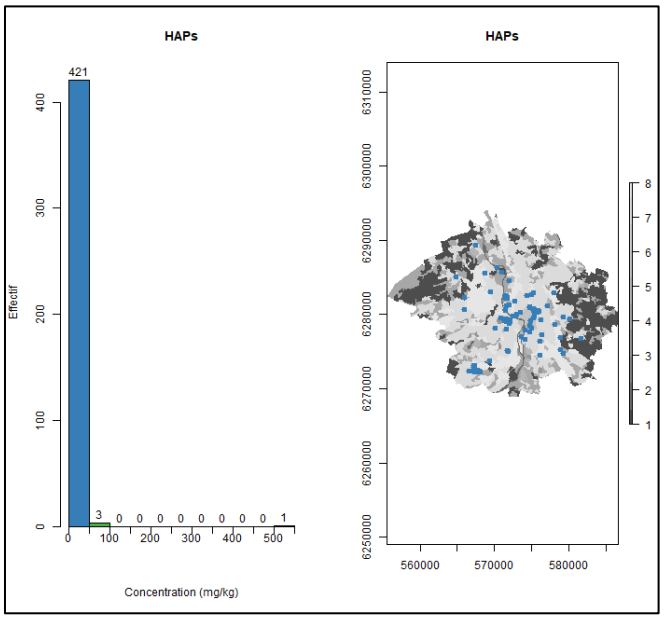


Le zinc est présent partout ; on le vérifie avec les résultats de la BDD (cf. points verts).

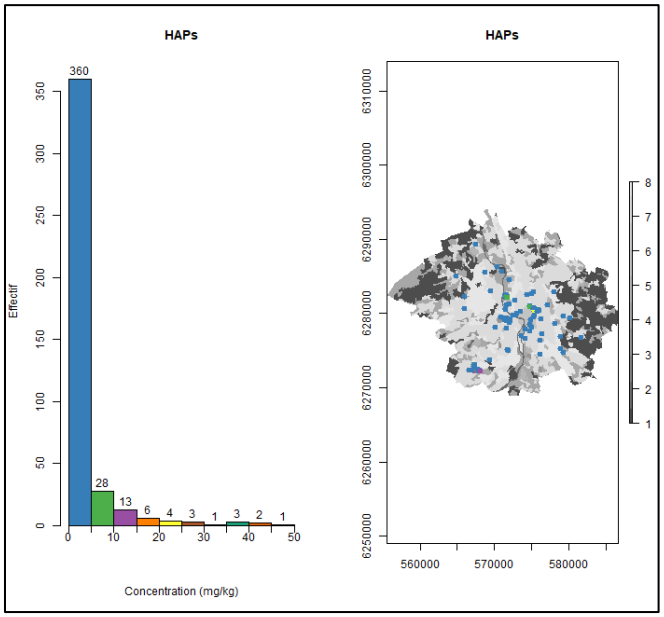
- HAPs : hydrocarbures aromatiques polycyclique

La concentration en HAPs correspond à la concentration totale des 16 hydrocarbures aromatiques polycycliques.

Les concentrations en HAPs sont principalement comprises entre 0 et 50 mg/kg.



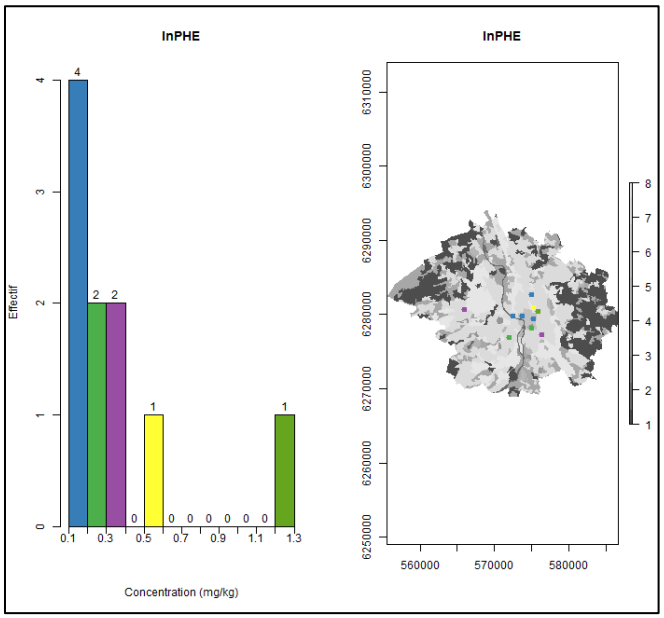
Nous décidons de zoomer cet intervalle de concentration afin d'obtenir une distribution plus précise. 85% des concentrations (360 parmi 425) sont comprises entre 0 et 5 mg/kg.



- InPHE : Indice de phénol

InPHE est l'indice de phénol dans les sols. Cette méthode ne donne qu'une information obtenue dans des conditions conventionnelles sur la teneur globale en certains dérivés du phénol.

Il ne reste que 10 prélèvements pour ce polluant. Les concentrations de ce polluant sont faibles : elles sont comprises entre 0,1 et 0,4 mg/kg.

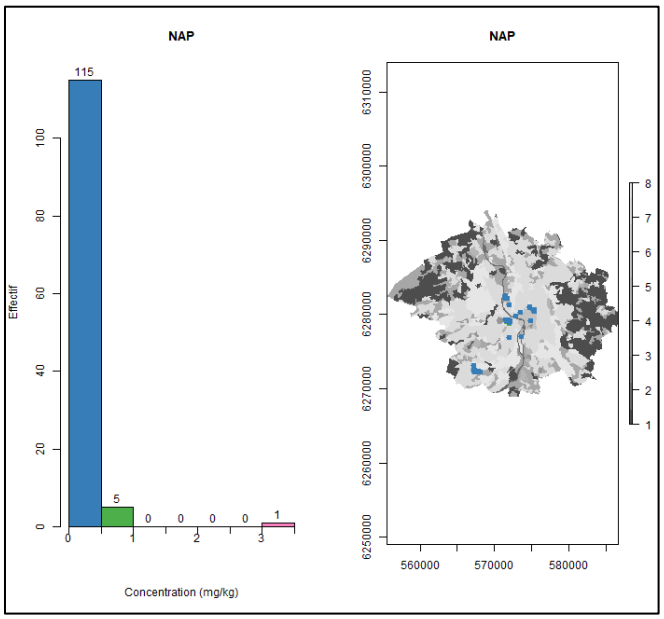


Les prélèvements ayant une concentration interprétable en phénol sont principalement au centre de l'agglomération.

- NAP : naphtalène

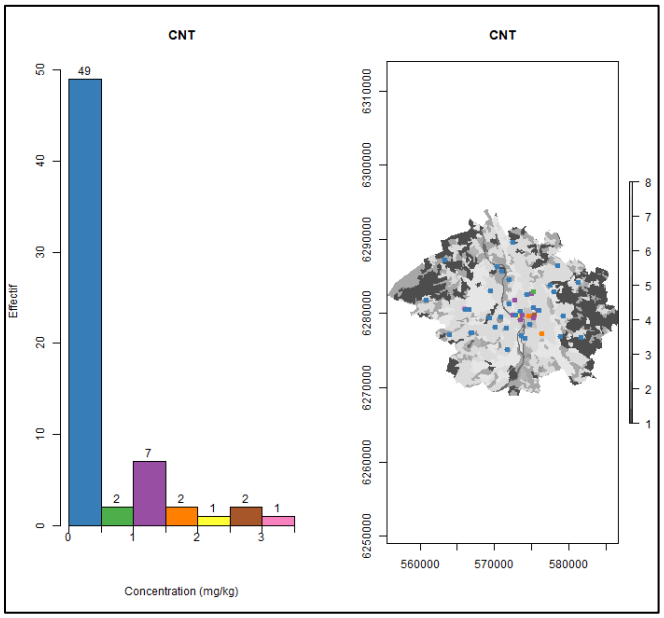
Le naphtalène fait partie des hydrocarbures aromatiques (HAP). C'est un élément très volatil généralement émis dans l'environnement lors de la combustion incomplète du bois. Il se retrouve donc principalement dans l'air. Il est classé comme cancérogène possible pour l'Homme.

Quasiment la totalité des prélèvements est supprimée : 1 235 sur 1 356, soit 91%. Les prélèvements restants ont, pour la majorité, une concentration inférieure à 0,2 mg/kg.



- CNT : concentration totale en cyanure

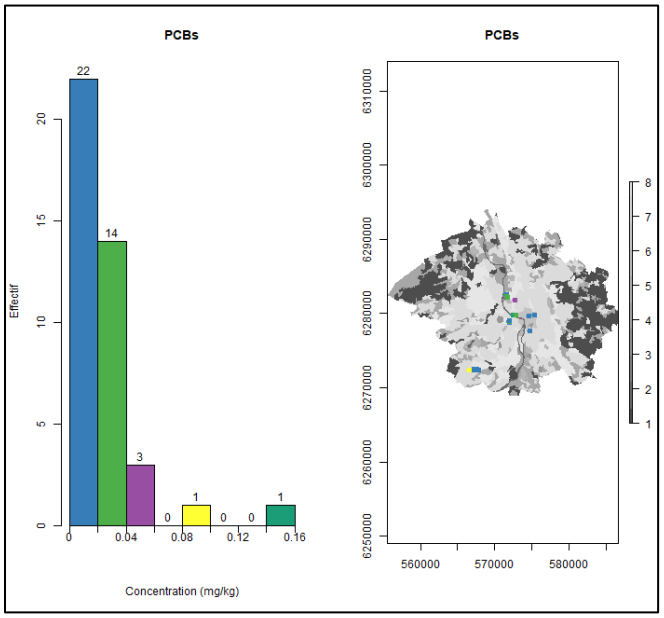
Le cyanure peut entraîner une pollution en deux temps. Transformé en gaz, il se répand dans un premier temps dans l'atmosphère. Ensuite, il réagit avec l'humidité de l'air et les pluies pour petit à petit retomber à terre et entraîner une pollution de l'eau et des sols. Sa proportion augmente avec l'acidité du sol.



Le cyanure est présent sur la totalité de la zone étudiée, majoritairement en petites concentrations.

- PCBs : polychlodobiphényles

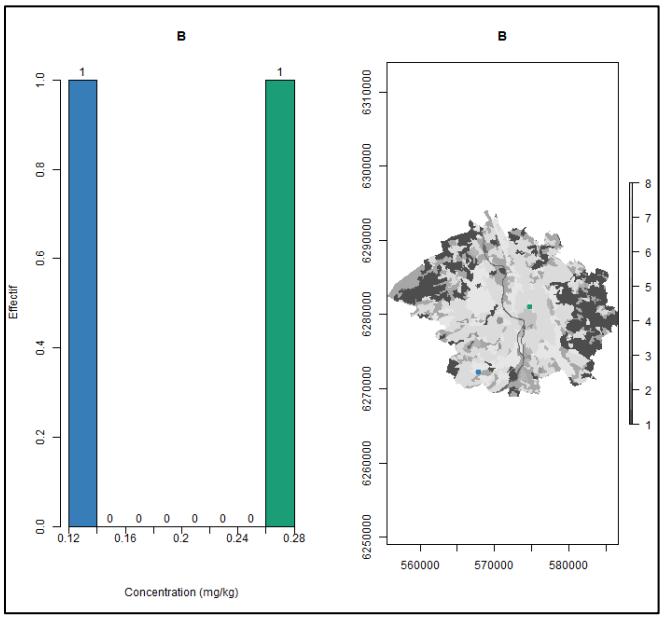
Les polychlorobiphényles sont au nombre de sept. Ce sont des polluants organiques persistants, ce qui signifie qu'ils se désagrègent très peu dans l'environnement et s'accumulent dans différents milieux, en particulier le sol.



- B : benzène

Le benzène appartient à la famille des hydrocarbures aromatiques. C'est l'un des constituants du pétrole brut.

Comme quasiment la totalité (547) des prélèvements de ce polluant présente une concentration inférieure à la LQ et sont supprimés, nous avons seulement 2 prélèvements.

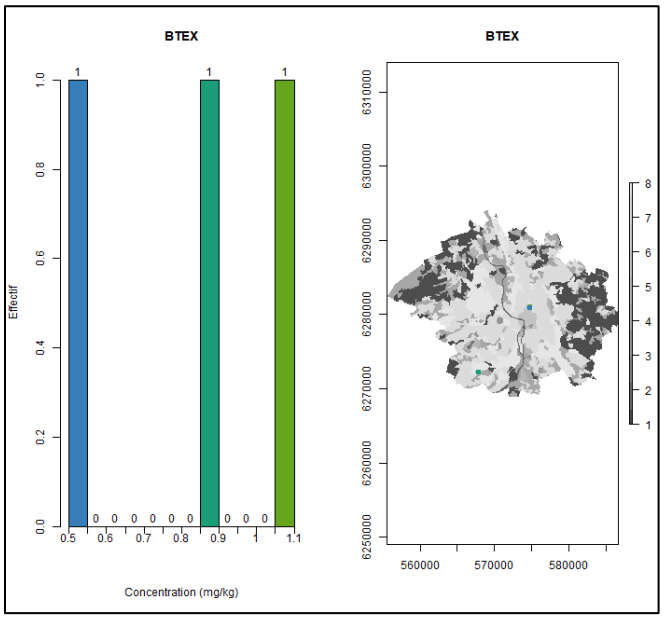


- BTEX : benzène, toluène, éthylbenzène et xylènes

Les BTEX (benzène, toluène, éthylbenzène et xylènes) sont des hydrocarbures monoaromatiques dont les origines sont majoritairement pétrolières ou pétrochimiques. Leur solubilité leur confère une bonne mobilité dans les eaux et dans les sols, pouvant entraîner d'importants impacts environnementaux du fait qu'ils deviennent facilement accessibles aux micro-organismes sous forme solubilisée.

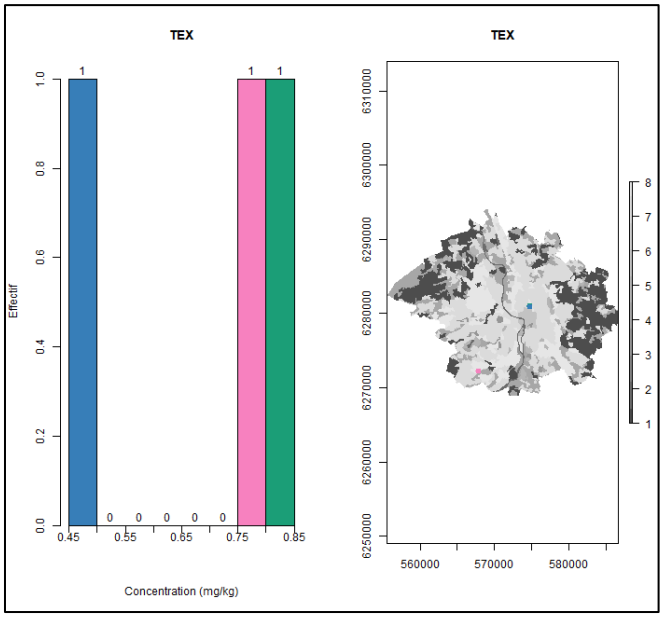
Les BTEX sont toxiques pour l'organisme humain, le benzène est cancérogène et le toluène, l'éthylbenzène et xylène sont toxiques et mutagènes.

De même que précédemment, quasiment la totalité de ces prélèvements (442 sur 445) présente une concentration inférieure à la LQ. Les concentrations de ce polluant sont donc ininterprétables.



- TEX : toluène, éthylbenzène et xylènes

Les résultats sont similaires aux précédents.

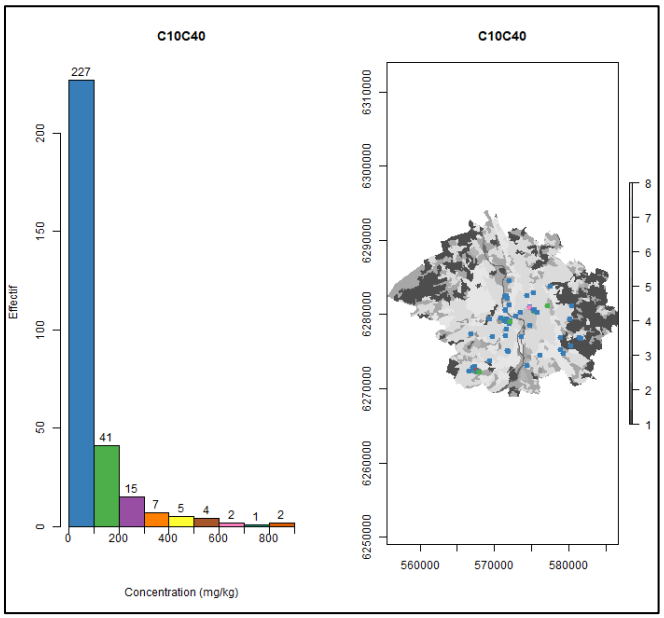


- C5C10 : hydrocarbure

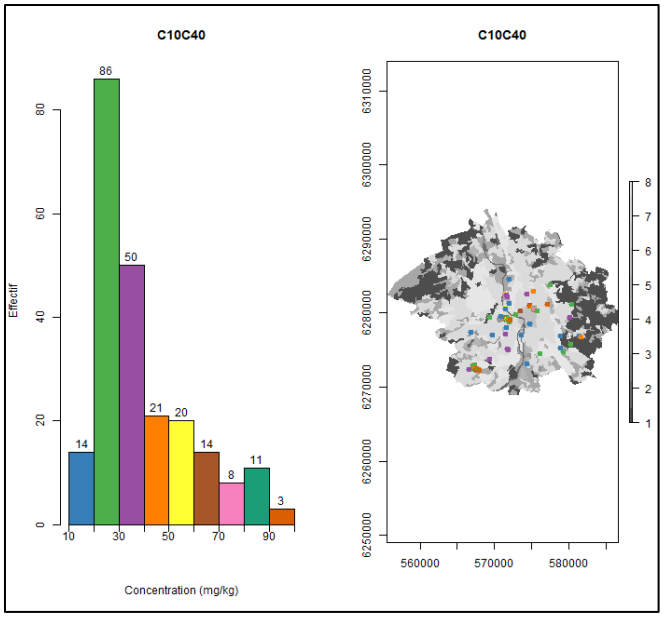
Nous n'avons aucun prélèvement exploitable (i.e. sans NA) dans la BDD. Toutes les mesures étant inférieures à la LQ selon le fichier "Fusion_Ech_150119_SO2_LD", elles sont à présent supprimées. Autrement dit, nous n'avons plus de données interprétables pour ce polluant.

- C10C40 : hydrocarbure

Les concentrations en C10C40 sont principalement comprises entre 0 et 100 mg/kg (cf. graphique ci-dessous).

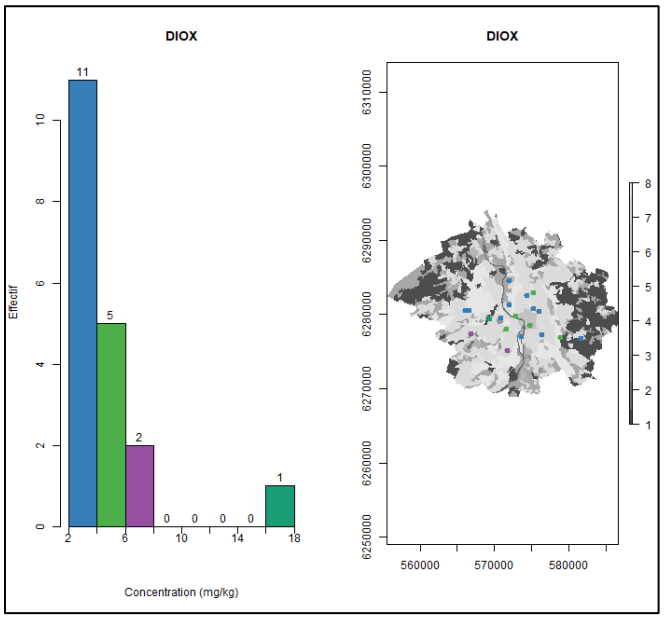


Nous décidons de zoomer cet intervalle de concentration afin d'obtenir une distribution plus précise.



- DIOX : dioxines

Les dioxines sont des polluants organiques persistants.



II. Exploitation des résultats

L'objet du rapport est de valider une méthode de calcul des seuils de pollution parmi quatre proposées : Centile 90%, Centile 95%, MED ± 2.MAD et Vibrisse interne supérieure du boxplot de Tukey. Notre objectif est de réaliser une comparaison critique des résultats obtenus avec chaque méthode, accompagnée d'éventuels commentaires.

Cette démarche nous amène à calculer des « seuils de pollution » dans le but d'établir quelle est la méthode la plus appropriée au besoin du BRGM ; le calcul de ces seuils ne sert qu'à la comparaison des méthodes, il n'est en aucun cas l'objet du contrat.

Nous travaillons avec les données triées, c'est-à-dire sans les outliers.

A. Présentation des méthodes

1. Méthodes 1 et 2 : Centile 90% et 95%

Le centile est un nombre, une valeur observée, qui partage la série ordonnée en deux sous-ensembles. Les valeurs du premier sous-ensemble sont inférieures à la valeur observée, et celles du second sous-ensemble sont supérieures à cette valeur.

Le centile 90% est tel que 90% des données se trouvent en dessous de cette valeur et 10% au-dessus.

Cette méthode suppose qu'il existe toujours un certain pourcentage d'outliers au sein d'un jeu de données. En effet, le pourcentage se trouvant au-dessus de la valeur observée est considéré comme contenant principalement des outliers.

C'est une méthode simple et efficace qui n'a besoin d'aucun postulat. Néanmoins, pour l'appliquer, il faut un nombre minimal de données. Par convention, on retient un nombre d'observations $n \geq 30$.

La définition du centile 95% est sensiblement la même. La seule différence est qu'au lieu d'avoir 90% des données en dessous de la valeur observée, il y en a 95% (et donc 5% au-dessus).

2. Méthode 3 : MED + 2MAD

a. Principe de la méthode

Cette méthode fait intervenir la médiane et la MAD (cf. le paragraphe b. suivant). C'est une approche robuste facilement compréhensible et largement utilisée pour évaluer un seuil dans les études géochimie environnementales.

MED correspond à médiane. La médiane est le centile 50%. Dans ce cas, les deux sous-ensembles ont des dimensions égales : la moitié des valeurs se trouvent en dessous de la valeur observée et l'autre moitié au-dessus.
MAD correspond à Median Absolute Deviation. Autrement dit, c'est la médiane des écarts absolus à la médiane. Elle est définie de la façon suivante :

$$MAD(X) = 1{,}4826 \times mediane(|X - mediane(X)|)$$

Le facteur 1,4826 est une constante. C'est un facteur d'échelle dépendant de la distribution des données (1,4826 correspond au facteur utilisé pour la distribution normale). Il est possible d'utiliser la valeur $1/Q75\%$ comme facteur d'échelle, afin de mieux s'adapter aux données.

La méthode "médiane ± 2.MAD" est sensible à l'ajout et au retrait de valeurs, en particulier aux valeurs extrêmes. En effet, pour pouvoir l'appliquer, il faut vérifier que moins de 50% des valeurs sont des valeurs extrêmes, sinon même les paramètres robustes tels que la médiane ou la MAD se retrouvent influencés par les valeurs extrêmes. Notre jeu de données présente beaucoup de données extrêmes, rendant l'application de cette méthode délicate.

De plus, cette méthode repose sur une hypothèse de normalité, si c'est le facteur 1,4826 qui est utilisé lors des calculs. Enfin, comme pour la méthode précédente, il faut un nombre de données suffisamment grand ($n \geq 30$).

La littérature scientifique concernant cette méthode recommande une transformation logarithmique avant son utilisation pour les données en géochimie environnementale asymétriques positives. Cependant, nous ne préconisons pas les transformations de données, nous préférons analyser les données brutes, c'est-à-dire sans transformations.

b. Analyse critique de la méthode "MED + 2MAD"

La médiane désigne le nombre m tel qu'il y ait autant de valeurs de X au-dessus qu'au-dessous. Cela n'a de sens que pour une variable aléatoire réelle (et cela n'en a pas pour une variable complexe, ou qualitative, par exemple). Il y a un problème de définition si le nombre de valeurs est pair : on se retrouve avec un intervalle (en général, on prend le milieu de cet intervalle).

Une fois la valeur de m déterminée, elle est invariante par éloignement des valeurs de X : les valeurs au-dessus de m peuvent s'éloigner à droite, les valeurs au-dessous peuvent s'éloigner à gauche, sans que cela change la médiane.

On peut dire grossièrement que la médiane est un concept qualitatif, permettant de partager les valeurs de X en deux classes d'égale importance ; ce n'est pas un concept quantitatif.

La MAD (Median Absolute Deviation) est définie par : $\mathrm{MAD} = cst * med|X - m|$.

Commençons par le cas simple où $m = 0$: X a autant de valeurs positives que négatives.

On regarde $|X|$; les valeurs sont d'une part les $x_i > 0$ et d'autre part les $y_i = -x_i$ pour les $x_i < 0$. La MAD sera le nombre A tel qu'il y ait autant de x_i, y_i de part et d'autre. Mais :

— Être au-dessus de A signifie pour X être $< -A$ ou $> A$;
— Être entre 0 et A signifie pour X être dans l'intervalle $-A, A$.

Par conséquent, la MAD est le nombre A tel que (pour X) il y ait autant de valeurs dans l'ensemble $]-\infty, -A] \cup [A, +\infty[$ que dans l'intervalle $[-A, A]$.

Si on revient au cas général, il faut appliquer ceci à $X - m$.

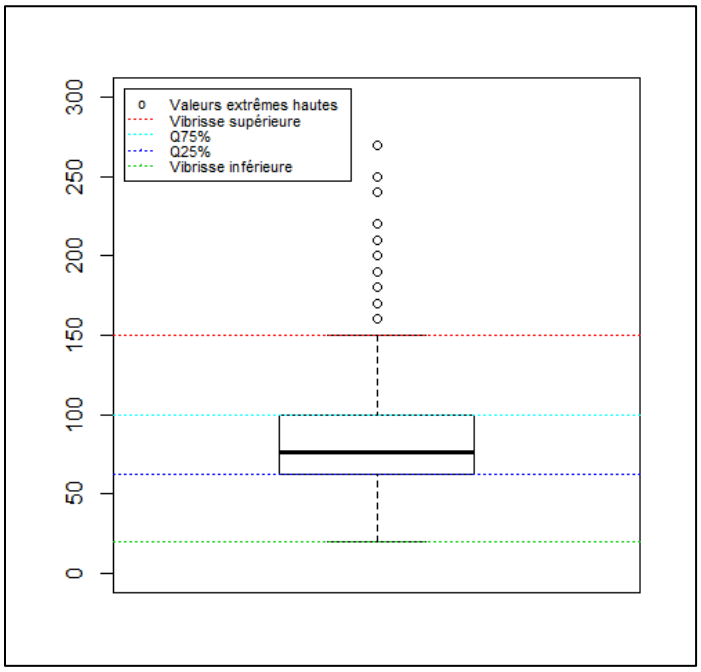
Concrètement, on part de $X - m$ et on regarde $X - m \pm \alpha$; on commence avec $\alpha = 0$ et on augmente $\alpha > 0$ progressivement. On s'arrête quand le nombre de points dans l'intervalle $[X - m - \alpha, X - m + \alpha]$ est égal au nombre de points hors de cet intervalle.

Ce n'est pas une vraie mesure de dispersion. On peut éloigner les points les plus grands ou les plus petits sans changer le résultat. Plus généralement, la MAD est invariante par modification des valeurs à l'intérieur des intervalles. Les points du premier intervalle peuvent être tous mis à m et les points du second ensemble éloignés indéfiniment (et même remplacés par leurs opposés) sans que cela change. A.

Aucune méthode à vocation quantitative, visant en particulier à fixer un seuil, ne peut reposer sur une utilisation de la médiane, puisque celle-ci est de nature qualitative.

3. Méthode 4 : Vibrisse interne supérieure du boxplot de Tukey

Le boxplot (ou boîte à moustaches ou diagramme en boîte ou boîte de Tukey) est une représentation graphique des statistiques de base (minimum, maximum, quantiles, médiane, valeurs extrêmes). Les vibrisses correspondent aux extrémités des moustaches (cf. graphique ci-dessous).



Cette méthode permet d'identifier les valeurs extrêmes hautes. La définition de la vibrisse interne supérieure est la suivante :

$$Vibrisse(X) = Q75\% + 1{,}5 \times (Q75\% - Q25\%)$$

La vibrisse est une méthode robuste, mais son calcul est basé sur la symétrie de la distribution des données. De plus, comme toutes les autres méthodes, il faut un nombre de données assez important ($n \geq 30$).

4. Robustesse des méthodes

Deux méthodes parmi nos quatre sont basées sur des estimateurs robustes, mais elles donnent de meilleurs résultats suite à des transformations. Ainsi, étant sensibles aux transformations, elles ne peuvent être considérées comme robustes à part entière.

> "*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*" [BRGM, p34].

Les données log-transformées n'ayant rien à voir avec les données originales, il est difficile de transposer les résultats obtenus aux données originales.

Ceci explique pourquoi nous travaillons qu'avec des données brutes. En effet, les transformations, et la log-transformation en particulier, peuvent être assez problématiques. Elles entraînent souvent un biais dans le résultat à interpréter : les données logarithmiques transformées ne peuvent généralement pas faciliter les interprétations concernant les données originales, puisqu'elles ont peu en commun avec celles-ci [Log].

Si une transformation est utilisée, il faut être attentif à ses limites, en particulier lorsqu'il s'agit d'interpréter la pertinence de l'analyse des données transformées pour l'hypothèse de départ concernant les données originales. Par exemple, une étude [Log] a démontré que, dans la plupart des cas, la log-transformation n'aide pas à rendre les données moins variables ou plus normales et peut, dans certaines circonstances, rendre les données plus variables et encore plus faussées. De plus, parfois, quelle que soit la transformation appliquée aux données brutes, il est quasiment impossible de parvenir à une distribution normale.

En plus de reposer sur la théorie de la loi normale [BRGM, p77], le calcul de la méthode de la vibrisse repose sur une hypothèse de symétrie de distribution qui n'est pas vérifiée ici. Plus une méthode a d'hypothèses, moins elle est robuste : il suffit que l'une d'entre elles ne soit pas vérifiée pour l'affaiblir.

Les méthodes des centiles sont plus robustes que les précédentes, car on peut les utiliser sur les données brutes et parce qu'elles ne sont pas sensibles à la présence de valeurs aberrantes extrêmes ou non.

5. Nombre d'échantillons

Pour ce rapport, un échantillon correspond à un prélèvement.

L'une des conditions d'utilisation de ces quatre méthodes est $n \geq 30$. C'est le nombre d'échantillons pratique sur lequel se sont arrêtés les statisticiens en début de siècle pour pouvoir utiliser l'hypothèse de normalité. Ce n'est pas une norme, il s'agit d'un consensus entre statisticiens.

En effet, pour un nombre d'échantillons supérieur à 30 et par application du théorème central limite, la moyenne de n variables aléatoires i.i.d (variables indépendantes et identiquement distribuées) converge vers une loi normale. Ainsi les conclusions obtenues à partir de ces échantillons sont identiques à celles résultant d'échantillons issus d'une population normale de mêmes paramètres que la population étudiée (normale ou non). Au contraire, si le nombre d'échantillons est inférieur à 30, les statistiques relatives aux paramètres de ces échantillons ne correspondent plus à celles des distributions normales et l'erreur standard de la moyenne augmente fortement. Dans ce cas, une bonne approximation de la distribution réelle est la loi de Student.

De plus, 30 est la limite inférieure acceptable de l'effectif des populations pour qu'un fond géochimique puisse être déterminé [BRGM p14].

Comme pour deux des méthodes l'hypothèse de normalité reste nécessaire (cf. § II.A.4), nous gardons la limite d'échantillon fixée à 30 en ce qui concerne leurs conditions d'utilisation. Néanmoins cette condition n'est pas toujours vérifiée, car le nombre de prélèvements du BRGM est lié au nombre d'établissements scolaires retenu par la démarche ETS [BRGM p14].

B. Seuils obtenus pour les quatre méthodes

Les données utilisées sont les prélèvements de la BDD du 15 janvier 2019 sans les outliers identifiés par le BRGM (OUT = 1) et avec les concentrations inférieures à la LQ supprimées.

Les calculs sont réalisés sur l'ensemble des données, c'est-à-dire avec toutes les profondeurs confondues. En effet, les distributions des concentrations en fonction des profondeurs sont globalement les mêmes (cf. Annexe 1).

Afin de s'affranchir des hypothèses de normalité, pour la méthode MED+2MAD nous n'utilisons pas le facteur normal, mais 1/Q95%.

Les seuils en mg/kg trouvés avec les quatre méthodes présentées précédemment sont dans le tableau ci-dessous pour chacun des polluants présents dans la base de données.

Polluant	Nombre de prélèvements	Q90% (seuil en mg/kg)	Q95% (seuil en mg/kg)	MED+2MAD (seuil en mg/kg)	Vibrisse (seuil en mg/kg)
As	1349	24	27	15,42	29,5
Sb	8	4,21	5,85	2,42	4,15
Ba	184	140	160	73,36	157,38
Cd	485	0,7	0,77	0,84	0,9
Co	23	10,8	11	9,2	13
Cr	1355	37	41,3	28,25	44
Cu	1355	53	73,3	25,41	57,75
Hg	535	0,63	0,95	0,6	0,63
Ni	1355	38	43	26,32	44,5
Pb	1305	60	77	22,43	68,5
Se	0				
Mo	23	0,94	0,95	1,01	1,38
Zn	1355	130	140	76,34	157
HAPs	425	8,98	17,4	1,48	5,75
InPHE	10	0,59	0,94	0,67	0,6
NAP	121	0,2	0,38	0,33	0,13
CNT	64	1,3	1,97	0,87	0,73
PCBs	41	0,04	0,05	0,55	0,06
B	2	0,26	0,27	0,87	0,36
BTEX	3	1,06	1,08	1,31	1,45
TEX	3	0,81	0,82	0,9	1,07
C5C10	0				
C10C40	304	237	374	40,4	212,5
DIOX	19	6,96	8,28	3,54	6,28

Les résultats sont d'ordre de grandeur proche pour les centiles et la vibrisse. La méthode MED+2MAD quant à elle procure des résultats assez différents des autres méthodes.

Ces méthodes sont applicables à condition d'avoir un échantillon assez conséquent ($n \geq 30$). Or, suite à la suppression des prélèvements dont les concentrations des polluants sont inférieures à la LQ, il y a beaucoup de polluants qui se retrouvent avec un nombre insuffisant de prélèvements (cf. données en rouge dans le tableau). Il faut rester vigilant lors de l'interprétation de ces résultats.

C. Proportion des prélèvements supérieurs aux seuils

À partir des seuils précédemment obtenus, nous définissons quelle méthode est la plus "punitive". Autrement dit, nous déterminons la proportion de prélèvements pollués pour un polluant et une méthode donnés. Ces proportions sont présentées dans le tableau ci-dessous.

Traitons l'exemple de l'arsenic : sur 1 349 prélèvements,

- 10% dépassent le seuil pour le test "Centile 90%" ;
- 5% dépassent le seuil pour le test "Centile 95%" ;
- 48% dépassent le seuil pour le test "MED+2MAD" ;
- 3% dépassent le seuil pour le test " Vibrisse interne supérieure".

Polluant	% > seuil			
	Centile 90%	Centile 95%	MED+2MAD	Vibrisse interne supérieure
As	10	5	48	3
Sb	12	12	38	12
Ba	8	4	49	7
Cd	9	5	3	2
Co	13	4	35	4
Cr	10	5	43	3
Cu	10	5	47	9
Hg	10	5	11	10
Ni	10	5	46	4
Pb	10	5	50	7
Se				
Mo	9	4	4	4
Zn	7	5	49	3
HAPs	10	5	42	14
InPHE	10	10	10	10
NAP	9	5	6	15
CNT	9	6	20	22
PCBs	10	7	0	5
B	50	50	0	0
BTEX	33	33	0	0
TEX	33	0	0	0
C5C10				
C10C40	10	5	46	11
DIOX	11	5	47	16

Remarque : pour les centiles Q90% et Q95%, nous devrions avoir respectivement 10% et 5% de valeurs supérieures au seuil. Or, nous obtenons des pourcentages supérieurs. Ceci s'explique par le nombre d'ex-æquos assez important pour certains polluants.

La méthode la plus "punitive" est sans conteste la "MED+2MAD". Elle reste la plus punitive même avec le facteur normal (cf. le rapport d'avancement "SCM_BRGM_R1_2018_12_05").

La vibrisse interne supérieure semble aussi performante que les centiles. Néanmoins, en plus de l'hypothèse de normalité, cette méthode repose sur l'hypothèse de symétrie des distributions. Or, comme nous n'appliquons aucune transformation aux données, aucune distribution n'est symétrique. Les conditions d'utilisation de cette méthode ne sont donc pas vérifiées.

Le choix d'une méthode se porte donc sur l'un des deux centiles. Les résultats montrent que ces deux méthodes sont robustes et peuvent être utilisées pour déterminer un seuil de pollution. Quant à savoir quel centile choisir, il s'agit du ressort du BRGM : soit ils préfèrent être punitifs, et choisiront le Q90%, soit ils veulent l'être un peu moins, et choisiront le Q95%. Il est aussi possible d'utiliser le centile 90% pour les polluants les plus nocifs et le centile 95% pour les autres.

D. Comparaison avec la variabilité naturelle

La comparaison n'est possible que pour les polluants dits « éléments à traces métalliques », car pour les autres polluants (hydrocarbures, polluants persistants, etc.), nous n'avons pas d'information sur les valeurs de concentration naturelle.

Nous comparons les seuils obtenus à partir des centiles aux concentrations naturelles présentes dans la littérature [BGEP, BRGMAs, BRGMHg, INERIS, INERISInfo] dans le tableau ci-dessous.

Polluant	Q90% (seuil en mg/kg)	Q95% (seuil en mg/kg)	Concentration naturelle (mg/kg)	Conclusion
As	24	27	1 à 50	Seuils inclus dans la variabilité naturelle
Sb	4,21	5,85	0,2 à 10	Seuils inclus dans la variabilité naturelle
Ba	140	160	562	Seuils inférieurs à la variabilité naturelle
Cd	0.7	0.77	0,05 à 0,45	Seuils supérieurs à la variabilité naturelle
Co	10.8	11	1 à 40	Seuils inclus dans la variabilité naturelle
Cr	37	41.3	10 à 90	Seuils inclus dans la variabilité naturelle
Cu	53	73.3	6 à 30	Seuils supérieurs à la variabilité naturelle
Hg	0.63	0.95	0,02 à 0,15	Seuils supérieurs à la variabilité naturelle
Ni	38	43	2 à 60	Seuils inclus dans la variabilité naturelle
Pb	60	77	9 à 50	Seuils supérieurs à la variabilité naturelle
Se			0,10 à 0,70	-
Mo	0.94	0.95	2	Seuils inférieurs à la variabilité naturelle
Zn	130	140	10 à 300	Seuils inclus dans la variabilité naturelle

Tous les cas de figure sont présents : soit les seuils sont compris dans l'intervalle de concentration naturelle, soit ils sont inférieurs (en vert), soit ils sont nettement supérieurs (en rouge). Les seuils Q90% et Q95% obtenus pour :

– Mo, Ba sont inférieurs à la variabilité naturelle ;
– As, Sb, Co, Cr, Ni et Zn sont inclus dans la variabilité naturelle ;
– Cd, Cu, Hg et Pb sont supérieurs à la variabilité naturelle.

Il est difficile d'utiliser les concentrations naturelles comme référence, d'autant que nous ne les avons pas pour tous les polluants. De plus, elles ne permettent pas de faire un choix entre les deux méthodes. Toutefois, si les seuils calculés sont nettement supérieurs aux concentrations naturelles, il peut être judicieux d'envisager une dépollution pour les polluants concernés.

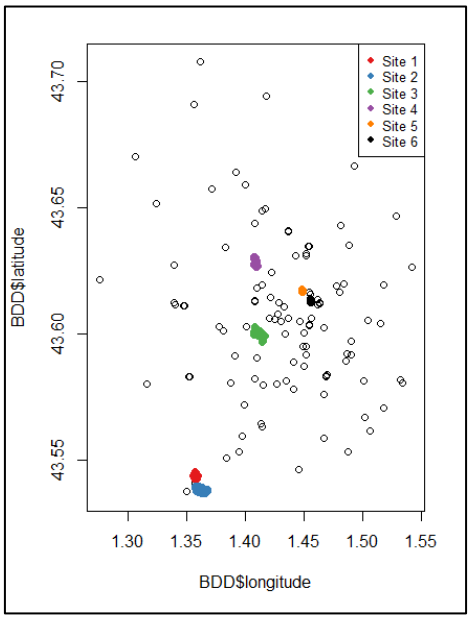
III. Influence et recommandation des sites

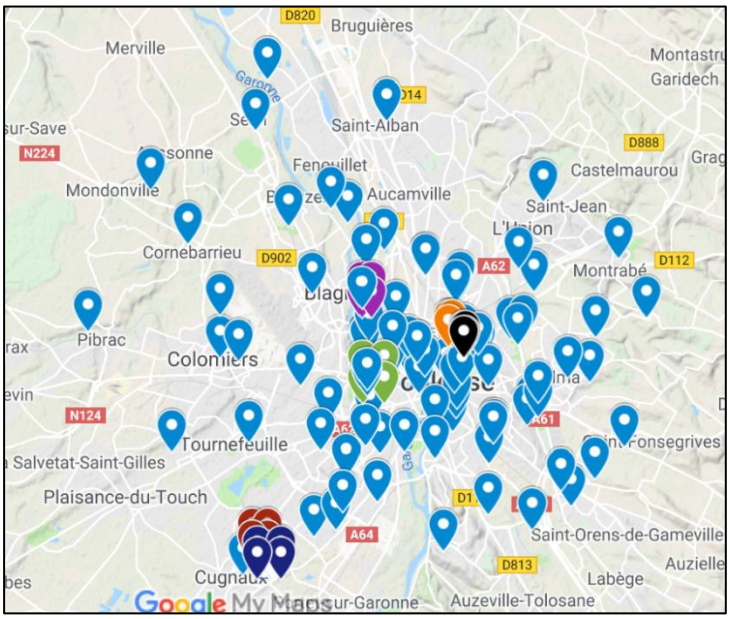
Pour cette partie nous travaillons toujours avec la même base de données BDD (sans outliers et avec les concentrations inférieures à la LQ supprimée).

Les prélèvements ne sont pas réalisés de façon uniforme sur l'ensemble de l'agglomération. En effet, certaines surfaces sont plus sondées que d'autres : nous les appelons « sites ». Cela est en lien avec une suspicion de pollution ou un souhait de construction sur un site en particulier.

Par exemple, un site au sud de l'agglomération (ancien site militaire) a été échantillonné avec plus de 100 forages, ce qui correspond à peu près à l'ensemble des forages des parcs et jardins de toute l'agglomération. Ce site militaire est probablement le « site 2 », représenté par les points bleus sur le graphique ci-dessous. Nous supposons qu'autant de forages ont été effectués sur ce site connu comme pollué afin de trouver les caractéristiques d'un site pollué.

Nous identifions 6 sites différents au vu des regroupements géographiques de forages (cf. graphique ci-dessous).



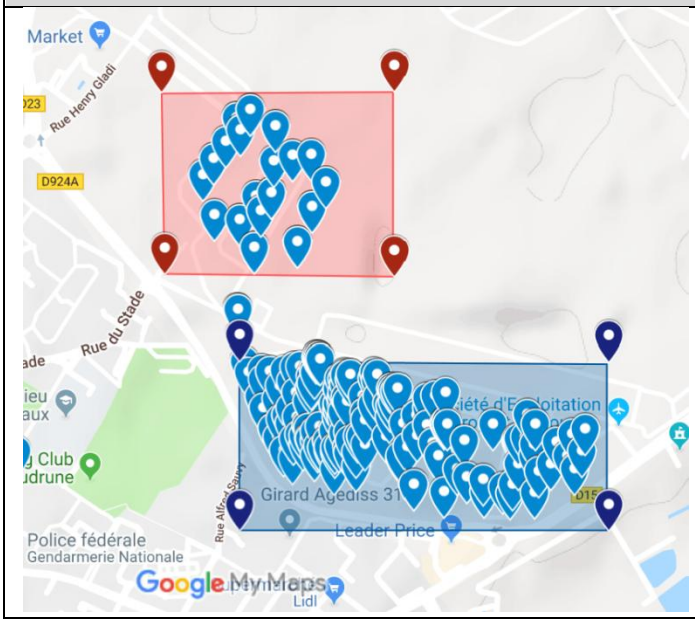


Lorsque plusieurs forages se trouvent à proximité et que cet ensemble de forages est lui-même isolé des autres forages, nous définissons cet ensemble comme un nouveau site.Le tableau ci-dessous donne la répartition des 1 356 prélèvements selon le site auquel ils appartiennent.

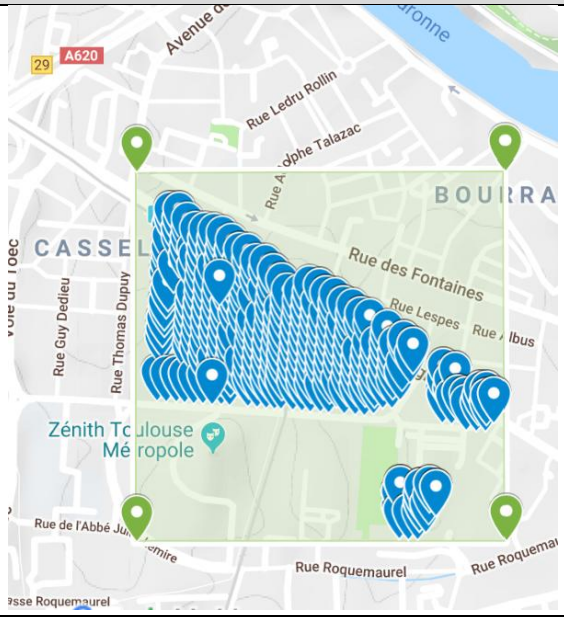
Site	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Aucun site
Nombre de prélèvements	47	312	562	80	31	112	212

Ci-dessous, la délimitation des 6 sites identifiés via GoogleMap.

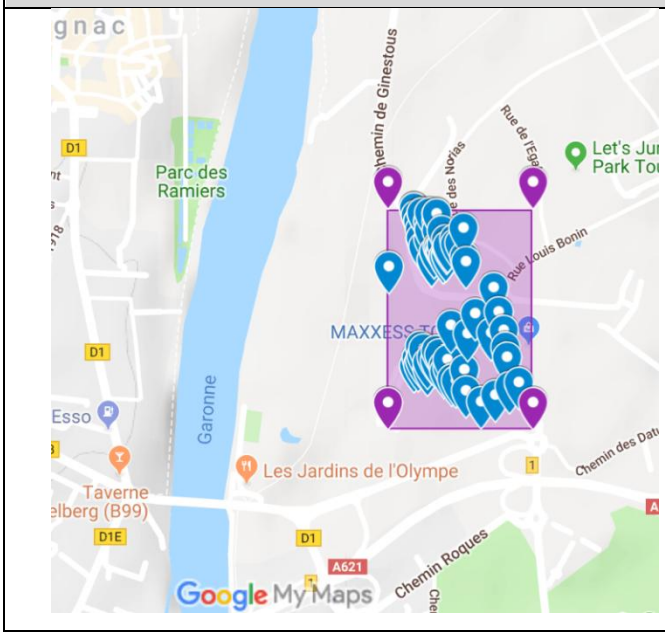
Sites 1 et 2



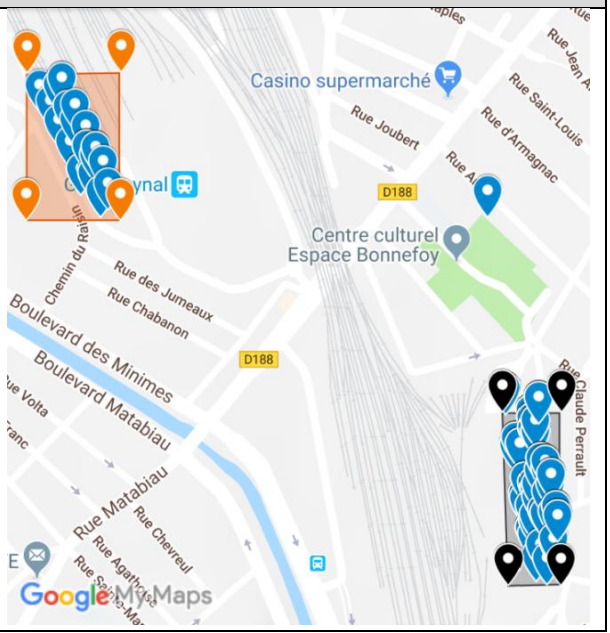
Site 3



Site 4



Sites 5 et 6



A. Influence des sites sur les seuils

Pour pouvoir caractériser de manière efficace la concentration des polluants dans le sol ou pour mettre en évidence des différences selon les zones géographiques, il aurait fallu un échantillonnage uniforme de l'agglomération.

Ici, le surplus de prélèvements sur quelques sites peut biaiser les résultats : si plus de prélèvements proviennent d'un site connu comme pollué que d'un site neutre, les seuils calculés auront tendance à être plus élevés que la normale. Pour parer à ce problème, nous décidons de moyenner les concentrations des sites de sorte à n'avoir qu'un seul prélèvement par site. Cette démarche permet de vérifier si les conditions d'échantillonnages impactent réellement les résultats obtenus concernant le calcul des seuils par la méthode des centiles.
Une nouvelle base de données BDDS est créée. Elle contient tous les prélèvements n'appartenant à aucun site (selon leur disposition géographique) ainsi qu'un unique prélèvement par site (qui se trouve être la moyenne de tous les prélèvements des sites).

Données	Description	Nombre total de prélèvements
BDD	Données de l'étude	1356
BDDS	Données de BDD avec les données des sites moyennées (i.e. un seul prélèvement par site)	218

Pour déterminer si les prélèvements des sites influencent les calculs, nous calculons les nouveaux seuils (cf. tableau ci-dessous). Les calculs ont également été réalisés avec les données dépourvues des données concernant l'ensemble des sites précédemment définies. Les résultats restent alors similaires à ceux de BDDS.

Polluants	BDD			BDDS		
	Effectif	Q90%	Q95%	Effectif	Q90%	Q95%
As	1349	24	27	213	23	27
Sb	8	4,21	5,85	2	2,75	2,8
Ba	184	140	160	137	150	170
Cd	485	0,7	0,77	35	0,7	0,83
Co	23	10,8	11	23	10,8	11
Cr	1355	37	41,3	218	44,3	48
Cu	1355	53	73,3	218	48,15	66,3
Hg	535	0,63	0,95	74	0,7	2,27
Ni	1355	38	43	218	37	42
Pb	1305	60	77	209	74	106
Se	0			0		
Mo	23	0,94	0,95	2	0,74	0,74
Zn	1355	130	140	218	120	141,5
HAPs	425	8,98	17,4	86	3,4	6,1
InPHE	10	0,59	0,94	10	0,59	0,94
NAP	121	0,2	0,38	14	0,21	0,33
CNT	64	1,3	1,97	64	1,3	1,97

Polluants	BDD			BDDS		
	Effectif	Q90%	Q95%	Effectif	Q90%	Q95%
PCBs	41	0,04	0,05	10	0,1	0,13
B	2	0,26	0,27	2	0,26	0,27
BTEX	3	1,06	1,08	2	0,88	0,89
TEX	3	0,81	0,82	2	0,76	0,76
C5C10	0			0		
C10C40	304	237	374	49	87,91	126
DIOX	19	6,96	8,28	19	6,96	8,28

Les polluants surlignés en rouge restent supérieurs aux variations naturelles, quelles que soient les données utilisées pour le calcul des seuils.

Les sites ont une influence sur le calcul des seuils pour quelques polluants (cf. valeurs surlignées en orange). Afin de vérifier cette influence, nous calculons les seuils pour chaque site ainsi que pour les prélèvements n'appartenant à aucun site. Les tableaux suivants présentent les seuils avec la méthode des centiles Q90% et Q95%.

Polluants	Q90% (seuil en mg/kg)							
	BDD	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Aucun site
As	24	23	20	27	26,2	17	17	23,4
Sb	4,21					2,36	5,15	
Ba	140					77,4	73,3	150
Cd	0,7	0,33	0,57	0,7	0,67	0,75	0,24	0,72
Co	10,8							10,8
Cr	37	37,4	40	34	37	22	31	44,9
Cu	53	52,4	33,9	65	46,6	54	24	48,45
Hg	0,63	0,1	0,13	0,69	0,26	0,42	0,08	0,82
Ni	38	35	31,9	44	37,1	17	26	37
Pb	60	24,4	30,6	70	56,9	77,8	19	74
Se						0	0	
Mo	0,94					0,72	0,94	
Zn	130	76,6	91	140	141	130	79,9	120
HAPs	8,98	4,8	12	8,8	5,44	32,5	14,48	2,14
InPHE	0,59							0,59
NAP	0,2	0,04	0,14	0,11	0,1	0,38	0,05	0,27
CNT	1,3							1,3
PCBs	0,04		0,02	0,02	0,04			0,12
B	0,26		0,12			0,28		
BTEX	1,06		0,89			1,04		
TEX	0,81		0,77			0,78		
C5C10								
C10C40	237	184	276	170	102	488	398	49,2
DIOX	6,96							6,96

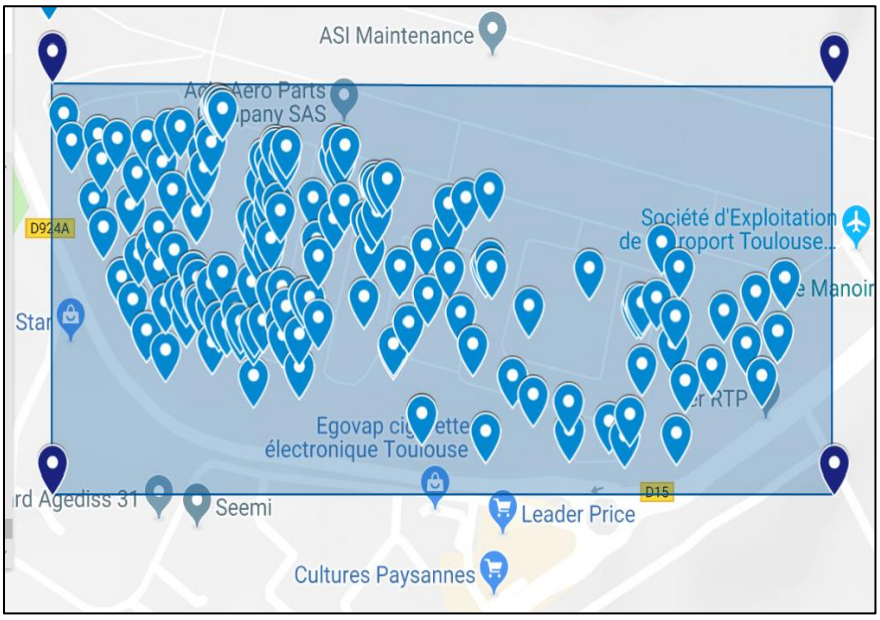
Polluants	Q95% (seuil en mg/kg)							
	BDD	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Aucun site
As	27	25,1	22	30	28	23,5	18	27
Sb	5,85					2,43	6,33	
Ba	160					78,2	81,15	170
Cd	0,77	0,39	0,75	0,76	0,79	0,77	0,25	0,86
Co	11							11
Cr	41,3	39,4	44	37	38	23,5	32,9	48
Cu	73,3	63,4	45	87	63	65,5	34,15	66,9
Hg	0,95	0,12	0,17	0,97	0,33	0,44	0,08	2,33
Ni	43	35	33	47	41	17,5	27	42
Pb	77	28,1	40,6	86,85	72,15	79	22	109
Se						0	0	
Mo	0,95					0,73	0,97	
Zn	140	81,7	110	150	170	165	82	144,5
HAPs	17,4	6	15,4	13,25	6,76	36,75	18,24	3,51
InPHE	0,94							0,94
NAP	0,38	0,04	0,37	0,21	0,11	0,56	0,06	0,4
CNT	1,97							1,97
PCBs	0,05		0,02	0,03	0,04			0,14
B	0,27		0,12			0,28		
BTEX	1,08		0,89			1,07		
TEX	0,82		0,77			0,8		
C5C10								
C10C40	374	217	509,5	298	126	568	483	64
DIOX	8,28							8,28

Nous analysons deux exemples particuliers de sites, les deux plus conséquents, soit le 2 et 3. Puis nous donnons les résultats globaux pour l'ensemble des sites.

- Analyse du site 2

Le site 2 est celui localisé au sud de l'agglomération. Il contient 312 prélèvements ce qui correspond à 23% de la BDD. Selon le BRGM, ce site est connu comme étant pollué.

Pour ce site, nous récupérons tous les forages contenus dans le rectangle bleu. Le positionnement des forages de ce site n'est pas uniforme, comme le montre la carte suivante.

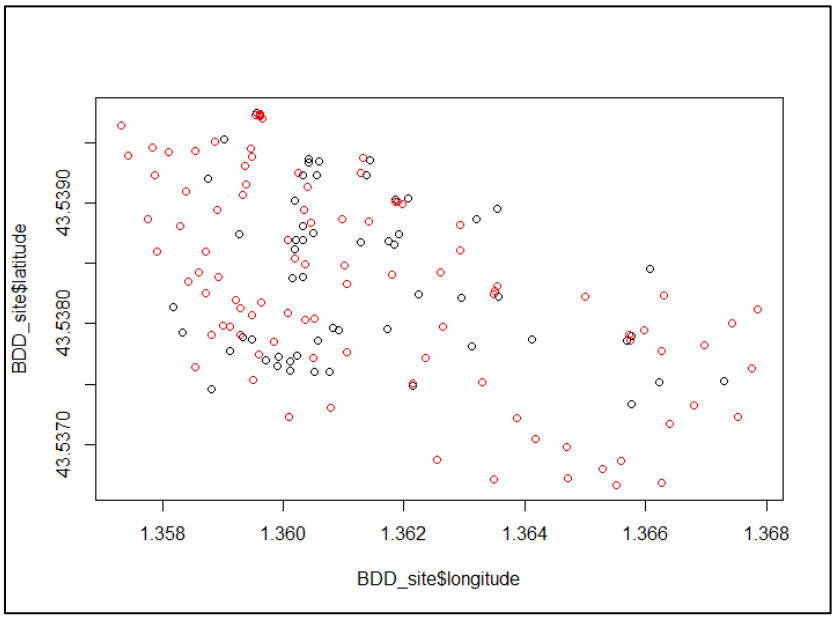


Sur ces 312 prélèvements, 134 sont pollués, soit 43% des données de ce site. Un prélèvement est dit pollué si au moins l'un des polluants a une concentration supérieure au seuil calculé à partir de la méthode des centiles. Pour cet exemple, nous nous basons sur des seuils déterminés à partir de la méthode Q90%.

Le tableau ci-dessous récapitule le nombre de polluants ayant une concentration supérieure au seuil. La moitié des prélèvements sont considérés comme pollués parce qu'ils présentent une seule concentration trop élevée.

Nombre de polluants supérieurs au seuil Q90%	1	2	3	4	5	6	Total
Nombre de prélèvements	85	30	14	3	1	1	134

Le graphique ci-dessous, explicite les positionnements des 134 prélèvements considérés comme pollués (cercles rouges sur le graphique ci-dessous).



Nous voulons déterminer quel est le polluant le plus présent au sein des prélèvements pollués. Pour les 85 prélèvements présentant un seul polluant avec une concentration élevée, nous avons listé les polluants mis en cause dans le tableau ci-dessous.

Polluant	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn	HAPs	NAP	C10C40	Total
Nombre de prélèvements	12	3	18	11	1	13	10	9	2	2	4	85

Les éléments à traces métalliques sont majoritaires.

- Analyse du site 3

Le site 3 est celui localisé au centre de l'agglomération. Il contient 562 prélèvements ce qui correspond à 41% de la BDD.

Pour ce site, nous récupérons tous les forages contenus dans le rectangle vert. Le positionnement des forages de ce site est assez uniforme comme le montre la carte suivante.



Sur ces 562 prélèvements, 225 sont pollués soit 40% des données de ce site. Pour cet exemple nous nous basons également sur des seuils déterminés à partir de la méthode Q90%.

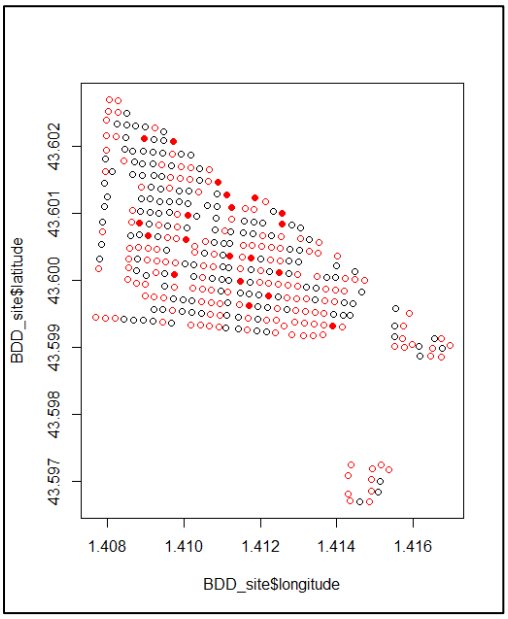
Le tableau ci-dessous récapitule le nombre de polluants ayant une concentration supérieure au seuil. La moitié des prélèvements sont considérés comme pollués parce qu'ils présentent une seule concentration trop élevée.

Nombre de polluants supérieurs au seuil Q90%	1	2	3	4	5	Total
Nombre de prélèvements	117	52	44	10	2	225

Nous voulons également déterminer quel est le polluant le plus présent au sein des prélèvements pollués de ce site. Pour les 117 prélèvements présentant un seul polluant avec une concentration élevée, nous avons listé les polluants mis en cause dans le tableau ci-dessous. Il s'agit des mêmes polluants que pour le site 2.

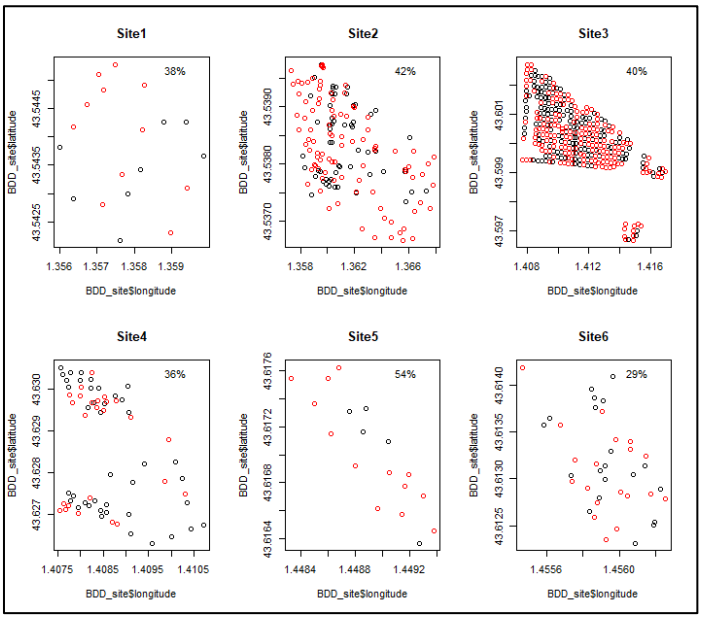
Polluant	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn	HAPs	NAP	PCBs	C10C40	Total
Nombre de prélèvements	5	13	11	20	18	7	20	7	8	1	1	6	117

Le graphique ci-dessous explicite les positionnements des 562 prélèvements. Par exemple, les 20 prélèvements pollués à cause du plomb parmi l'ensemble des prélèvements pollués du site (cercles rouges) sont représentés par des points rouges. Il est possible que deux prélèvements se superposent s'ils sont pollués au même endroit, mais à des profondeurs différentes.



- Résultats pour l'ensemble des sites

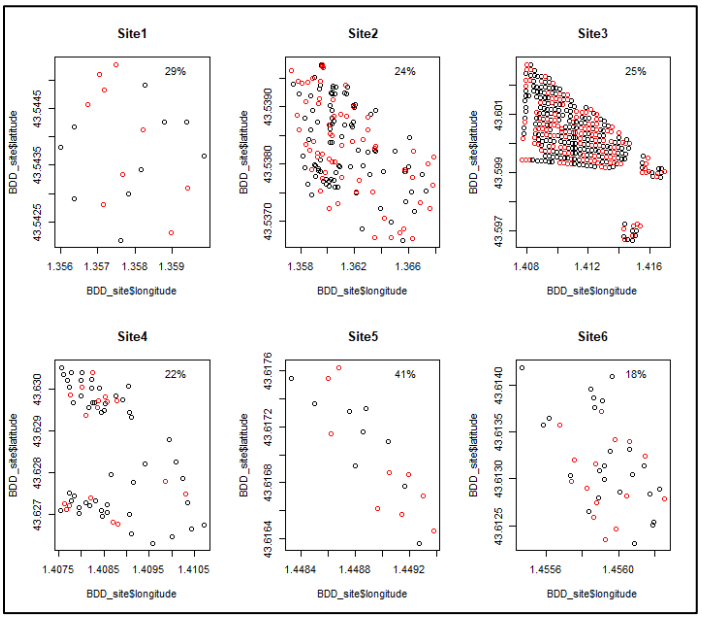
Les graphiques ci-dessous renseignent en haut à droite le pourcentage de prélèvements considérés comme pollués selon la méthode Q90%. Ces prélèvements pollués sont représentés en rouge pour chaque site.



Avec cette méthode, aucun site n'est recommandable tel quel. En effet, les pourcentages de pollution sont plutôt élevés.

Néanmoins, lorsqu'il s'agit d'un polluant peu nocif, un tel taux de pollution peut être acceptable, même s'il serait préférable d'envisager des mesures de dépollution lorsque le taux est proche de 50%. En revanche, si des polluants tels que l'arsenic, le cadmium ou le plomb (qui font partie des polluants les plus nocifs) sont la cause de pollution d'un grand nombre de prélèvements (comme nous avons pu le constater avec les sites 2 et 3), alors ces sites restent non recommandables.

Les graphiques ci-dessous renseignent en haut à droite le pourcentage de prélèvements considérés comme pollués selon la méthode Q95%. Ces prélèvements pollués sont représentés en rouge pour chaque site.



Les taux de pollution sont moins élevés pour la plupart des sites avec cette méthode, ce qui reste logique. Cependant, il n'est pas préférable d'utiliser cette méthode pour les polluants dits « nocifs » ou « toxiques » : elle n'est pas assez restrictive.

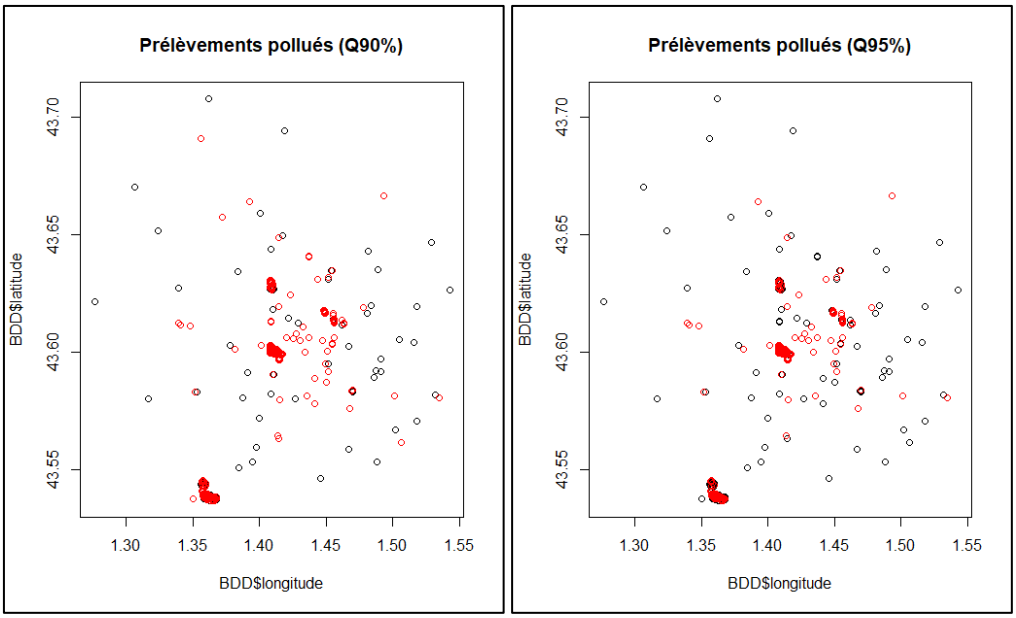
B. Détermination des prélèvements pollués

Un prélèvement est dit pollué si au moins l'un des polluants a une concentration supérieure stricte au centile 90 % (reps. 95%). Le tableau ci-dessous renseigne le nombre, ainsi que la proportion, de prélèvements pollués selon la base de données et la méthode utilisées.

Données	Nombre total de prélèvements	Nombre de prélèvements pollués avec Q90%	Nombre de prélèvements pollués avec Q95%
BDD	1356	528 (39%)	327 (24%)
BDDS	218	86 (40%)	53 (24%)

Il n'y a pas d'influence des sites sur la proportion de prélèvements pollués : les pourcentages sont similaires, que les données concernant les sites soient prises en compte en totalité ou en moyenne.

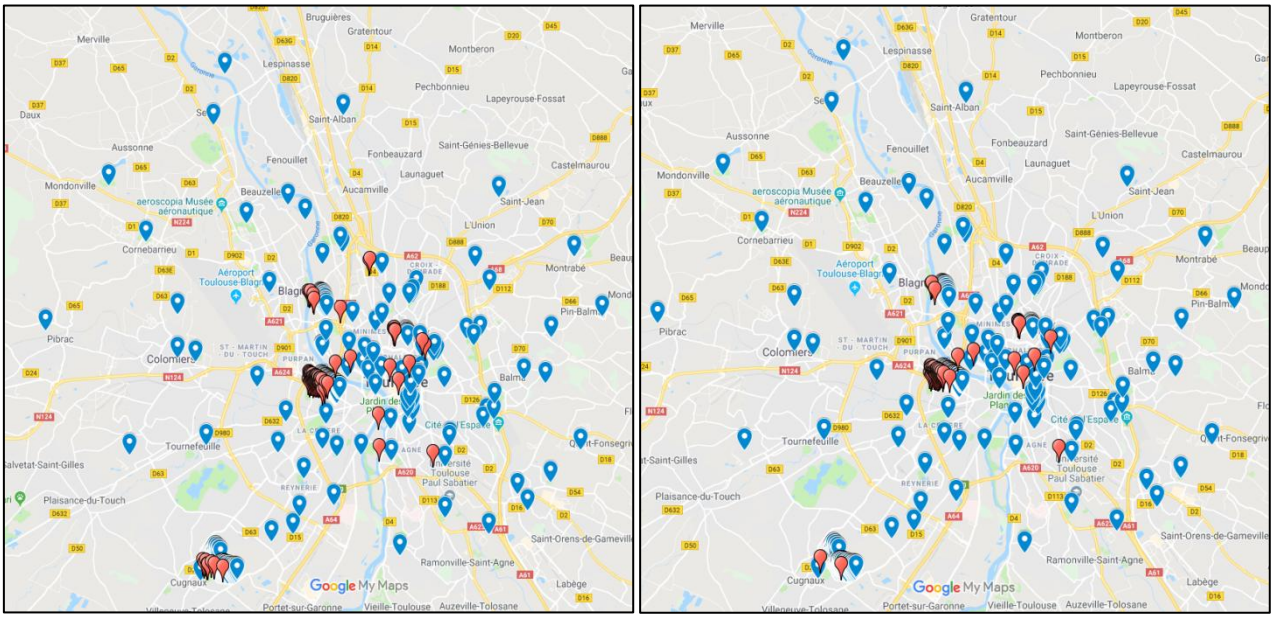
Les graphiques ci-dessous représentent en rouge les prélèvements pollués, avec la méthode Q90% et Q95%.



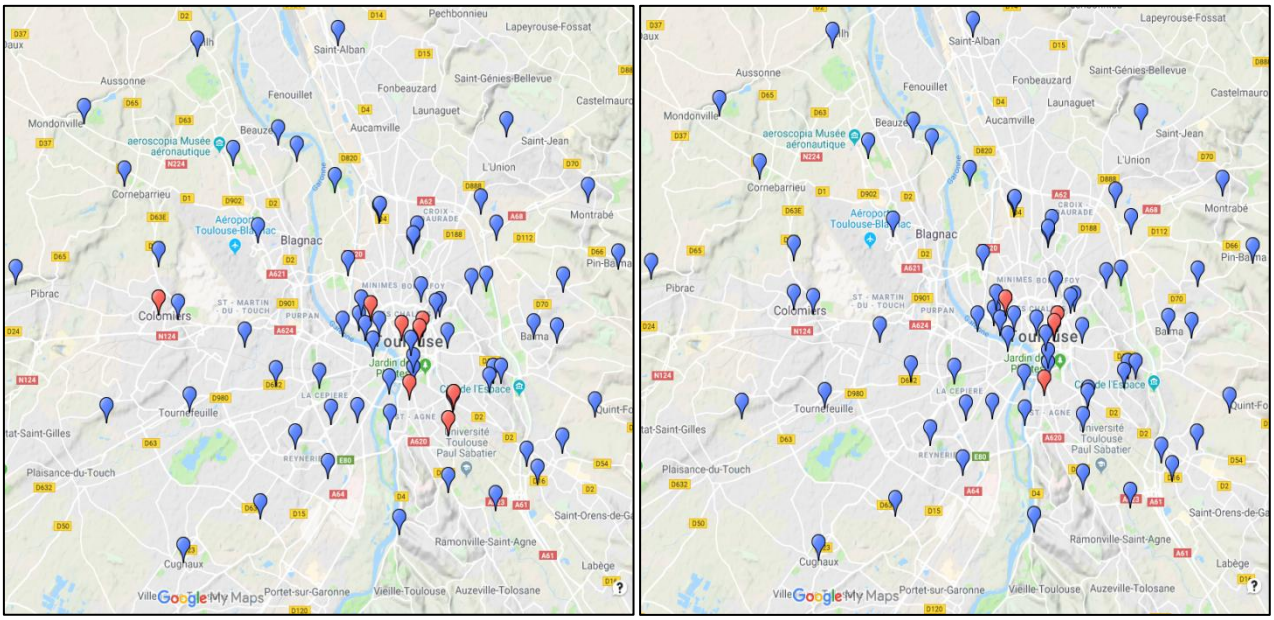
La conclusion est la même, que nous prenions Q90% ou Q95% : ce sont principalement les sites et le centre de l'agglomération qui sont pollués.

Le choix de Q90% ou Q95% dépend du niveau de sécurité désiré. Pour un polluant plus nocif, il est préférable que le seuil soit plus faible que pour un polluant mieux toléré (dans ce cas, il faut privilégier le centile 90%).

Voici un exemple de pollution dû au plomb (points rouges) : prélèvements pollués selon la méthode Q90% à gauche et selon la méthode Q95% à droite. Pour ce polluant, connu comme l'un des plus nocifs, il est préférable d'utiliser la méthode Q95%.



Voici un autre exemple de pollution : pollution au cyanure (points rouges). Les prélèvements pollués selon la méthode Q90% sont ceux identifiés sur le graphique de gauche et selon la méthode Q95% ceux identifiés sur le graphique de droite.



IV. Classement des prélèvements

Nous voulons répondre aux questions du BRGM suivantes : les sols des champs sont-ils différents des sols des villes ? Y a-t-il des variations intra métropole ? Pour cela nous calculons un indice de similarité qui nous permet ensuite de classer les prélèvements.

Nous travaillons avec une extraction de la BDD (sans outliers et avec les concentrations inférieures à la LQ supprimées). Nous conservons seulement les paramètres donnant les concentrations des polluants (soit 24 paramètres pour 1 356 prélèvements). Nos paramètres sont quantitatifs réels ce qui implique que nous ne pouvons pas calculer d'indices binaires tels que l'indice de similarité de Jacard ou de Dice.

Du fait de la suppression des valeurs de concentrations inférieures à la LQ, les données présentent beaucoup de valeurs manquantes. Ceci pose problème pour le calcul des indices de similarité : il y a trop de données manquantes, même pour les indices qui peuvent les prendre en compte (e.g. indice de Gower). Pour parer à ce problème, nous utilisons une méthode d'imputation simple présente dans le package « missMDA » du logiciel R.

Le package « missMDA » permet de gérer les données manquantes à partir de l'imputation simple. L'imputation simple consiste à remplacer les valeurs manquantes par des valeurs plausibles (cela revient à compléter le jeu de données). Les valeurs manquantes sont imputées de sorte que les valeurs imputées n'aient aucune influence sur les résultats (i.e. les valeurs imputées n'ont aucun poids donc les résultats de l'analyse sont obtenus uniquement avec les valeurs observées). Ce sont des méthodes de réduction de données qui sont utilisées lors de l'imputation, comme l'ACP (Analyse en Composantes Principales), en prenant en compte à la fois les similarités entre prélèvements et les liens entre variables. Ces méthodes permettent d'imputer de façon satisfaisante de gros jeux de données. Elles sont plus efficaces que les méthodes classiques (e.g. forêts aléatoires).

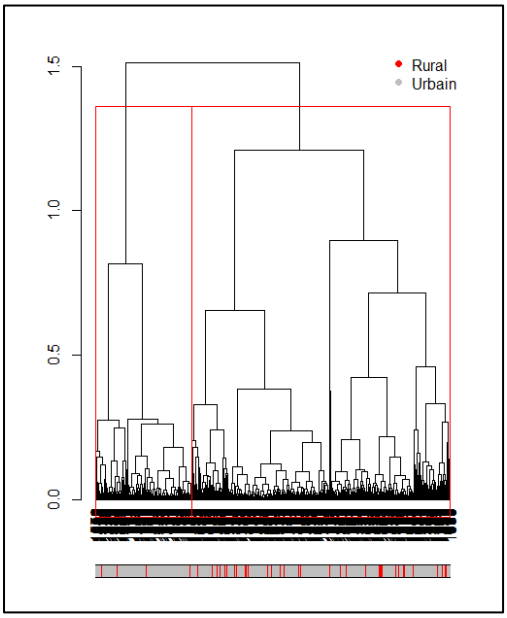
Nous préférons travailler avec les données brutes lorsque cela est possible. Or, ce n'est pas le cas. En effet, avec les données brutes, il n'est pas possible d'obtenir un indice de similarité (que ce soit la distance euclidienne, euclidienne normalisée, de Manhattan, de Mahalanobis, de Gower, etc.) à cause du grand nombre de valeurs inférieures à la LQ qui sont supprimées. Nous avons recours à cette méthode seulement pour pouvoir apporter des éléments de réponse aux questions du BGRM.

Une fois les polluants Se et C5C10 supprimés (ils n'ont plus aucune donnée) et nos données complétées obtenues, nous travaillons avec l'indice de similarité de Gower. L'indice de Gower consiste également à mesurer dans quelle mesure deux prélèvements sont semblables. Il varie entre 0 et 1. Si l'indice vaut 1, les deux prélèvements sont identiques. Au contraire, s'il vaut 0, les deux prélèvements n'ont pas de point commun. Les résultats sont similaires avec les indices basés sur la distance euclidienne ou de Manhattan.

Nous utilisons cet indice afin de classer les prélèvements selon qu'ils appartiennent à la ville ou aux champs. Nous connaissons déjà la répartition "Urbain" / "Rural" des prélèvements (146 "Rural" contre 1 210 "Urbain", cf. § I.B.3, paramètre "GEOBAP"), mais nous souhaitons déterminer si les prélèvements montrent bien une similitude au sein de chaque classe.
Autrement dit, nous souhaitons vérifier si nous retrouvons bien ces deux classes à partir du calcul de similarité et de la classification des prélèvements.

Nous travaillons avec la CAH (Classification Ascendante Hiérarchique) qui relève de la classification non supervisée (i.e. les classes des prélèvements sont déjà connues).

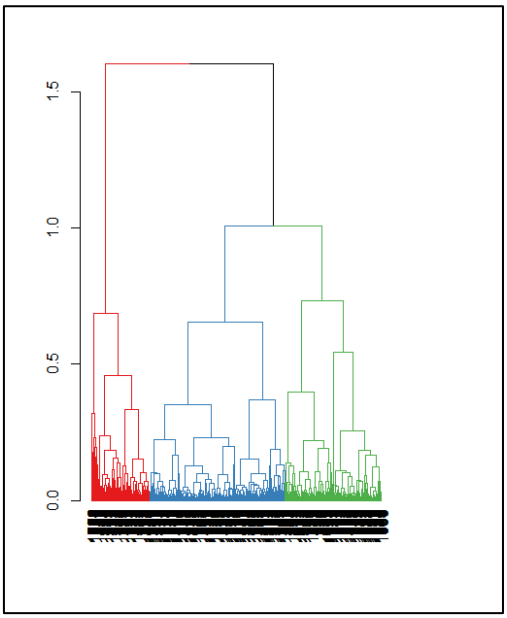
La CAH est une méthode de regroupement par étapes qui, à chaque étape successive, réunit les deux groupes dont la dissimilarité est la plus petite. Les résultats de la CAH sont présentés ci-dessous.

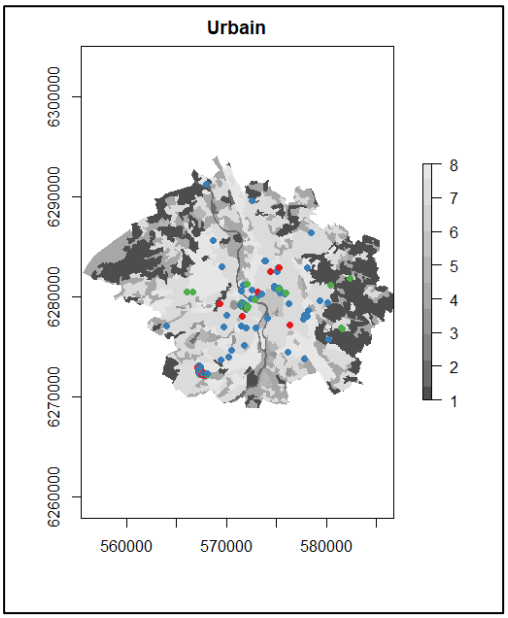


Le graphique correspond au dendrogramme réalisé à partir de la méthode de Ward sur les données de similarité de Gower. La méthode de Ward consiste à regrouper les classes de façon à avoir une inertie intra classe la plus faible possible. Les deux rectangles rouges représentent les deux classes trouvées à partir de cette méthode. L'échelle de couleur sous le graphique permet de visualiser les vraies classes des prélèvements ("Urbain" / "Rural") par rapport à leur classification. Par exemple, tous les prélèvements appartenant à la catégorie "Rural" (traits rouges parmi les gris) ne sont pas regroupés dans une même classe, mais sont dispersés dans les deux classes.

Ainsi, il n'existe pas de différence significative entre les sols des champs et les sols des villes au vu de leur concentration en différents polluants.

Pour déterminer l'existence de variations intra métropole, nous réalisons également une CAH sur l'ensemble des données urbaines (soit 1 210 prélèvements). Les résultats de la CAH montrent trois groupes distincts (cf. graphiques ci-dessous).





Une fois représentés sur la carte, il est visible que les prélèvements de chaque groupe ne sont pas tous situés à un même endroit. Au contraire, ils restent dispersés (cf. points bleus).

Nous calculons la moyenne des concentrations selon les classes obtenues.

	Rouge	Bleu	Vert
As	17,03	13,77	17,49
Sb	3,30	2,16	
Ba	77,25	80,20	94,97
Cd	0,38	0,39	0,45
Co	7	14,17	
Cr	30,92	26,73	28,06
Cu	26,81	28,74	36,76
Hg	0,26	0,34	0,27
Ni	28,61	24,34	28,94
Pb	25,06	27,29	36,36
Se			
Mo	0,60	0,74	0,80

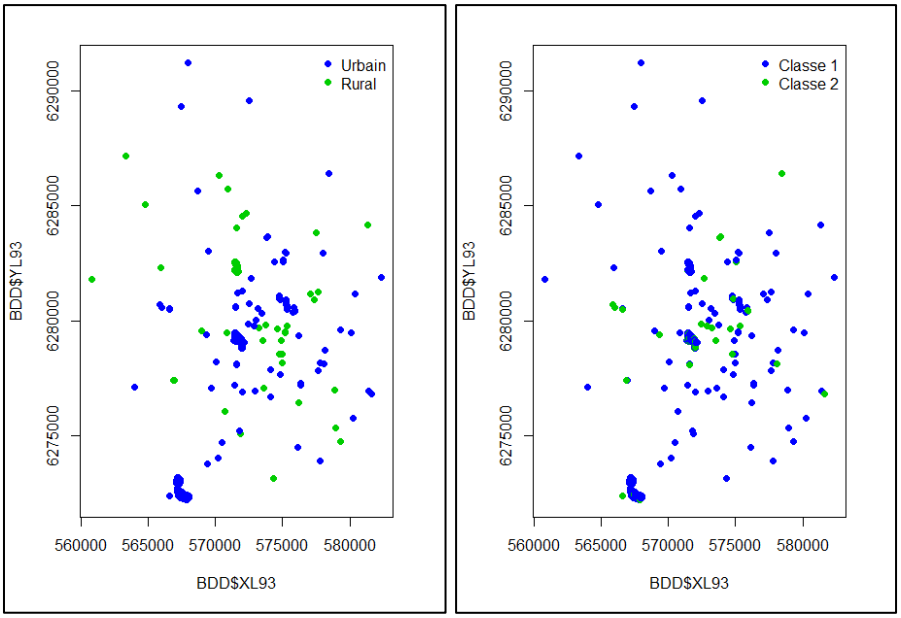
	Rouge	Bleu	Vert
Zn	76,54	76,06	94,75
HAPs	4,36	4,32	6,54
InPHE	0,34	0,55	0,14
NAP	0,06	0,10	0,15
CNT	0,66	0,66	0,25
PCBs	0,01	0,03	0,03
B	0,12	0,28	
BTEX	0,89	0,80	
TEX	0,77	0,64	
C5C10			
C10C40	82,53	116,67	84,09
DIOX	3,40	6,26	3,74

Les résultats appuient la présence de variations intra métropole avec des différences inter classes plus ou moins importantes selon le polluant concerné.

Une seconde méthode envisageable pour classer les prélèvements, et par conséquent identifier d'éventuelles variations de concentration, est la méthode des « kmeans » (ou moyennes mobiles en français). Cependant les données manquantes ne sont pas prises en charge par cette méthode, ce qui la rend inapplicable à nos données.

Une solution est de faire tourner l'algorithme des kmeans sur les données de dissimilarité calculées précédemment. Nous retrouvons les mêmes résultats qu'avec la CAH. Autrement dit, il n'existe pas de différence significative entre les sols des champs et les sols des villes.

Par exemple, les vraies classes "Urbain" / "Rural" du graphique de gauche ci-dessous se retrouvent complètement mélangées dans les classes obtenues avec l'algorithme kmeans (graphique de droite). Cela signifie que l'algorithme n'arrive pas à distinguer efficacement ces deux catégories parce qu'il n'existe pas suffisamment de différence entre elles.



V. Conclusion

Pour des raisons théoriques et pratiques, nous ne favorisons pas l'utilisation de la méthode MED+2MAD ni de la vibrisse interne supérieure. En effet, certaines conditions d'utilisation ne sont pas vérifiées (normalité ou symétrie des données) ; et pour la MED+2MAD, les résultats ont des ordres de grandeur totalement différents des résultats des trois autre méthodes.

Le choix de Q90% ou Q95% dépend du niveau de sécurité désiré. Pour un polluant plus nocif, il est préférable que le seuil soit plus faible que pour un polluant mieux toléré.

Par exemple, comme l'arsenic (As), le cadmium (Cd) et le plomb (Pb) font partie des polluants les plus nocifs, il est préférable d'utiliser le centile à 90%, d'autant plus si les valeurs sont supérieures aux variations naturelles. Il en est de même pour le benzène (B) et les hydrocarbures (HAPs). De plus, il faut peut-être envisager une mesure de dépollution avant d'établir une construction lorsque ces polluants sont présents en forte concentration.

Pour les polluants inclus dans la variabilité naturelle, comme le cuivre (Cr), le nickel (Ni) ou le zinc (Zn), il n'est pas forcément nécessaire de les considérer comme dangereux ni problématiques pour une construction donc le centile 95% peut suffire pour la détermination des seuils.

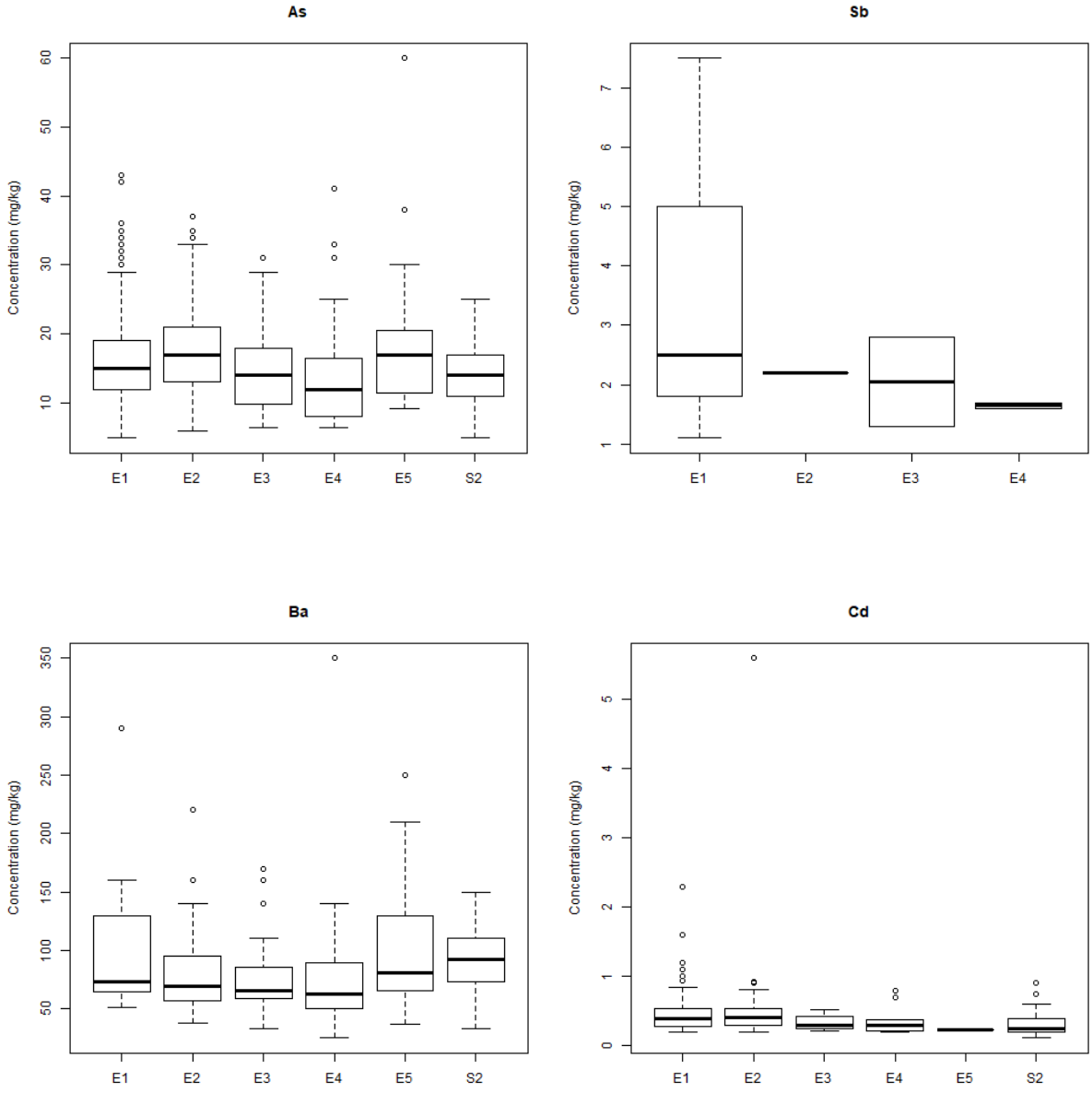
Enfin, au vu des résultats, il n'existe pas de différence de variations de concentration entre les sols urbains et les sols ruraux. Néanmoins il existe quelques variations intra métropole, même si les prélèvements présentant les mêmes variations ne sont pas regroupés au même endroit sur la carte.

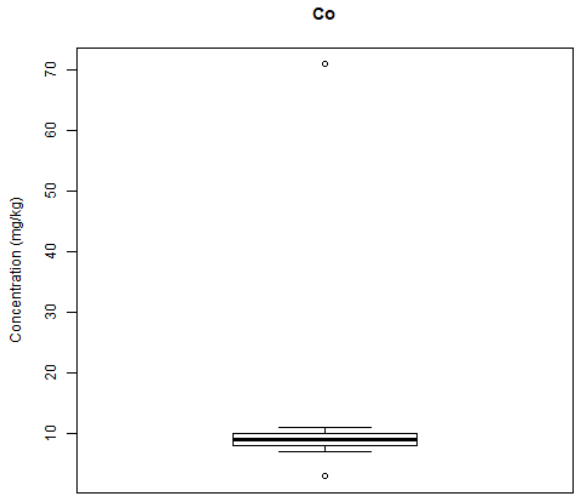
Annexe : Concentration en fonction de la profondeur

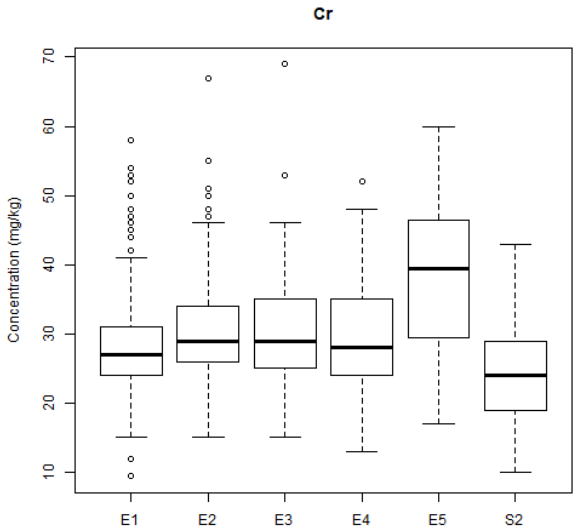
Les graphiques de cette annexe présentent les boxplots des concentrations en fonction de la profondeur de forage pour l'ensemble des polluants de BDD.

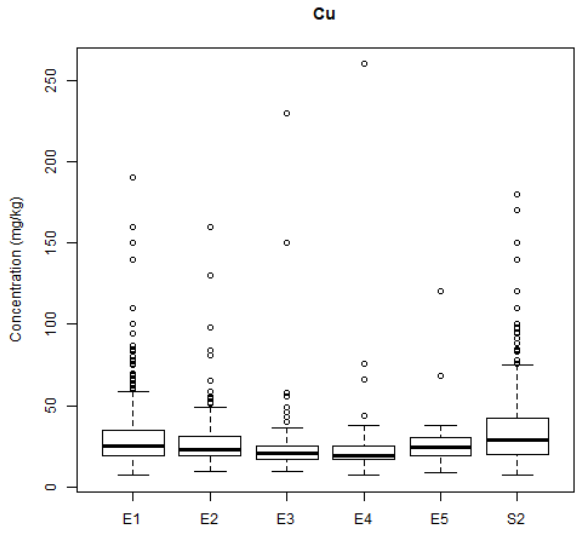
Lorsqu'il n'y a pas de profondeur précisée en abscisse, cela signifie que les prélèvements proviennent de forages de la couche appelée « S2 », soit entre 0 – 55 cm (e.g. Co, InPHE, B, BTEX, TEX, DIOX).

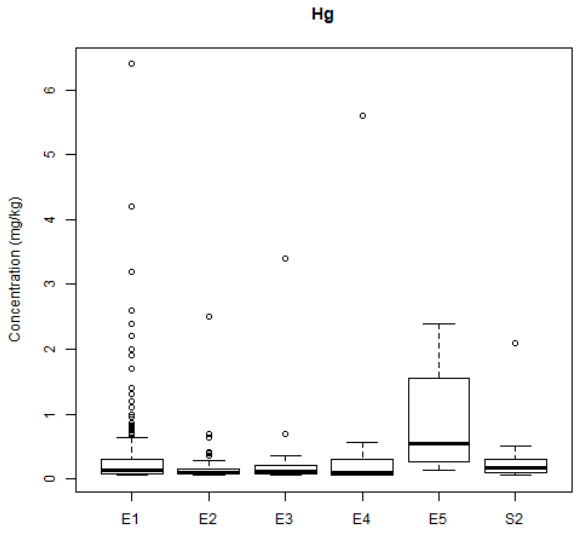
Sauf cas particulier, e.g Mo ou les polluants à faible effectif, et les valeurs extrêmes mises à part, les distributions des concentrations en fonction des profondeurs sont globalement les mêmes.

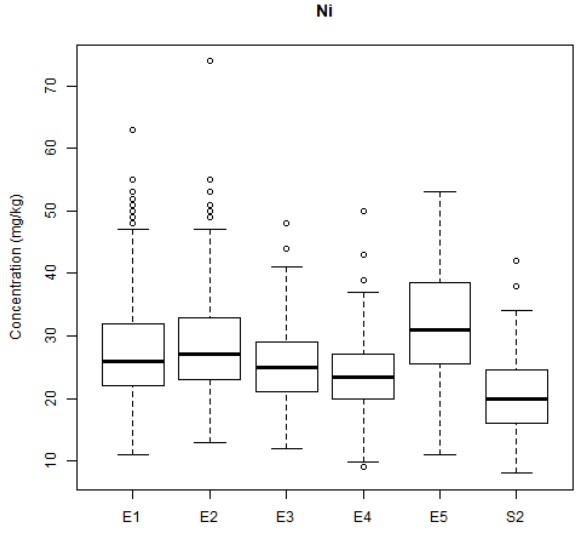


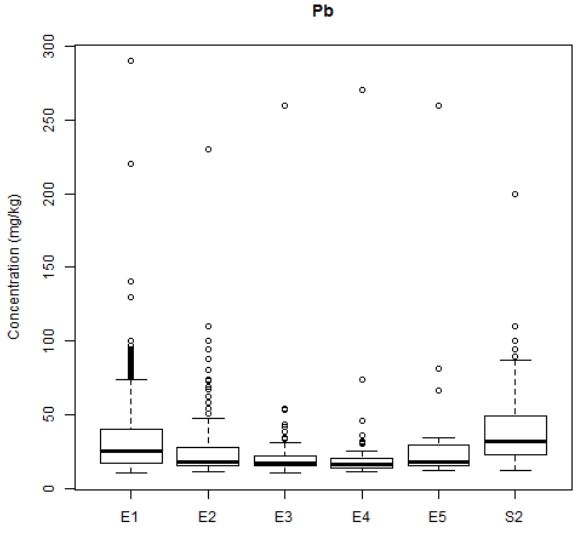

Co
Concentration (mg/kg)
10
20
30
40
50
60
70

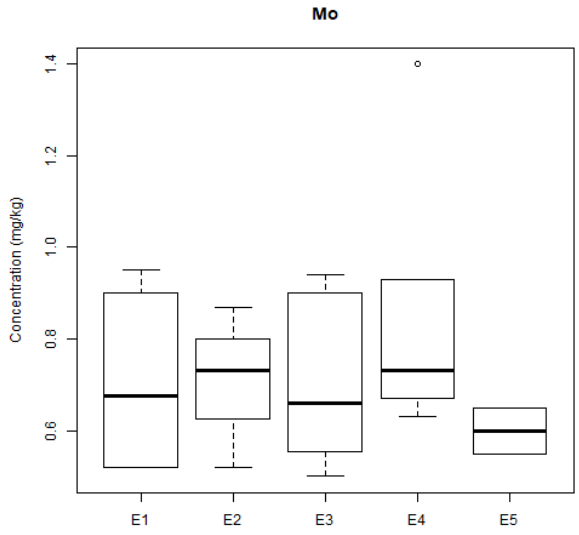

Cr
Concentration (mg/kg)
10
20
30
40
50
60
70
E1
E2
E3
E4
E5
S2

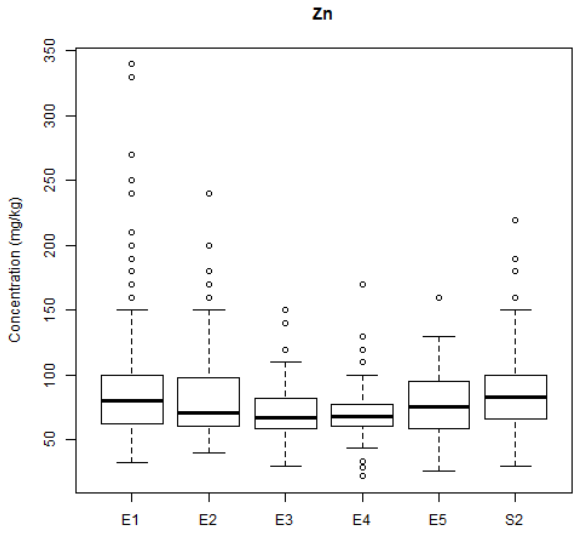

Cu
Concentration (mg/kg)
0
50
100
150
200
250
E1
E2
E3
E4
E5
S2

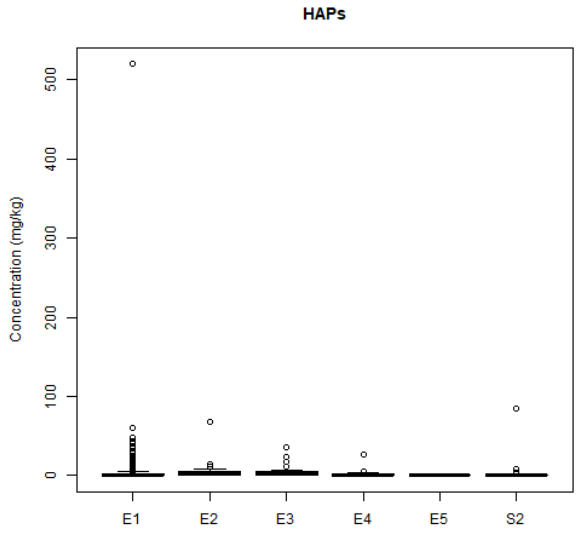

Hg
Concentration (mg/kg)
0
1
2
3
4
5
6
E1
E2
E3
E4
E5
S2

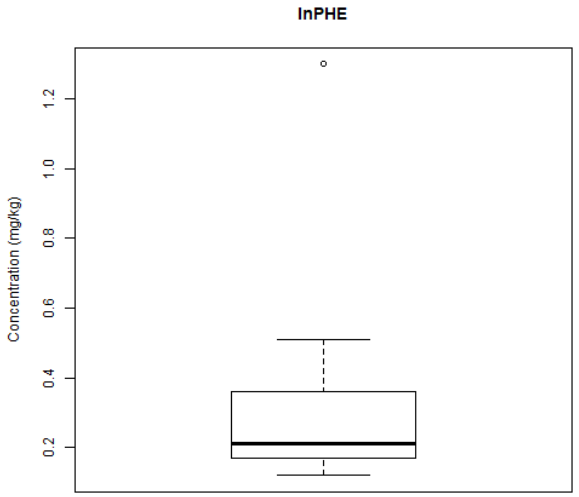

Ni
Concentration (mg/kg)
10
20
30
40
50
60
70
E1
E2
E3
E4
E5
S2

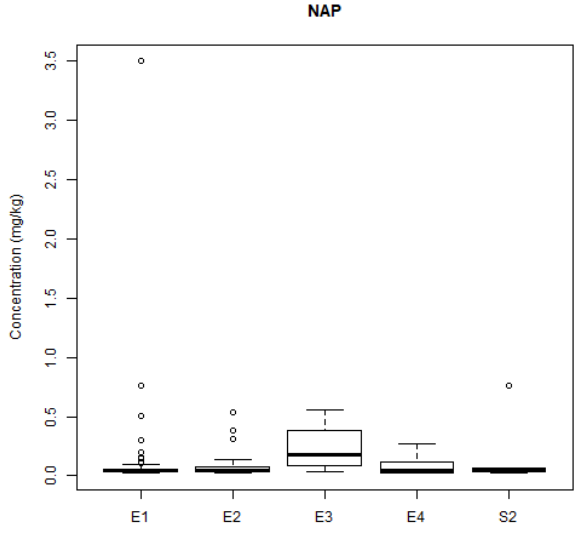

Pb
Concentration (mg/kg)
0
50
100
150
200
250
300
E1
E2
E3
E4
E5
S2

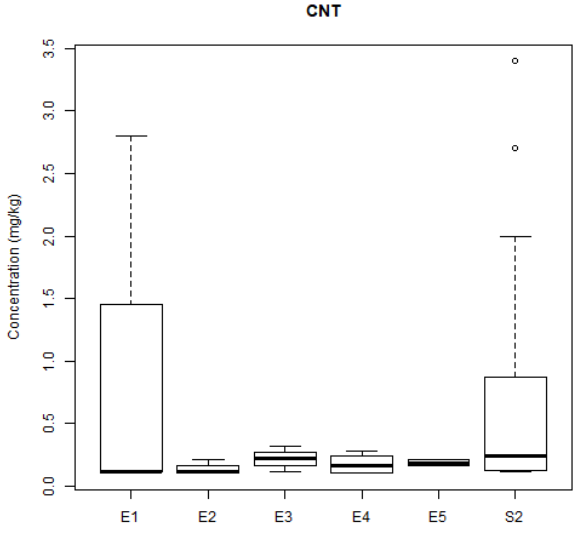

Mo
Concentration (mg/kg)
0.6
0.8
1.0
1.2
1.4
E1
E2
E3
E4
E5

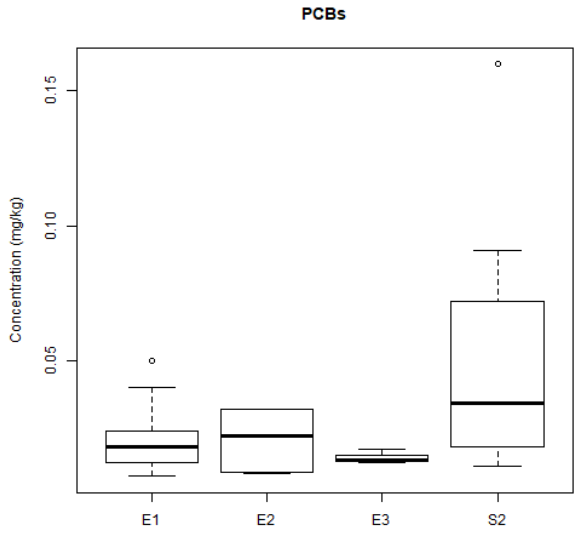

Zn
Concentration (mg/kg)
50
100
150
200
250
300
350
E1
E2
E3
E4
E5
S2

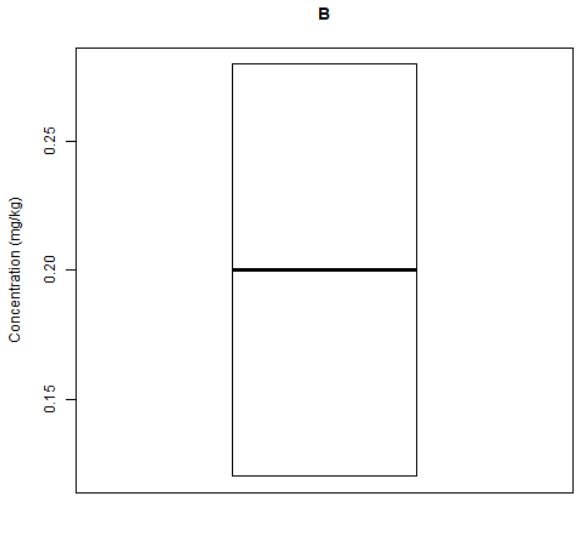

HAPs
Concentration (mg/kg)
0
100
200
300
400
500
E1
E2
E3
E4
E5
S2

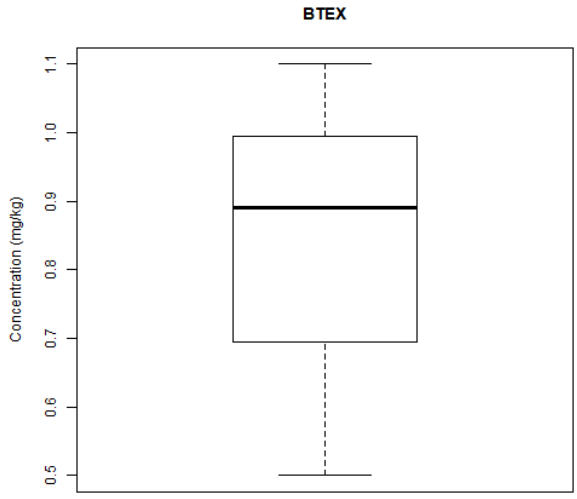

InPHE
Concentration (mg/kg)
0.2
0.4
0.6
0.8
1.0
1.2

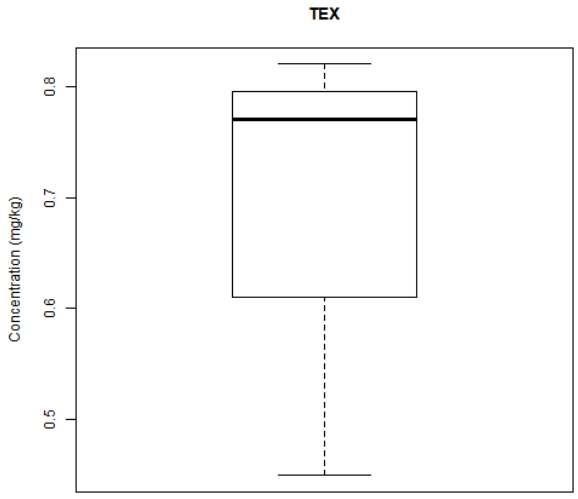

NAP
Concentration (mg/kg)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
E1
E2
E3
E4
S2

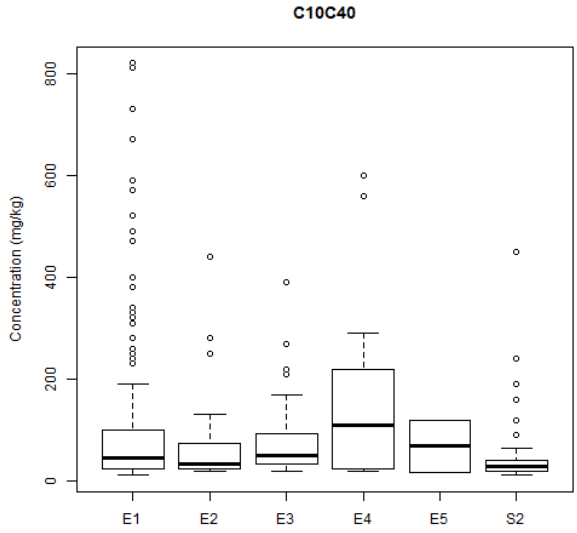

CNT
Concentration (mg/kg)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
E1
E2
E3
E4
E5
S2

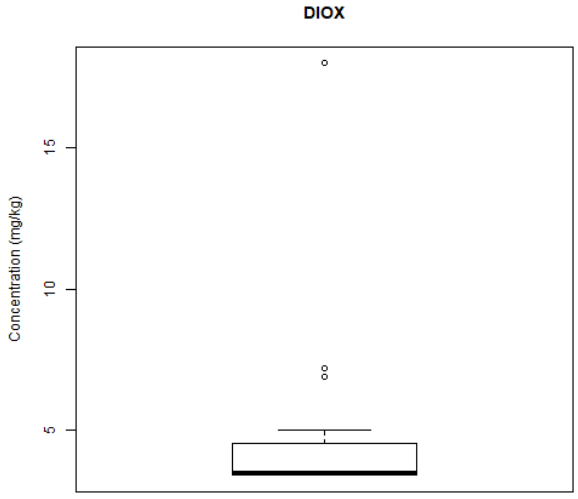

PCBs
Concentration (mg/kg)
0.05
0.10
0.15
E1
E2
E3
S2


B
Concentration (mg/kg)
0.15
0.20
0.25


BTEX
Concentration (mg/kg)
0.5
0.6
0.7
0.8
0.9
1.0
1.1


TEX
Concentration (mg/kg)
0.5
0.6
0.7
0.8


C10C40
Concentration (mg/kg)
0
200
400
600
800
E1
E2
E3
E4
E5
S2


DIOX
Concentration (mg/kg)
5
10
15

Bibliographie

[BGEP] *Diagnostic initial de la qualité des sols mission A200,* CEBTP, geoterre, BGEP.F0033, décembre 2015

[BRGM] *Protocole d'analyse statistique pour la contribution d'un fond pédo-géochimique anthropisé des sols urbains*, Rapport final, RP-66501-FR, décembre 2016

[BRGM1] *Etablissement de fonds pédogéochimiques urbains en parallèle à l'Opération ETS du Ministère de l'Ecologie, Rapport intermédiaire*, RP-66306-FR, octobre 2016

[BRGMAs] *Guide méthodologique de l'arsenic appliqué à la gestion des sites et des sols pollués*, RP-52066-FR, janvier 2003

[BRGMHg] *Le mercure et ses composés, Comportement dans les sols, les eaux et les boues de sédiments*, Rapport final, RP-51890-FR, juin 2003

[INERIS] A. BURNOL, L. DURO, M. GRIVE, *Eléments traces métalliques, Guide méthodologique, Recommandations pour la modélisation des transferts des éléments traces métalliques dans les sols et les eaux souterraines*, Rapport d'étude, INERIS-DRC-06-66246/DESP-R01a, 28/08/2016

[INERISInfo] https://substances.ineris.fr (recherche de chaque polluant)

[Log] C. Feng et al, *Log-transformation and its implications for data analysis*, Shanghai Archives of Psychiatry, 2014, Vol. 26, No. 2

[Vibrisse] E. Villanneau et al, *Détection de valeurs anomaliques d'éléments traces métalliques dans les sols à l'aide du Réseau de Mesure de la Qualité des Sols*, Etude et Gestion des Sols, Volume 15, 2008, p. 183 à 200