Communauté française de Belgique
Faculté universitaire des Sciences agronomiques de Gembloux

Détection de valeurs aberrantes dans des mélanges de distributions dissymétriques pour des ensembles de données avec contraintes spatiales

Viviane PLANCHON

Dissertation originale présentée en vue de l'obtention du grade
de docteur en sciences agronomiques et ingénierie biologique

Promoteurs : Pr. J.J. CLAUSTRIAUX
Dr. R. OGER

2007

Planchon Viviane. (2007). **Détection de valeurs aberrantes dans des mélanges de distributions dissymétriques pour des ensembles de données avec contraintes spatiales** (Thèse de doctorat). Gembloux, Faculté Universitaire des Sciences Agronomiques, 237 p., 43 tabl., 80 fig.

Résumé. Dans le cas des analyses chimiques de sols, les distributions de fréquences des résultats présentent, pour certains éléments étudiés, un caractère très dissymétrique avec un étalement très marqué à droite ou à gauche. Une fréquence importante de valeurs extrêmes est également observée et un mélange éventuel de plusieurs distributions au sein d'une même entité géographique, lié à la présence de divers types de sols, peut être rencontré. Dès lors, pour la détection des valeurs aberrantes et la fixation des limites de détection, une méthode originale, permettant d'estimer des quantiles extrêmes au-dessus et en dessous desquelles les observations sont considérées comme aberrantes, a été élaborée. L'estimation des limites de détection est établie de manière distincte à partir des queues des distributions droite et gauche. Une première estimation par entité géographique élémentaire est réalisée afin de déterminer un niveau de troncature adéquat. Une classification spatiale permet ensuite de créer des groupes d'entités homogènes contiguës, de manière à estimer des valeurs limites robustes basées sur un nombre d'observations optimal.

Planchon Viviane. (2007). **Outliers detection in mixtures of dissymmetric distributions for data sets with spatial constraints** (Thèse de doctorat). Gembloux, Belgium, Faculté Universitaire des Sciences Agronomiques, 237 p., 43 tabl., 80 fig.

Summary. In the case of soil chemical analyses, frequency distributions for some elements show a dissymmetrical aspect, with a very marked spread to the right or to the left. A high frequency of extreme values is also observed and a possible mixture of several distributions, due to the presence of various soil types within a single geographical unit, is encountered. Then, for the outliers detection and the establishment of detection limits, an original outliers detection procedure has been developed; it allows estimating extreme quantiles above and under which observations are considered as outliers. The estimation of these detection limits is based on the right and the left of the distribution tails. A first estimation is realised for each elementary geographical unit to determine an appropriate truncation level. Then, a spatial classification allows creating adjoining homogeneous groups of geographical units to estimate robust limit values based on an optimal number of observations.

REMERCIEMENTS

La réalisation d'une thèse de doctorat est un travail de longue haleine, qui est le fruit de nombreux échanges et de collaborations diverses ; je désire ici remercier les personnes qui m'ont aidée et encouragée à la réaliser. Mes remerciements s'adressent plus particulièrement à :

Monsieur Jean-Jacques Claustriaux, promoteur de ce travail, pour sa disponibilité, son accueil toujours chaleureux et ses nombreux conseils.

Monsieur Robert Oger, promoteur également, pour les longues discussions et les conseils statistiques toujours pertinents.

Monsieur Rodolphe Palm, rapporteur, dont les avis et suggestions ont été très appréciables à chaque étape d'avancement du travail.

Monsieur Jean-Paul Gouet (ARVALIS, Institut du végétal - France), rapporteur, pour ses commentaires constructifs.

Mesdames et Messieurs Catherine Charles, Sylvia Dautrebande, Laurent Bock, Philippe Lejeune, André Toussaint, membres du jury, dont les remarques m'ont permis d'améliorer le manuscrit.

Monsieur Patrick Meeùs et Monsieur Roger Piscaglia pour leurs encouragements tout au long de ce travail.

Monsieur Pierre Dagnelie qui m'a à nouveau fait profiter de son expérience statistique et m'a amicalement fourni de précieux conseils.

Messieurs Jan Beirlant et Yuri Goegebeur pour leur collaboration au début du travail et pour les discussions très intéressantes sur la *théorie des valeurs extrêmes*.

Mesdames Delphine Pontegnies et Gisèle Bazier (MET) ainsi que Monsieur Bernard Mohymont (IRM) pour les échanges d'idées à propos des *événements rares*.

Madame Valérie Genot pour son aide très précieuse dans l'interprétation des résultats liés aux analyses de sols.

Monsieur Dany Weverberg pour ses qualités de pédagogue en mathématiques.

La Direction Générale de l'Agriculture du Ministère de la Région Wallonne (Monsieur Mohamed Mokadem) pour la mise à disposition de données relatives au projet de cartographie numérique des sols de la Région Wallone (PCNSW).

Mes remerciements sont également destinés aux membres du personnel de :

la Section Biométrie, Gestion des données et Agrométéorologie du Centre wallon de Recherches agronomiques (CRA-W) ; merci particulièrement à Marie-Thérèse Marique-Zimmer pour son aide au niveau de la bibliographie ;

l'asbl RéQuaSud ; merci à Marie-Julie Goffaux et Michel Martinez pour leur précieuse collaboration ;

l'Unité de Statistique, Informatique et Mathématique appliquées de la Faculté Universaire des Sciences Agronomiques de Gembloux (FUSAGx) ; je pense spécialement à Josiane Austraet et à Guylaine Delaplace-Mélon ;

la Direction du CRA-W qui m'a apporté son soutien. Merci à Mathilde Davister pour son aide au niveau administratif et à Bernadette Dury pour ses nombreuses attentions.

Merci à tous ceux et celles qui, au sein des départements du CRA-W et de la FUSAGx, m'ont assuré de leur soutien ou aidée ponctuellement ; je pense particulièrement à Dominique Vrebos, Christian Roisin, Georges Sinnaeve, Gilles Colinet, Hugues Prévot, Billo Bah Boubacar et Vincent Leemans.

Mes plus vifs remerciements s'adressent à Clotilde qui, par ses nombreux encouragements dans les moments difficiles, m'a toujours permis d'avancer.

Je tiens enfin à remercier ma famille toujours présente … en toutes circonstances.

Merci à Jean-Marc et à mes « petits et grand bonhommés » ; ces derniers m'ont, sans le savoir, changé les idées de toutes les manières possibles et imaginables, même s'ils n'ont qu'une idée très approximative de ce que signifient les termes *repos* et *patience.*

TABLE DES MATIERES

INTRODUCTION GENERALE

Cadre général de l'étude

Au cours de cette dernière décennie, divers domaines d'études, tels que l'agriculture de précision et les systèmes d'informations géographiques (SIG) se sont très fortement développés. Dès lors, la capture automatique des informations et la constitution de bases de données se sont nettement étendues avec, comme conséquence, l'acquisition de très grands ensembles de données. Par ailleurs, suite au développement de ces techniques, la demande en analyses de tout type s'est aussi accrue pour les laboratoires. Ces analyses concernent, par exemple, la composition chimique des sols, la détermination de la qualité du froment, etc.

L'automatisation de l'acquisition des données crée une situation où l'utilisateur perçoit de moins en moins facilement la signification et la grandeur réelle des données. Il éprouve des difficultés à appréhender l'adéquation de certaines d'entre elles dans le contexte du phénomène étudié. Quant aux laboratoires d'analyses, ils doivent gérer de grandes quantités de données provenant parfois d'origines différentes et qu'il convient de rassembler ; ceci peut mettre en cause la qualité finale de l'information générée.

Comme il est d'usage classique dans tout traitement statistique des données, par exemple, pour des représentations graphiques, l'exploitation de l'information à partir de bases de données nécessite en premier lieu la recherche de valeurs aberrantes. En effet, cette opération est indispensable car les observations peuvent être introduites au sein des bases de données à partir de sources différentes. Malgré toutes les mesures mises en place pour standardiser l'information, des erreurs concernant les unités, les ordres de grandeur, etc., sont souvent rencontrées.

Si on introduit une qualité supplémentaire aux données, à savoir leur caractère spatial, la détection des valeurs aberrantes engendre des difficultés supplémentaires pour lesquelles il convient de mettre au point des outils spécifiques.

Pour mieux faire comprendre la nécessité d'intégrer la notion de composante spatiale lors de la recherche de valeurs aberrantes, nous présentons l'exemple qui a été à l'origine de l'étude, à savoir la base de données SOLS de *RéQuaSud*. Cette base de données est très importante car les données vont être exploitées à plusieurs reprises dans le cadre de cette étude.

Présentation de la base de données de RéQuaSud

Le réseau *RéQuaSud* a été créé en 1989 afin d'offrir un service d'analyses à la disposition des praticiens (agriculteurs, vulgarisateurs, etc.) pour leur fournir des conseils les plus efficaces possibles dans les secteurs agricole et agroalimentaire. Le réseau a notamment pour but l'amélioration et la promotion de la qualité des analyses et des produits. Il se structure en plusieurs chaînes, rassemblant différents laboratoires, en fonction des types d'analyses et des activités effectuées par ceux-ci (Anonyme, 2003).

Le réseau *RéQuaSud* s'occupe notamment d'aspects pédologiques à travers ce qui a été appelé la *chaîne Minérale sols de RéQuaSud* et que nous nommons dans ce travail *chaîne SOLS*. Depuis 1982, des analyses d'échantillons[1] de terre provenant de l'ensemble du territoire de la Région wallonne (partie Sud de la Belgique) ont été réalisées dans différents laboratoires. Depuis 1994, les résultats de ces analyses sont centralisés au sein de la base de données SOLS qui contient, à l'heure actuelle, plus de 300.000 enregistrements.

Chaque enregistrement comprend les résultats des analyses suivantes : pHKCl, C organique, N, Na, K, Mg, Ca, Mn, Fe, Cu, Zn[2], P ainsi que des variables relatives au lieu d'extraction de l'échantillon de sol, à la texture et à l'occupation des sols, aux types de cultures (actuelles et précédentes), etc.

Le laboratoire de Géopédologie de l'Unité « Sol, Ecologie, Territoire » de la Faculté universitaire des Sciences agronomiques de Gembloux est le laboratoire de référence pour les analyses de sol tandis que la base de données est centralisée à la Section Biométrie, Gestion des données et Agrométéorologie du Centre wallon de Recherches agronomiques (Gembloux).

Une étude approfondie de l'ensemble de ces données permet d'appréhender les propriétés physiques et chimiques des terres agricoles par type de sols et donc d'optimiser la gestion des sols grâce à la connaissance de leur potentiel de fertilité. Les agriculteurs possèdent ainsi les informations utiles pour décider judicieusement de l'affectation de leurs parcelles et de l'application des fertilisants.

L'extraction d'informations de la base de données passe, avant toute analyse, par une validation des données qui comporte une phase d'identification ou de détection des valeurs suspectes. Cette phase est parfois complexe, tel est le cas par exemple pour l'élément calcium (Ca) qui présente, d'une part, une très forte variabilité d'un type de sols à l'autre et, d'autre part, des valeurs particulièrement élevées.

---

[1] Pour le pédologue et le statisticien, le terme d'échantillon recouvre des concepts différents. Dans ce travail, lorsque le contexte ne permettra pas d'établir clairement dans quel sens le terme est employé, nous lèverons l'ambiguïté en parlant d'*échantillon de sol* pour désigner les quantités de terre qui ont été soumises à l'analyse chimique et nous conserverons le terme de d'*échantillon* pour la notion statistique correspondante.

[2] Les analyses relatives au Fe, Cu, Zn ne sont pas effectuées sur tous les échantillons.

Lors de l'introduction de nouvelles observations dans la base de données, celles-ci doivent faire l'objet de comparaison par rapport à des limites préalablement fixées par commune ou par groupe de communes présentant une homogénéité au niveau des sols.

Problématique

De manière plus précise, examinons les difficultés rencontrées dans le cas de la base de données SOLS de *RéQuaSud* pour la détection de valeurs aberrantes.

1° Le premier problème concerne la détection classique des valeurs aberrantes. Comme nous le verrons dans la première partie de ce travail, les démarches de détection de valeurs aberrantes montrent toute l'importance de l'hypothèse de normalité des distributions des populations-parents (Barnett et Lewis, 1994).

2° Cette hypothèse de normalité n'est pas nécessairement vérifiée pour les différents éléments étudiés lors d'analyses de sols. Les **distributions** pour les éléments chimiques considérés sont très **dissymétriques** avec un étalement vers la droite très marqué (Sichel, 1973; Houghton, 1988; Sichel *et al.*, 1995; Caers *et al.*, 1996). La présence d'un grand nombre de valeurs très élevées ou extrêmes à droite de la distribution rend difficile l'estimation des paramètres nécessaires à la réalisation des tests de détection de valeurs aberrantes. Avec les tests classiques de discordance, basés généralement sur la distribution normale, la plupart de ces valeurs sont considérées comme aberrantes. Il en résulte un sentiment d'insatisfaction de la part des acteurs de terrain qui n'acceptent pas le rejet de valeurs extrêmes pouvant avoir une réelle signification. Ce nombre élevé de valeurs extrêmes, caractéristique majeure des distributions observées, est un élément crucial à prendre en compte dans notre recherche.

3° Sachant que la majeure partie des travaux sur les valeurs extrêmes concernent rarement la partie gauche des distributions et se concentrent essentiellement sur la partie droite (Beirlant *et al.*, 1996), les **valeurs aberrantes situées à gauche** des distributions doivent également faire l'objet de toute l'attention.

4° Venons-en au problème qui concerne la **contrainte de contiguïté spatiale** ou *contrainte spatiale*. Elle correspond par définition, dans le domaine de la géographie, à la nécessité de respecter le regroupement d'unités géographiques contiguës, c'est-à-dire des lieux géographiquement voisins (Beghin, 1979 ; Foguenne, 1994). Cependant, la contrainte spatiale n'est pas forcément liée à une unité géographique bien définie. Tel est le cas en géostatistique dans le cas du krigeage ; un point du territoire peut être comparé à l'ensemble des points du territoire qui l'entoure constituant ainsi une contrainte spatiale (Legendre et Legendre, 1984a).

Dans le cas général d'unités géographiques contiguës, l'exemple de la carte *des associations de sols* de Belgique (Maréchal et Tavernier, 1974) et de sa numérisation[3] ou de la *carte des principaux types de sols de la Région wallonne*[4] (Belgique), actuellement en cours de finalisation, illustre encore mieux la problématique rencontrée. En effet, des groupements spatiaux de sols y ont été réalisés à partir de caractéristiques identiques et ils ont été rassemblés au sein d'associations de sols, pour la première carte citée, ou au sein des principaux types de sols, pour la seconde. Ainsi, ces cartes permettent de distinguer clairement les associations ou les types de sols présents dans chaque commune.

La carte des principaux types de sols de la Région wallonne de la figure 1 permet de montrer qu'au sein d'une même commune (délimitée par un contour noir) d'une région agricole de Wallonie (Condroz), plusieurs types de sols peuvent être rencontrés. Par contre, un même type de sols peut se retrouver sur plusieurs communes, principalement si celles-ci sont voisines.

Ces types de sols constituent donc des contraintes spatiales à inclure dans notre réflexion afin de construire un système cohérent de détection de valeurs aberrantes. En effet, des observations liées à un type de sols donné (par exemple pour un type de sols à pH élevé) peuvent être tout à fait aberrantes pour un autre type de sols (type de sols à pH faible). D'autre part, tout élément lié, par exemple à la composition chimique des sols, et étudié au sein de chaque type de sols, se présente sous la forme d'une distribution de probabilité. Au sein d'une même commune, des distributions différentes sont rencontrées d'un endroit à l'autre, et, pour des communes voisines, situées sur des types de sols similaires, des distributions présentant des paramètres comparables sont observées.

Malheureusement, l'information relative aux types de sols n'apparaît pas dans la base de données SOLS de *RéQuaSud* et aucune information précise dans un système de coordonnées géographiques n'est donc disponible. En effet, les analyses ne sont pas référencées à la parcelle dans laquelle l'échantillon de sol a été prélevé mais elles le sont dans la commune dans laquelle se situe cette parcelle. Cependant, les types de sols existants sont connus par commune, grâce à la carte des principaux types de sols de la Région wallonne.

---

[3] Carte numérique réalisée à partir de la numérisation des différentes associations de sols selon la légende de la carte de base établie par R. Maréchal et R. Tavernier en 1974. Les associations de sols correspondent à des unités pédo-cartographiques constituées d'un sol principal (dominant) et d'un certain nombre de sols associées, se présentant dans une configuration géographique identifiable avec des proportions définies (FAO, 1994 ; Borgers, 2005). Soixante-deux associations de sols sont répertoriées en Belgique dont quarante-sept en Région wallonne. Echelle : 1/500.000.

[4]Projet de création de la carte des principaux types de sols de la Région wallonne piloté par la Direction Générale de l'Agriculture (Ministère de la Région wallonne) et mise en œuvre par convention par la Faculté universitaire des Sciences agronomiques de Gembloux (Bracke et Veron, 2002; Bah et Veron 2005, Bah *et al.*, 2005).

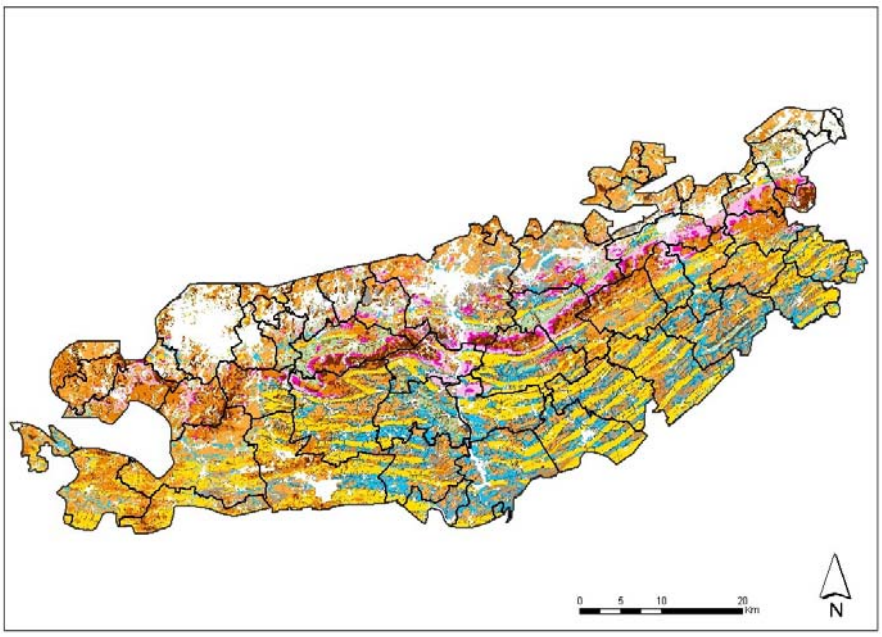


Figure 1. Entités communales (contour noir) et principaux types de sols pour une région agricole de la Région wallonne (Condroz).

5° Un problème nouveau qui découle de la présence de divers types de sols dans une même commune est celui **du mélange éventuel de plusieurs distributions** dissymétriques au sein de la base de données. Par exemple, si la « commune », identifiée par son code postal, est la source de l'information spatiale, alors, il est fort probable qu'une seule distribution apparaisse dans le cas d'une commune très homogène, c'est-à-dire où l'on trouve un seul type de sols. Par contre, lorsque, pour une commune dans laquelle plusieurs types de sols sont présents, il est probable qu'il existe un mélange de plusieurs distributions.

Dans la littérature, des études théoriques tenant compte de mélanges de distributions existent. C'est notamment le cas pour les distributions exponentielles et log-normales (Rocke, 1992; Barnett et Lewis, 1994). Cependant, pour d'autres distributions dissymétriques, peu d'informations sur les mélanges sont publiées.

En présence de mélanges de distributions, il ne nous semble pas correct de travailler à partir de l'ensemble de la distribution résultant du mélange des distributions initiales. En effet, quand plusieurs populations sont mélangées, la variance de la population globale est en général surestimée. Les limites calculées sont alors plus larges et moins fiables pour détecter les valeurs aberrantes. Il existe également un biais entre les limites calculées pour les distributions initiales et les limites calculées à partir du mélange de distributions.

Le calcul des paramètres, à partir de l'ensemble des observations, a pour conséquence d'attacher trop d'importance aux valeurs centrales de la distribution et de rejeter beaucoup de valeurs extrêmes issues, entre autres, du mélange de distributions. Pour éluder ce problème de mélanges de distributions, il est envisageable de calculer les paramètres à partir des queues des distributions et de ne considérer que la distribution située dans la partie droite ou gauche du mélange. Cette manière d'aborder le problème nécessite la recherche du nombre optimal d'observations à prendre en compte pour estimer les paramètres au niveau de la queue de la distribution. Ceci entraîne également de trouver la meilleure méthode d'estimation des paramètres dans le cas de mélanges. Cette méthode optimale permet d'obtenir le biais et la variance les plus faibles.

6° Comme signalé précédemment (4°), la recherche de valeurs aberrantes est basée sur une partie non connue (types de sols) et une partie connue (signalétique : code postal, région agricole, occupation du sol, types de cultures, etc.). De cette information connue, liée à la signalétique des enregistrements, on peut trouver une autre contrainte dont il faudrait aussi tenir compte et qui correspond à la ***contrainte liée à la structuration des données***, ou tout simplement à la *structure des données*. En tenant compte de la structure des données, des distinctions devraient être établies. A titre d'exemple, les terres de cultures et les prairies correspondent à deux types distincts d'occupations de sols.

Cette contrainte liée à la structure des données doit être intégrée à la contrainte spatiale. En effet, les informations telles que la « région agricole » (Ardenne, Condroz, Hesbaye, etc.) ou la « commune » correspondent à des contraintes de données structurées *a priori* mais cette signalétique présente une connotation spatiale, liée à la présence des types de sols. Par exemple, dans le cas de la région agricole de l'Ardenne, une cohérence spatiale pour les données structurées de cette région agricole doit exister. Ceci signifie que pour un type de sols présentant un pH élevé sur une terre de prairie en Ardenne doit correspondre un type de sols sur une terre de culture avec pH élevé. Il existe donc une cohérence entre les types d'occupation de sol, par exemple culture–prairie, dont il faut tenir compte au sein d'une même région agricole, celle-ci étant liée à la présence de types de sols déterminés.

Ce problème de la structure des données ne sera cependant pas pris en compte dans ce travail. Il pourrait faire l'objet d'une étude complémentaire.

7° Etant donné la nature des problèmes posés, le dernier point à retenir concerne la **facilité de mise en oeuvre** de la méthode que nous proposons. La méthode de détection des valeurs aberrantes doit être facilement et rapidement applicable au vu de la quantité de données contenues dans les bases de données, et en tenant compte des problèmes envisagés.

– Introduction générale –

Objectifs et plan du travail

1° L'objectif général de ce travail est de proposer une méthode opérationnelle de détection de valeurs aberrantes, applicable sur de grands ensembles de données avec contraintes spatiales.

Nous cherchons plus particulièrement à développer une méthode originale de détection de valeurs aberrantes au sein de distributions dissymétriques issues de mélanges de populations ; le type de distribution devant être identifié grâce à ce travail méthodologique. Cette méthode doit tenir compte de contraintes spatiales liées, par exemple, à la présence de types de sols au sein de communes.

Afin d'atteindre l'objectif du travail, il est nécessaire de mettre en place une méthode qui permette de déterminer, de manière optimale, les valeurs limites à partir desquelles une valeur à intégrer dans une base de données est statistiquement acceptée ou rejetée en suivant une cohérence spatiale. Afin d'éluder le problème des mélanges, ces valeurs sont déterminées à partir des queues de distributions dissymétriques. Avant toute intégration dans la base de données, les échantillons de sol font l'objet d'un contrôle sur base de leur situation géographique. L'objectif ultime est de constituer un référentiel par entité géographique, telle que des communes ou des groupements de communes voisines.

La partie innovante de ce travail réside en l'étude des queues de distributions pour répondre au problème de mélanges en présence de la contrainte spatiale.

Complémentairement, nous formulons de manière synthétique, dans le tableau 1, les hypothèses émises lors de cette recherche et les objectifs spécifiques qui y sont associés.

2° Ce travail est scindé en deux parties. La *première partie* est consacrée à l'étude bibliographique et se concentre autour des trois thèmes suivants :

- la détection de valeurs aberrantes dans le cas général et choix méthodologiques (chapitre 1) ;
- la théorie portant sur les distributions à forte dissymétrie qui, combinée aux démarches classiques, offre de nouvelles possibilités de détection de valeurs aberrantes dans le cas de distributions très dissymétriques (chapitre 2) ;
- le problème des contraintes spatiales (chapitre 3).

La *deuxième partie* est consacrée, d'une part, à la recherche d'une méthode appropriée de détection de valeurs aberrantes (méthodologie), (chapitre 4) et, d'autre part, à l'application de la méthode retenue à un sous-ensemble de la base de données de *RéQuaSud* pour la partie droite (chapitre 5) et pour la partie gauche (chapitre 6) des distributions étudiées.

Enfin, nous clôturons ce document par une discussion générale et des conclusions ainsi que la suggestion de nouvelles voies de recherche.

Tableau 1. Hypothèses de recherche et objectifs spécifiques associés.

Hypothèses	Objectifs spécifiques
L'utilisation de distributions dissymétriques différentes à droite et à gauche des distributions rend plus performante la détection de valeurs aberrantes.	Traiter les parties droite et gauche des distributions de manière séparée afin de trouver la distribution la plus appropriée pour chaque partie.
Il est possible de détecter des valeurs aberrantes dans des populations mélangées en fixant des valeurs limites au-dessus desquelles les valeurs sont considérées comme aberrantes. Ces valeurs limites peuvent être déterminées à partir des paramètres des queues des distributions dissymétriques.	- Considérer la distribution située dans la partie droite ou gauche du mélange et calculer les paramètres à partir des queues de distributions avec des niveaux de troncature différents - utiliser ces paramètres pour classer les entités communales spatialement. - Choisir la distribution qui fournit les meilleures propriétés d'ajustement (« robustesse »). - Choisir le niveau de troncature qui donne le meilleur potentiel de détection de valeurs aberrantes – vérifier si le niveau de troncature doit être adapté à l'effectif des populations.
La mise en place d'un système de détection de valeurs aberrantes en intégrant la contrainte spatiale est plus robuste, plus cohérente que les méthodes qui considèrent que les populations sont homogènes dans l'espace (méthode actuellement appliquée au sein de *RéQuaSud*).	- Créer une matrice de contiguïté, matrice *communes-types de sols* qui permette de tenir compte du voisinage des entités communales et de la présence de types de sols différents. - Classer spatialement les entités communales par groupes d'entités qui présentent des caractéristiques similaires. - Valider la nouvelle procédure par rapport à la méthode utilisée au sein de *RéQuaSud.*

I. PREMIERE PARTIE : APPROCHE BIBLIOGRAPHIQUE

1. DETECTION DES VALEURS ABERRANTES

1.1. Introduction

Afin de garantir la qualité des informations extraites à partir de bases de données, une recherche de valeurs suspectes ou aberrantes doit être effectuée avant toute analyse ou exploitation de la base de données, quel que soit le domaine concerné. Les méthodes de détection des valeurs aberrantes sont donc essentielles dans la gestion des bases de données, et dans le cas qui nous préoccupe, spécialement pour l'intégration de nouvelles observations, de manière à fournir des informations de très grande qualité.

Les observations *non représentatives* ou *aberrantes*, appelées en anglais *outliers*, font l'objet d'un grand intérêt dans la littérature. Ces données ont été considérées comme une source de contamination, déformant l'information obtenue à partir des données brutes. Ce problème est incontournable pour toutes les personnes qui manipulent des données et doivent juger de la manière de traiter ces valeurs anormales. Il est naturel de rechercher les moyens d'interpréter ou de caractériser les valeurs aberrantes et de mettre au point des méthodes pour les traiter, soit en les rejetant afin de restaurer les propriétés initiales des ensembles de données, soit en adoptant des méthodes qui diminuent leur impact au cours des analyses statistiques (Barnett et Lewis, 1994).

Depuis plus d'un siècle, un large éventail de méthodes d'analyses statistiques de plus en plus précises a été élaboré pour tester des hypothèses concernant des paramètres déterminés ou pour estimer la validité de certains modèles. Cette grande sophistication dans la conception et l'utilisation de méthodes statistiques nécessite une évaluation fiable de l'intégrité d'ensembles de données.

L'évolution dans la manière d'appréhender le problème du traitement des valeurs aberrantes est très nette. En 1852, Peirce, le premier auteur à s'intéresser au problème des valeurs anormales disait, de manière très naïve et restrictive, que *dans presque toutes les séries de données, il y a des observations qui diffèrent tellement des autres, qu'elles servent uniquement à rendre l'expérimentateur perplexe et à l'induire en erreur* (Barnett et Lewis, 1994). Dans certains cas, l'expérimentateur peut être tenté de ne pas rejeter la valeur aberrante mais de l'accepter comme une indication intéressante. Tel est le cas lors d'essais variétaux très prometteurs ou lors de prospections minières. Il n'est pas approprié d'adopter une attitude radicale, soit de rejet, soit d'inclusion systématique des valeurs aberrantes. La première attitude peut entraîner la perte d'informations réelles tandis que, dans le cas de l'acceptation des valeurs aberrantes, il y a un risque de contamination. En fonction des circonstances, il existe des méthodes, dites *robustes*, qui prennent en compte toutes les données mais minimisent

l'influence des valeurs aberrantes. Ces méthodes sont considérées comme *s'adaptant*[5] *aux valeurs aberrantes* ou les *accommodant*.

Ces considérations nous mènent à réaliser des distinctions bien claires entre les objectifs des analyses statistiques et la manière de considérer les données (paragraphe 1.2). Barnett et Lewis (1994) dressent une classification des types de questions auxquelles il faut réfléchir lors de l'étude de valeurs aberrantes. Il est nécessaire de faire les distinctions suivantes :

- entre les causes déterministes ou aléatoires d'apparition de valeurs aberrantes, ces notions sont développées au paragraphe 1.2.3 concernant les sources de variabilité ;
- entre les différents objectifs à atteindre lors de l'étude des valeurs aberrantes (paragraphe 1.2.4) ;
- entre les différents modèles de probabilité spécifiques (paragraphe 1.2.5) ;
- entre les données univariées (paragraphe 1.3) et multivariées ;
- entre les valeurs aberrantes simples ou multiples (paragraphe 1.3.2).

Comme nous ne nous intéressons qu'au domaine univarié lors la mise en place de la méthode de détection de valeurs aberrantes (1.4), le problème multivarié n'est pas développé dans ce document. De plus amples informations à ce sujet ont été exposées dans l'article de Planchon (2005). De même, les notions relatives au traitement des valeurs aberrantes dans le cas de situations plus structurées (modèles de régression, analyse de la variance, séries chronologiques, contraintes spatiales) sont exposés dans la publication citée ci-dessus.

1.2. Considérations générales sur l'étude de valeurs aberrantes

1.2.1. Introduction

Comme nous l'avons présenté au paragraphe précédent, la manière d'aborder le problème des valeurs aberrantes a évolué au cours du temps. En découlent les difficultés liées à la définition d'une valeur aberrante en fonction de l'avancement des théories statistiques (paragraphe 1.2.2).

Toute étude sur les valeurs aberrantes se doit de prendre en compte la nature et l'origine de celles-ci afin de les identifier et de les traiter de la manière la plus adéquate. De manière générale, les valeurs aberrantes sont de nature aléatoire ou déterminées et peuvent trouver leur origine à différents niveaux (paragraphe 1.2.3.a). Comme nous nous préoccupons de données chimiques géoréférencées, des valeurs aberrantes spécifiquement liées à ces caractéristiques sont rencontrées (paragraphe 1.2.3.b).

L'objectif poursuivi lors de l'étude des valeurs aberrantes est déterminant pour le traitement statistique de celles-ci. Divers objectifs à envisager sont présentés au paragraphe 1.2.4.

[5] En anglais : *to accommodate the outlier*.

Dans les échantillons univariés, les observations suspectes qui peuvent être déclarées comme valeurs aberrantes sont évidentes par simple inspection car la valeur aberrante correspond à l'extrême dans l'échantillon. Néanmoins, des valeurs anormales pour la distribution normale ne le sont pas nécessairement pour une distribution dissymétrique. L'importance du modèle de probabilité pris en compte est détaillée au paragraphe 1.2.5.

1.2.2. Définitions

De nombreux auteurs ont cherché à définir le terme de *valeur aberrante* et les définitions fournies ont évolué au cours du temps. Grubbs (1969) définit une valeur aberrante comme étant *une observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel il apparaît*. Carletti (1988) s'intéresse aux *valeurs anormales* qu'il définit comme étant *une valeur qui paraît suspecte parce qu'elle s'écarte d'une façon importante des autres valeurs de la variable étudiée ou ne semble pas respecter une norme ou une relation bien définie*. Munoz-Garcia *et al.* (1990) proposent également une définition du terme *valeur aberrante* et tentent d'éviter le côté subjectif en ajoutant la condition que l'observation devrait dévier nettement du comportement général par rapport au critère sur lequel l'analyse est réalisée (Barnett et Lewis, 1994).

Barnett et Lewis (1994) définissent une valeur aberrante dans un ensemble de données comme étant une *observation (ou un ensemble d'observations) qui semble être inconsistante avec le reste des données* ou d'une autre manière, il y a une valeur aberrante *lorsque l'une ou l'autre observation d'un ensemble de données, détonne ou n'est pas en harmonie avec les autres observations*. Ce qui caractérise la valeur aberrante, c'est son impact sur l'observateur. Selon les auteurs, l'observation ne va pas sembler extrême mais va apparaître, dans un certain sens, comme étant « *étonnamment extrême »*.

L'expression « *semble être inconsistante »* est cruciale car elle émane d'un jugement subjectif de la part de l'observateur qui s'intéresse aux données. Ce qui est important c'est de savoir si les données font vraiment partie de la population principale. Si ce n'est pas le cas, elles sont alors considérées comme des *contaminants*, définis comme étant des *observations issues d'autres populations*. Les contaminants peuvent poser des problèmes lors de l'application de méthodes inférentielles à partir de la population d'origine. Il est clair que tout contaminant se trouvant au milieu d'un ensemble de données ne va pas être « visible » et il est improbable qu'il affecte sérieusement le processus d'inférence. Néanmoins, si de telles observations, étrangères à la population principale, sont situées dans les queues des distributions, elles peuvent, par leur nature de contaminants, causer des difficultés dans la tentative de décrire la population et déformer l'estimation des paramètres de la population.

Barnett et Lewis (1994) ont affiné leur définition en faisant intervenir la notion de modèle de probabilité : une valeur aberrante *est une observation qui apparaît douteuse dans le contexte d'un modèle de probabilité, désigné initialement pour expliquer le processus de génération des données*. Des considérations sur l'importance des modèles de probabilité sont exposées au paragraphe 1.2.5 qui traite des valeurs aberrantes en relation avec ceux-ci.

Everitt (2002) tient également compte des modèles de probabilité sous-jacents dans la définition suivante : les valeurs aberrantes correspondent à des *observations qui semblent dévier de manière importante des autres observations de la population de laquelle elles proviennent, ces observations semblent être inconsistantes avec le reste des données, en relation avec un modèle supposé connu*.

A partir de ces définitions, on se rend compte qu'il est nécessaire de définir également d'autres termes qui sont utilisés de manière courante et qui ont tendance à semer la confusion dans les esprits.

Le terme *valeurs extrêmes*[6] est défini par Everitt (2002) comme *les valeurs les plus grandes et les plus petites parmi un ensemble d'observations*. Notons que cette définition ne donne aucune idée sur le nombre de valeurs à prendre en compte, soit à droite, soit à gauche de la distribution.

Barnett et Lewis (1994) ont distingué les notions de valeurs aberrantes, d'observations extrêmes et de contaminants à l'aide d'une figure dont une adaptation est présentée à la figure I.1. Soit x_1, x_2,..., x_n, un échantillon aléatoire univarié de taille n, provenant d'une distribution F, et soit $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$, les données ordonnées dans l'ordre croissant. Les valeurs $x_{(1)}$ et $x_{(n)}$ sont respectivement l'observation extrême inférieure et supérieure. Les auteurs présentent les valeurs extrêmes comme étant l'observation minimale et l'observation maximale de l'échantillon.

Le fait de déclarer qu'une observation extrême est une valeur aberrante dépend de la manière par laquelle elle apparaît en fonction du modèle F. En effet, dans la figure I.1(a), ni la valeur $x_{(1)}$, ni $x_{(n)}$ ne semblent correspondre à une valeur aberrante. Par contre, dans la figure I.1(b), $x_{(n)}$ est une valeur aberrante *supérieure* ou située au niveau de la queue droite de la distribution. La valeur $x_{(1)}$ cause également quelques problèmes et peut être considérée comme suspecte pour la queue gauche de la distribution. Ainsi, on voit que les valeurs extrêmes peuvent être ou ne pas être des valeurs aberrantes. Toute valeur aberrante est par contre toujours une valeur extrême de l'échantillon dans le cas univarié.

[6] En anglais : *extreme values.*

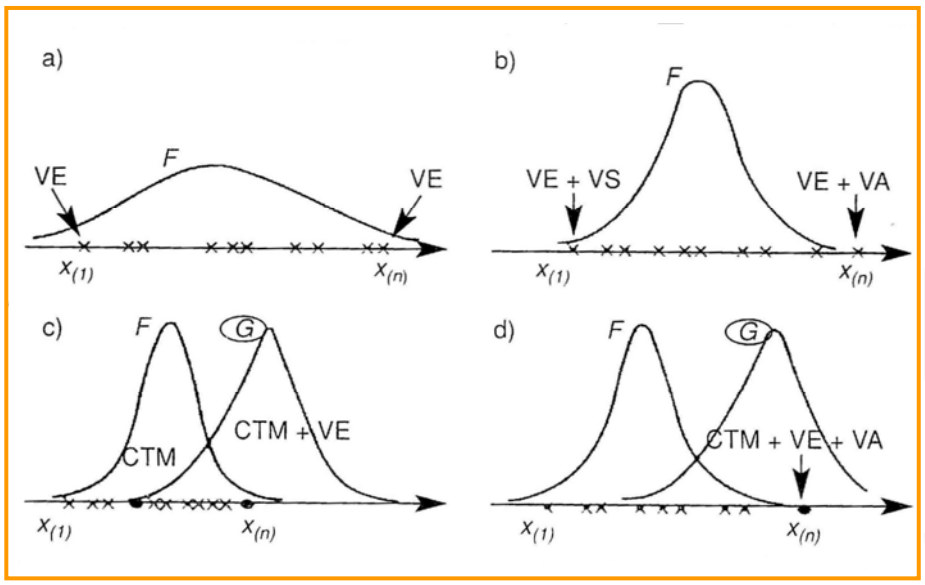


Figure I.1. Définition des termes : valeurs extrêmes (VE), valeurs aberrantes (VA), valeurs suspectes (VS) et contaminants (CTM) à partir de deux distributions notées F et G (figure adaptée à partir de Barnett et Lewis, 1994).

Si toutes les observations ne proviennent pas de la distribution *F* mais que l'une ou l'autre est issue de la distribution *G*, de moyenne plus élevée que *F*, les observations de *G* sont considérées comme des contaminants. De tels contaminants peuvent apparaître comme étant extrêmes mais ce n'est pas forcément le cas. La figure I.1(c) montre deux contaminants indiqués par un rond noir ; celui situé à droite est l'extrême supérieur tandis que celui de gauche se trouve au milieu de l'échantillon. Néanmoins, $x_{(n)}$, bien qu'il soit extrême et contaminant, n'est pas une valeur aberrante.

Enfin, dans la figure I.1(d), la valeur extrême $x_{(n)}$, correspond à un contaminant qui est également une valeur aberrante. Une valeur aberrante peut donc être la manifestation de la présence d'un contaminant. Ces diverses situations indiquent la complexité de l'étude de valeurs anormales et la difficulté de définir le type d'observation rencontré de manière précise.

Nous désirons ici introduire une notion qui est à la base de notre démarche. Dans le cas des échantillons de sol, de nombreuses valeurs extrêmes sont rejetées alors que les données obtenues suite aux analyses sont sûres et fiables. Une valeur extrême correspond alors à une valeur qui est statistiquement discordante mais qui n'est pas surprenante pour le praticien (dans le cas de l'hypothèse de la distribution normale). Par rapport à ce qui vient d'être cité ci-dessus, notons que l'analyse en laboratoire peut être sûre et fiable mais en fait, c'est l'échantillon qui ne l'est peut-être pas. Dans le cas des analyses de sols, c'est l'échantillonnage qui est la source la plus importante d'erreurs.

La notion d'*observation influente*[7] est définie par Everitt (2002) comme étant une *observation qui a une influence disproportionnée sur un ou plusieurs aspects de l'estimateur d'un paramètre*, en particulier, les coefficients de régression. D'après Cook et Weisberg (1980), les observations influentes sont *celles pour lesquelles les caractéristiques de l'analyse sont altérées de manière considérable quand elles sont supprimées*. Cette influence, peut être due à des différences par rapport aux autres observations de la variable explicative, à une valeur extrême pour la variable à expliquer ou à une combinaison des deux. Barnett et Lewis (1994) signalent que les valeurs aberrantes sont souvent des observations influentes. Néanmoins, les notions de valeurs aberrantes et d'observations influentes ne sont pas issues de concepts semblables. Une valeur aberrante peut altérer nettement l'estimation d'un paramètre ou le résultat d'un test spécifique mais cette éventualité n'est pas la base de l'identification d'une valeur aberrante. A la différence des observations influentes, une valeur aberrante tout à fait évidente n'a pas d'effet net sur une estimation particulière ou un test lorsqu'une méthode d'accommodation appropriée est utilisée.

Le terme *valeur suspecte*[8] correspond, selon Barnett et Lewis (1994), à une valeur moins extrême qu'une valeur jugée aberrante de manière statistique. Les définitions de valeurs suspectes et aberrantes sont complétées au paragraphe 1.3.2.b qui concerne les *tests de discordance*. Par définition, ces tests correspondent à des méthodes statistiques permettant de tester une valeur aberrante afin de déterminer si elle doit être gardée ou rejetée.

Anscombe (1960) définit également les valeurs aberrantes en fonction des sources d'apparition de celles-ci (paragraphe 1.2.3).

Quand on traite des échantillons à caractère multivarié, les méthodes statistiques univariées montrent bien vite leurs limites. Le problème est de définir quel échantillon est aberrant. Il est facile de détecter quelles valeurs d'une variable bien spécifique de l'échantillon sont aberrantes mais il est difficile de déterminer quels sont les échantillons aberrants. La définition de valeurs aberrantes inclut donc les notions de valeurs aberrantes pour une variable et les échantillons aberrants ; les échantillons aberrants correspondant à ceux qui possèdent un nombre trop élevé de valeurs aberrantes et qui ne partagent pas les relations entre les variables de la population. A titre d'exemple, Zhang *et al.* (1998) ont examiné les données d'une étude géologique en Suède, initiée en 1982 par un programme national de cartographie dont l'objectif était de produire un atlas géochimique détaillé du pays. Lors de l'analyse préliminaire des données, les auteurs ont jugé que les échantillons de plus de deux variables aberrantes étaient aberrants. Cependant, il reste encore le problème du traitement des échantillons comprenant une ou deux valeurs aberrantes. Des techniques

[7] En anglais : *influential observation.*
[8] En anglais : *straggler.*

d'analyse en composantes principales ont été utilisées pour résoudre ce problème.

Davies et Gather (1993) définissent les valeurs aberrantes dans le cas multivarié comme étant des observations qui se classent dans une région extrême de la distribution F, caractérisée par (θ, α_n), où θ correspond à la vraie valeur, inconnue, du paramètre qui permet de déterminer F et α_n est une petite valeur, spécifiée par l'observateur, équivalente à la probabilité qu'une observation tombe dans la région extrême. Il est possible qu'une observation unique de la distribution F soit désignée comme étant une valeur aberrante. Ce concept est examiné par rapport au *point de rupture*[9] d'un échantillon fini (Barnett et Lewis, 1994).

1.2.3. Nature et origine des valeurs aberrantes

a. Nature et origine des valeurs aberrantes dans le cas général

Une classification des différentes manières par lesquelles les valeurs aberrantes peuvent apparaître a été discutée dans la littérature par divers auteurs tels que Anscombe (1960), Grubbs (1969), Barnett (1978), Hawkins (1980), Barnett (1983) et Beckman et Cook (1983). Des exemples concrets issus de la base de données de *RéQuaSud* permettent d'illustrer ces diverses origines.

Lors de la collecte de données, différentes sources de variabilité peuvent être rencontrées dont (a) la variabilité inhérente, (b) l'erreur de mesure et (c) l'erreur d'exécution (Barnett et Lewis, 1994).

(a) La *variabilité inhérente* correspond à l'expression de la manière par laquelle les observations varient de manière aléatoire à travers la population. Une telle variation est une caractéristique naturelle de la population. Elle est incontrôlable et reflète les propriétés de la distribution d'un modèle de base qui décrit correctement la génération des données. Par exemple, les différentes teneurs en magnésium d'échantillons issus d'une commune donnée vont refléter la variabilité propre à la population de cette commune. Cette notion de distribution des données est détaillée au paragraphe 1.2.5 lié aux modèles de probabilité dans le cas univarié.

(b) En ce qui concerne l'*erreur de mesure*, ou l'erreur liée à la méthode de mesure, des inadéquations au niveau des instruments de mesure surimposent un degré plus élevé de variabilité au facteur inhérent. L'arrondi des valeurs obtenues ou les erreurs d'enregistrement correspondent également à des erreurs de mesure. Cette erreur est liée à des circonstances bien déterminées. L'erreur de mesure peut également être de nature aléatoire, cette variabilité correspond alors à l'incertitude de la méthode de mesure.

[9] Le *point de rupture* d'un échantillon correspond à une mesure de l'insensibilité d'un estimateur à la présence de plusieurs valeurs aberrantes au sein de l'échantillon. En bref, cela correspond à la plus petite fraction de contaminants nécessaires pour causer une modification plus ou moins grande lors d'estimation. En anglais : *breakdown point*.

Par exemple, dans le cas des modèles biologiques de croissance, soit de croissance de rameaux, pour des mesures effectuées à intervalles réguliers dans le temps, les valeurs observées doivent être, d'une fois à l'autre, supérieures à l'observation précédente. Des considérations biologiques conduisent ainsi à la conclusion que certaines observations suspectes sont clairement impossibles. Un autre exemple d'erreur d'enregistrement de données relatives à des observations de pH consisterait à avoir des valeurs comprises entre 0 et 3 ou supérieures à 9, ce qui est tout à fait impossible pour des échantillons de sol.

Ces exemples illustrent l'effet de facteurs non statistiques et, dans une plus ou moins grande mesure, au manque d'attention dans l'enregistrement ou la présentation des données. Quelques contrôles de ce type de variabilité sont possibles et facilement réalisables.

(c) Une autre source de variabilité apparaît dans la collecte imparfaite des données, c'est l'*erreur d'exécution*, qui est également liée à des circonstances bien déterminées. Par inadvertance, un échantillon peut être biaisé ou peut inclure des individus qui ne sont pas vraiment représentatifs d'une population-parent déterminée. Des erreurs d'exécution lors de la manipulation ou dans l'assemblage des données peuvent aussi mener à des valeurs aberrantes de nature déterministe. De même, des erreurs lors du traitement informatique ou des erreurs de gestion des données peuvent conduire à des observations erronées. De telles situations se présentent quand les erreurs humaines mènent à l'enregistrement évident de données incorrectes ou quand le manque de critiques vis-à-vis des facteurs pratiques entraîne des interprétations erronées. Le traitement de telles valeurs aberrantes dans ces situations n'est pas du domaine de l'analyse statistique mais du bon sens tout simplement. Il n'y a pas besoin de procédure statistique pour enlever ou remplacer la donnée suspecte qui peut être très facilement identifiée. De simples précautions peuvent réduire une telle variabilité mais il n'est pas toujours possible d'être au courant de ces erreurs d'exécution.

Une illustration d'erreur d'exécution possible, est obtenue à partir des boxplots de la figure I.2. Cette figure a été réalisée à partir de données issues de la base de données de la chaîne SOLS de *RéQuaSud* et concerne 11 communes de la zone agricole du Condroz, dont des échantillons de sols ont fait l'objet d'analyses pour la teneur en calcium disponible[10] (exprimée en grammes par 100 grammes de terre sèche ou g/100 g T.S.) (Laroche et Oger, 1999). On observe une première valeur aberrante pour la commune ayant le code postal 4530 et une deuxième valeur aberrante pour le code postal 4560. Le nombre d'observations pour la première commune est de 1002 et pour la teneur en calcium, on trouve l'observation extrême de

[10] La teneur en calcium disponible, c'est-à-dire utilisable par la plante au niveau des racines, est analysée à partir d'acétate d'ammonium 0.5N avec le chélatant EDTA ($C_{10}H_{16}N_2O_8$), 0.02M, à pH=4,65, par opposition au calcium échangeable qui est analysé à l'acétate à pH=7.

6553,0 mg/100 g T.S. On peut imaginer que cette valeur est une erreur d'encodage, c'est-à-dire une erreur d'exécution, étant donné l'écart énorme par rapport aux autres valeurs (Planchon, 2005). La valeur correcte aurait peut-être été 655,3 mg/100g ou 6553,0 mg/1000g. Malheureusement, il ne nous est pas possible de vérifier *a posteriori* cette donnée. Le même raisonnement peut être suivi pour l'observation extrême 3225,0 mg/1000g rencontrée pour la deuxième commune.

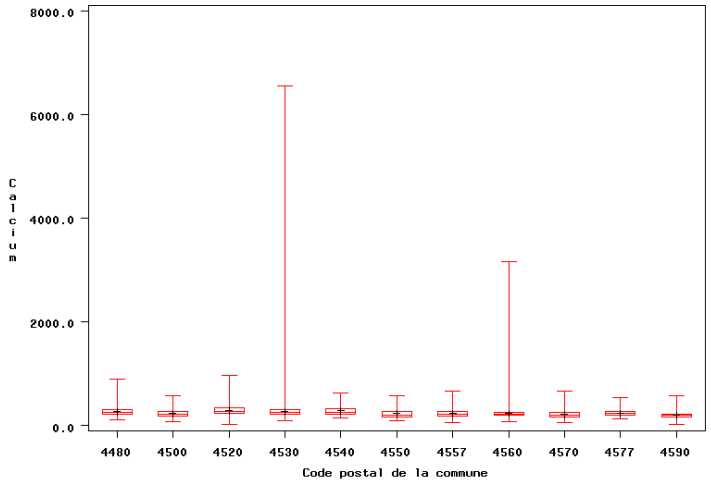


Figure I.2. Boxplots des teneurs en calcium (mg/100 g T.S.) pour 11 communes de la Région condruzienne. Identification de deux valeurs particulièrement élevées.

Dans le cas de la base de données de *RéQuaSud*, des observations mal géoréférencées au niveau du code postal peuvent induire des erreurs importantes au niveau de la base de données. Un autre exemple concernerait les mesures de pH d'échantillons de sols de la région limoneuse dans lequel se trouve, par erreur, une valeur de pH d'un échantillon de sol provenant de la région ardennaise ou en encodant des valeurs relatives à la teneur en calcium à la place des valeurs de pH. Il peut y avoir également confusion entre le code postal de la parcelle et celui de l'habitation de l'agriculteur.

De même, une signalétique qui n'est pas correcte, telle que le type de culture, la texture du sol ou l'occupation du sol peut entraîner des erreurs d'interprétation importante lors de l'analyse des résultats issus de la base de données. Ce type d'erreur lié à la structure des données est cependant très difficile à détecter mais la prise en compte de la contrainte spatiale peut compenser ce type d'erreur de signalétique.

Comme cité au paragraphe 1.2.2, en fonction du type de variabilité rencontré, Anscombe (1960), distingue, dans la terminologie, les valeurs aberrantes issues d'une large variabilité de type inhérente et les mesures élevées provenant d'erreurs de mesure ou d'erreurs d'exécution. Il appelle les

valeurs aberrantes, celles qui proviennent de la variabilité de type inhérente tandis que les autres, il les nomme les *observations fausses*. Il nous semble tout à fait évident que suite à l'exposé des différentes définitions, que cette distinction n'a pas réellement de sens et que toutes ces valeurs étonnamment extrêmes correspondent à des valeurs aberrantes.

b. Valeurs aberrantes dans le cadre d'études géochimiques

Dans une étude relative à des données de type géochimiques géoréférencées, Lalor et Zhang (2001) classent les valeurs aberrantes en trois catégories présentées ci-dessous.

Les *valeurs aberrantes d'amplitude*[11] sont considérées comme trop élevées ou trop basses comparées à la population de la majorité des échantillons. Dans le cas de valeurs aberrantes trop élevées, ces valeurs aberrantes sont issues de l'enrichissement naturel ou d'activités humaines locales.

Les *valeurs aberrantes spatiales*[12] sont généralement définies comme des observations qui sont extrêmes par rapport aux valeurs voisines (Cerioli et Riani, 1999). Ceci correspond à des observations qui ne sont pas cohérentes avec les contraintes spatiales. Dans le cas de la base de données de *RéQuaSud*, on s'attend à observer au sein d'une même zone agricole des observations très liées. De même, pour des communes avoisinantes, des caractéristiques similaires sont attendues et correspondent à des zones pédologiques semblables.

Les *valeurs aberrantes relationnelles*[13] sont définies comme des observations non conformes aux relations qui existent entre les éléments. Selon Lalor et Zhang (2001), il existe des corrélations entre les concentrations de beaucoup d'éléments, principalement pour les métaux lourds. Les valeurs qui ne suivent pas de telles relations ne sont pas toujours détectables par les méthodes classiques et peuvent conduire à des erreurs. La recherche de ce type de valeurs aberrantes est considérée par les auteurs comme étant un bon point de départ pour la détection des valeurs aberrantes. Pour la base de données qui nous préoccupe, les corrélations entre les différents éléments peuvent fournir de bonnes indications sur la nature des relations entre ceux-ci. Cependant, il n'est pas facile d'établir des relations directes entre les éléments analysés à partir d'échantillons de sols de la base de données *RéQuaSud*. En effet, d'une zone agricole, d'un type de sols ou d'une exploitation agricole à l'autre, les relations entre les éléments peuvent être différentes et ne sont dès lors pas faciles à appréhender de manière globale (Colinet, 2004).

[11] En anglais : *range outlier.*
[12] En anglais : *spatial outlier*.
[13] En anglais : *relationship outlier.*

1.2.4. Objectifs poursuivis lors de l'examen de valeurs aberrantes

Au cours des paragraphes précédents, les diverses sources de variation (variation inhérente, erreur de mesure, erreur d'exécution) qui provoquent l'apparition de valeurs aberrantes de nature différente (aléatoire ou déterminée) ont été identifiées et ont montré la complexité de l'examen des valeurs aberrantes. Les objectifs de l'étude des valeurs aberrantes dépendent très largement de l'origine et de la nature de celles-ci. Afin d'aider à la compréhension, la figure I.3 permet de visualiser clairement le schéma général de traitement des valeurs aberrantes et des objectifs poursuivis.

Pour les valeurs aberrantes de nature aléatoire, la réalisation d'un **test de discordance** doit être perçue uniquement comme la première étape de l'étude de valeurs aberrantes. En effet, en fonction des facteurs étudiés et de l'intérêt pratique de l'étude, il peut être décidé, suite à la réalisation du test, soit de *rejeter* les valeurs discordantes et de procéder à l'analyse à partir de l'échantillon modifié. D'autres possibilités sont également intéressantes. En effet, on peut choisir d'utiliser un autre modèle que celui choisi initialement. Les données de type géochimique nécessitent souvent l'utilisation des modèles de distributions dissymétriques (Sichel, 1973; Houghton, 1988; Sichel *et al.*, 1995; Caers *et al.*, 1996). Ces modèles permettent d'*incorporer* la valeur aberrante de manière non discordante.

On peut également concentrer son attention sur les valeurs aberrantes et *identifier* des facteurs non pris en compte initialement mais qui ont une grande importance pratique. Dans le cas d'expérimentations, dont le but est de rechercher des effet importants de facteurs expérimentaux, les valeurs aberrantes peuvent permettre d'identifier des caractéristiques importantes du point de vue pratique plutôt que de refléter une possible inadéquation du modèle.

Dans la pratique statistique courante, si des valeurs aberrantes sont discordantes sur base de la distribution normale, on a très vite tendance à supprimer les observations qui ne collent pas au modèle avant de procéder à une étude plus approfondie (Kruskal, 1960). Un modèle plus sophistiqué, non normal, les incorporerait peut-être de manière non-discordante. L'analyse ultérieure des données peut également faire l'objet de l'une ou l'autre forme d'**accommodation**. Ce choix est réalisé en fonction des objectifs de l'analyse statistique, car si on s'intéresse spécifiquement aux caractéristiques inférentielles d'un modèle de base, quelles que soient la présence et la nature des contaminants, les valeurs aberrantes n'ont qu'un effet de nuisance. Il est alors nécessaire d'utiliser des méthodes robustes pour minimiser leur impact. Dans ce cas, l'objectif est l'accommodation en tant que telle et aucun test de discordance n'est approprié. Le but est alors de trouver des procédures statistiques qui ne recherchent pas les valeurs aberrantes en elles-mêmes mais qui cherchent à les rendre moins importantes quant à leur influence lors de l'estimation de paramètres.

Il faut bien reconnaître que le rejet inconsidéré des valeurs aberrantes a des conséquences statistiques non négligeables pour l'analyse ultérieure de l'échantillon qui n'est plus aléatoire mais qui devient un échantillon censuré. Le remplacement des données rejetées par des équivalents statistiques implique des conséquences similaires. Les pratiques de *winsorization* par lesquelles les extrêmes les plus faibles et les plus grands sont remplacés par leurs plus proches voisins ou la réalisation de *censure/rognage*[14] vont également avoir des implications sur les distributions. Le processus de rognage consiste à utiliser un échantillon[15] dans lequel une fraction fixée, soit α, des valeurs extrêmes, basses et élevées, de l'échantillon initial sont totalement mises de côté.

La décision extrême de rejeter une valeur aberrante avant d'estimer ou de tester, repose donc sur des explications claires et tangibles de la présence de la valeur aberrante ou sur les résultats d'un test de discordance basé sur un modèle supposé connu.

Quant aux valeurs aberrantes dont la nature est déterminée, c'est-à-dire les erreurs de mesure ou d'exécution, les valeurs aberrantes peuvent être rejetées ou faire l'objet de corrections dans la mesure où celles-ci sont encore réalisables. Dans le cas de la base de données de *RéQuaSud*, aucune correction n'est possible *a posteriori* car les données sont centralisées et le retour aux données est un processus trop long et coûteux pour les laboratoires. Comme nous l'avons présenté dans l'introduction générale, l'objectif de ce travail est de détecter les valeurs aberrantes afin de les rejeter de la base de données de *RéQuaSud*. Quelle que soit la nature des valeurs anormales, les observations à introduire dans la base de données doivent dès lors faire l'objet de tests de discordance en relation avec un modèle de probabilité à déterminer pour chacun des éléments étudiés (paragraphe 1.2.5).

Il existe également de nombreuses méthodes graphiques qui permettent de signaler la présence de valeurs aberrantes. Ces méthodes peuvent avoir un impact important par la révélation de caractéristiques aberrantes des données d'une manière plus claire que les valeurs numériques apparentes.

Suite à cette présentation, on peut dire que les objectifs, lors de l'examen des valeurs aberrantes, peuvent être le rejet, l'incorporation, l'identification, l'accommodation pour les valeurs aberrantes de nature aléatoire. Pour les valeurs aberrantes de nature bien déterminée, le rejet ou le correction des données sont les deux solutions possibles. La figure I.3 permet d'illustrer la structure complexe entre les différentes sources de variation (variation inhérente, erreur de mesures, erreur d'exécution) menant à des valeurs aberrantes de nature différente (aléatoire ou déterministe) et de détailler les objectifs lors de l'examen des valeurs aberrantes.

[14] En anglais : *trimming*.
[15] En anglais : *α-trimmed sample*.

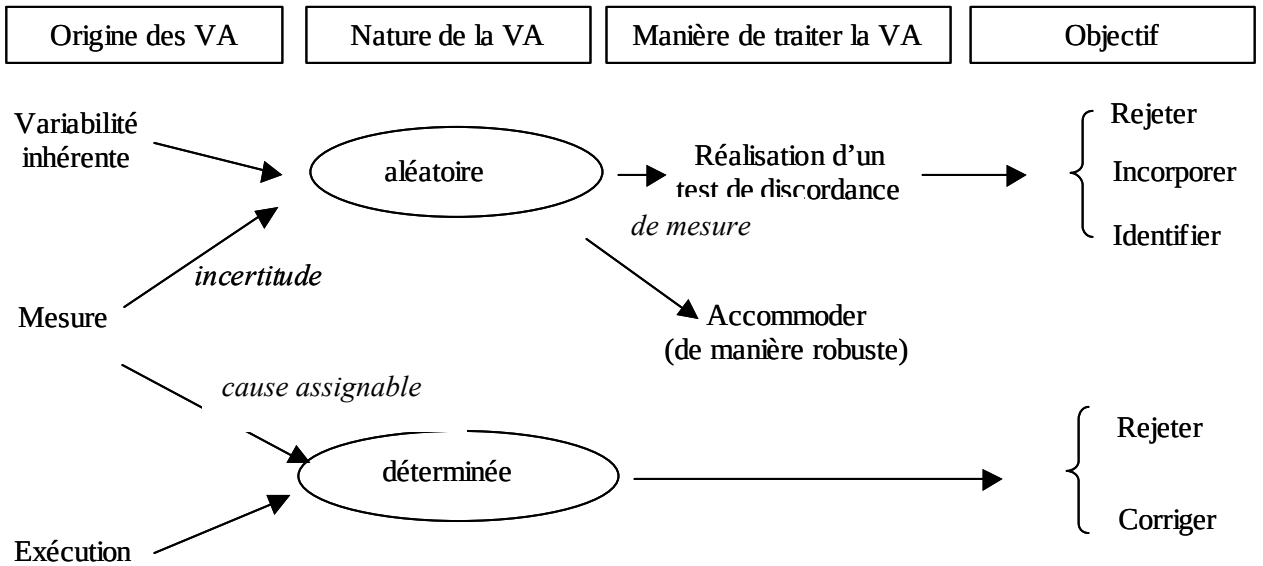


Figure I.3. Schéma général de traitement des valeurs aberrantes (VA) et objectifs poursuivis lors de l'examen des valeurs aberrantes (Barnett et Lewis, 1994).

1.2.5. Valeurs aberrantes en relation avec les modèles de probabilité

Dans les échantillons univariés, les observations susceptibles d'être déclarées comme valeurs aberrantes sont identifiées de manière évidente puisqu'il s'agit toujours des valeurs extrêmes de l'échantillon. Néanmoins, des valeurs anormales pour la distribution normale ne le sont pas nécessairement pour une distribution dissymétrique. Par définition, les valeurs aberrantes sont inconsistantes avec le reste des données en relation avec un modèle supposé connu. Ainsi, la distribution des données est une notion primordiale lors de l'application de méthodes statistiques car le traitement des valeurs aberrantes est directement lié au choix de cette distribution.

Notons que lorsqu'on émet l'hypothèse de l'existence d'une seule population, la valeur anormale correspond à une réelle valeur aberrante liée au choix du modèle de base. Le problème se complique lorsque plusieurs populations sont mélangées, par exemple dans le cas de F et G de la figure I.1.

Barnett et Lewis (1994) ont développé des tests d'hypothèses sur les problèmes de contamination des distributions. Néanmoins, ces tests restent très théoriques et sont liés à la connaissance *a priori* des proportions des distributions en mélange.

L'exemple présenté dans la suite de ce travail est également issu de la base de données de *RéQuaSud* et concerne une commune de la région condruzienne. Afin de vérifier si la distribution des données suit une distribution normale, le graphique des quantiles normaux a été réalisé (figure I.4). Lorsqu'une relation linéaire est obtenue à partir de ce graphique, les données suivent une distribution normale (Dagnelie, 1998a). Notons que d'autres possibilités pour tester la normalité des données sont proposées par Thode (2002).

Cette figure indique clairement la non-normalité de la distribution des valeurs de calcium étant donné qu'il n'existe pas de relation linéaire. Le traitement des valeurs aberrantes de manière classique selon l'hypothèse d'une distribution normale entraînerait clairement le rejet pur et simple de nombreuses valeurs supérieures à 700 mg/100 g T.S car celles-ci sont élevées par rapport à l'ensemble des données. Une valeur se démarque néanmoins très clairement, c'est la valeur 3158,5 mg/100 g T.S. qui, de toute évidence, présente les caractéristiques d'une valeur aberrante parce qu'elle est « étonnamment élevée » par rapport aux 1341 autres observations.

Dans le cas de cet exemple, il est nécessaire de se tourner vers des distributions dissymétriques (exponentielles, log-normale, de Weibull, de Pareto) pour vérifier si les valeurs supérieures à 700 mg/100 g T.S. sont bien aberrantes.

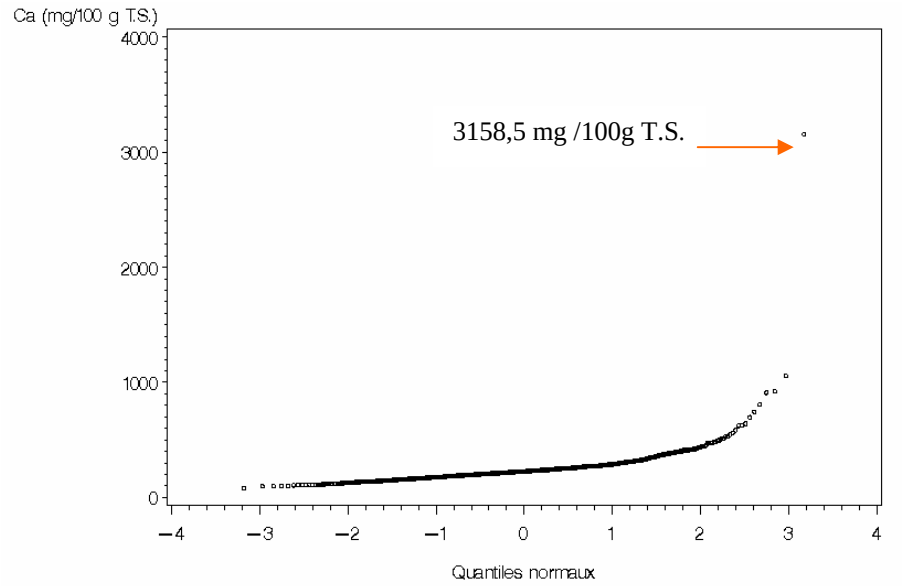


Figure I.4. Graphique des quantiles normaux pour les valeurs de calcium (mg/100 g T.S.) d'une commune de la zone agricole du Condroz (n=1342).

Comme pour la distribution normale, des graphiques des quantiles peuvent être réalisés afin de vérifier l'adéquation des observations à ces distributions. La manière de calculer les quantiles et de réaliser ce type de graphique est détaillée au paragraphe 2.2.2 pour diverses distributions dissymétriques telle que la distribution de Pareto.

La figure I.5 présente le graphique des quantiles pour la distribution de Pareto pour ce même échantillon de sol (Planchon, 2005). En considérant la partie droite de ce graphique, une relation linéaire est observée et permettrait d'inclure l'ensemble des données exceptée la valeur la plus élevée. Celle-ci serait considérée comme aberrante selon la distribution de Pareto. En prenant en compte une distribution dissymétrique, telle que la distribution de Pareto, il a donc été possible d'inclure la majorité des observations, excepté la valeur la plus extrême correspond clairement à une valeur aberrante. La théorie relative à cette distribution est détaillée au paragraphe 2.3.

Des techniques sophistiquées ont été développées par Beirlant *et al.* (1996) pour traiter des distributions dissymétriques et principalement en ne tenant compte que de la partie droite des distributions dissymétriques.

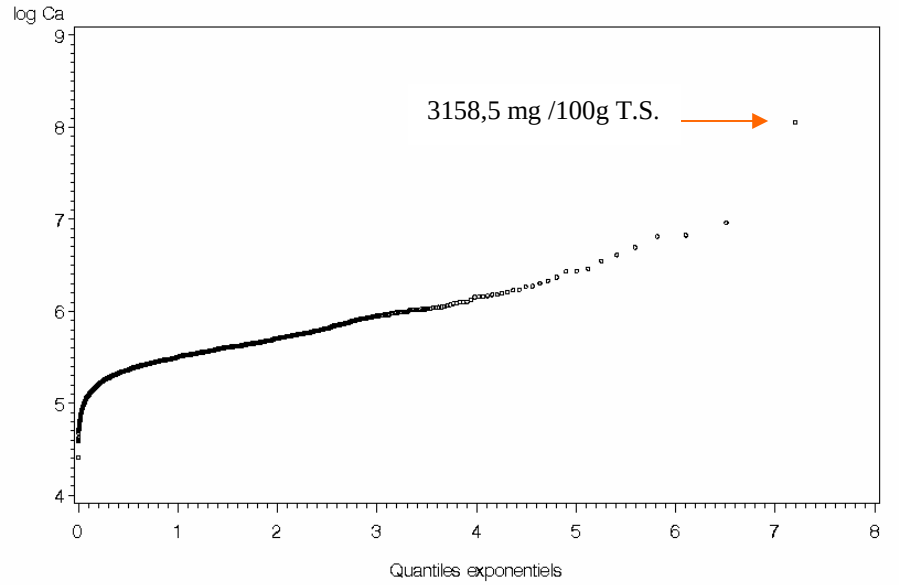


Figure I.5. Graphique des quantiles exponentiels pour le logarithme des valeurs de calcium (mg/100 g T.S.) d'une commune de la zone agricole du Condroz (n=1342) – *la linéarité de la relation pour les données situées dans la partie droite, indique que celles-ci suivent une distribution de Pareto, la dernière observation semble être une valeur aberrante selon cette distribution.*

1.3. Valeurs aberrantes dans le cas univarié

1.3.1. Introduction

Deux manières de traiter les valeurs aberrantes ont été citées au paragraphe précédent (figure I.3). Après quelques généralités (paragraphe 1.3.2.a), les tests de discordance sont présentés au paragraphe 1.3.2.b tandis que des généralités sur les procédures d'accommodation sont exposées au paragraphe 1.3.2.c.

1.3.2. Méthodes statistiques de traitement des valeurs aberrantes

a. Généralités

D'une manière générale, l'objectif d'une méthode statistique destinée à l'examen de valeurs aberrantes de nature aléatoire est de fournir des moyens pour vérifier si une déclaration *subjective* de la présence d'une valeur aberrante dans un ensemble de données possède des implications *objectives* importantes pour l'analyse future des données. Barnett et Lewis (1994) présentent deux catégories de méthodes statistiques distinctes. La première méthode (a) est basée sur l'idée de *tester* une valeur aberrante afin de déterminer si elle doit être gardée ou rejetée en utilisant un *test de discordance*. La deuxième méthode (b) est l'*accommodation* qui consiste en la manière de construire des procédures pour estimer les valeurs des paramètres de la distribution de base de façon relativement libre par rapport à toute influence néfaste d'une valeur aberrante. Cette dichotomie dans l'approche est fondamentale et se trouve à la base des améliorations récentes dans l'élaboration de méthodes statistiques. Cette deuxième approche, aussi intéressante soit-elle, ne correspond cependant pas à l'objectif poursuivi dans ce travail.

Notons que lors de l'utilisation d'un test statistique, une valeur aberrante peut simplement correspondre à une manifestation de l'erreur de type I, c'est-à-dire qu'il existe une certaine probabilité de mettre en évidence des valeurs aberrantes qui ne le sont pas. L'erreur de deuxième espèce consisterait par contre à ne pas déceler des valeurs aberrantes qui existent réellement. Une erreur de troisième espèce est liée à un choix inadéquat du modèle de base. Cette dernière erreur est peu connue et est rarement citée de la sorte (Dagnelie, 2003).

b. Tests de discordance

L'objectif poursuivi par ce premier type de méthodes statistiques est de tester la valeur aberrante afin de la rejeter de l'ensemble des données ou de l'identifier comme étant une caractéristique d'un intérêt particulier. Le test de discordance correspond à une procédure de détection qui permet de décider si une valeur aberrante peut être considérée comme faisant partie de la population principale.

Comme précédemment, soit l'échantillon x_1, x_2, …, x_n dont les valeurs extrêmes sont $x_{(1)}$, $x_{(n)}$. L'une de ces valeurs, par exemple $x_{(n)}$, peut être déclarée aberrante si elle engendre un effet de surprise en fonction de ce qu'on attend de manière informelle du modèle de base F. Supposons que toutes les observations sont bien issues de la distribution F. Un test statistique ou test de discordance peut être réalisé pour examiner si $x_{(n)}$ doit être considérée comme significativement plus grand, c'est-à-dire statistiquement inacceptable, en fonction de la distribution de $X_{(n)}$ sous F. Lorsque le résultat du test indique que $x_{(n)}$ n'est pas acceptable de manière statistique, on peut dire que $x_{(n)}$ est une valeur aberrante supérieure discordante pour le niveau du test. De manière similaire, on peut démontrer des discordances pour les valeurs aberrantes inférieures $x_{(1)}$ ou pour une paire de valeurs aberrantes ($x_{(1)}$,$x_{(n)}$), etc.

Aidé par la notion de test de discordance, il est possible de se rendre compte des différences dans la manière de définir les termes de valeur aberrante et valeur suspecte par les divers auteurs. Une *valeur suspecte* correspond, selon Barnett et Lewis (1994), à une valeur douteuse qui n'est pas jugée comme aberrante suite à la réalisation d'un test de discordance tandis que le terme *valeur aberrante* correspond à une valeur étonnamment extrême qui est statistiquement discordante. La valeur suspecte correspond donc à une valeur moins extrême qu'une valeur aberrante. Ce terme de valeur suspecte est exploité dans la norme ISO concernant l'utilisation de tests de détection de valeurs aberrantes (test de Cochran ou test de Grubbs) lors d'application de méthodes statistiques pour la maîtrise de la qualité et spécifiquement pour la détermination de la répétabilité et la reproductibilité d'une méthode de mesure (Anonyme, 1995). Si la statistique du test est supérieure à sa valeur critique au seuil de 1%, l'observation est une valeur aberrante tandis que si elle est supérieure à sa valeur critique au seuil de 5% et inférieure ou égale à sa valeur critique au seuil de 1%, l'observation est considérée comme suspecte.

Parmi les tests de discordance, une distinction peut être réalisée en fonction du type de distribution de la population-parent dont provient l'échantillon analysé. On peut distinguer les tests selon qu'ils sont appliqués dans le cas d'une population normale ou d'une autre distribution.

Barnett et Lewis (1994) classent les tests de discordance en sept types différents en tenant compte du critère retenu pour effectuer les tests. Certains tests ont des hypothèses très restrictives telles que la connaissance *a priori* du nombre de valeurs anormales ou la position relative de celles-ci (valeur inférieure ou supérieure). Les sept types de tests sont exposés ci-dessous.

1. Les statistiques liées au rapport *excès/étalement*. Ces statistiques correspondent au rapport des différences entre la valeur aberrante et la valeur de l'observation la plus proche, ou tout autre mesure d'étalement de l'échantillon, par rapport à une valeur de dispersion. Tel est le cas par exemple pour le test suivant où le modèle est supposé normal et s correspond à l'écart-type de l'échantillon, celui-ci peut être remplacé par toute mesure de la dispersion de l'échantillon :

$$\frac{x_{(n)}-x_{(n-1)}}{s}.$$

   Le test classique de Dixon entre dans cette catégorie où le dénominateur ne correspond non pas à s mais à $x_{(n)}-x_{(1)}$.

2. Les statistiques liées au rapport *amplitude/étalement* pour lesquelles, le numérateur est remplacé par l'amplitude de l'échantillon :

$$\frac{x_{(n)}-x_{(1)}}{s}.$$

3. Les statistiques liées au rapport *écart/étalement,* pour lequel le numérateur correspond à une mesure de distance entre une valeur aberrante et une mesure descriptive des données telle que la moyenne. En supposant que les données sont distribuées selon une loi normale, le test classique de Grubbs (1950) et utilisé par Carletti (1988) correspond à ce type de test et est destiné à tester soit une valeur aberrante inférieure $x_{(1)}$, soit une valeur supérieure $x_{(n)}$, soit les deux simultanément :

$$\frac{\bar{x}-x_{(1)}}{s}, \frac{\bar{x}-x_{(n)}}{s}.$$

   La moyenne $\bar{x}$ peut être remplacée par toute autre mesure de position. Des modifications peuvent être apportées et par exemple, une variante de ce test prend en compte le rapport de l'écart maximum au niveau du numérateur :

$$\frac{\max|x_i-\bar{x}|}{s}.$$

4. Les statistiques liées au rapport extrêmes/position qui correspondent au rapport de valeurs extrêmes par rapport à des mesures de position. Un exemple de ce type est le suivant :

$$\frac{x_{(n)}}{\bar{x}}.$$

5. Les statistiques liées au rapport de sommes de carrés qui sont légèrement différentes des statistiques précédentes et qui expriment le rapport entre les sommes des carrés pour l'échantillon dont on a extrait la valeur aberrante et l'échantillon global. Un test de ce type est proposé par Grubbs (1950) pour tester deux valeurs aberrantes supérieures. Ce test est également utilisé pour tester les valeurs aberrantes supérieures issues d'échantillons de valeurs extrêmes (Fung et Paul, 1985).
6. Les statistiques liées aux moments d'ordre supérieurs correspondent à des rapports de mesures de symétrie et d'aplatissement. Les tests développés par Ferguson (1961) utilisent ces statistiques.
7. Les statistiques *W* de Shapiro-Wilks sont également très utilisées pour tester des valeurs aberrantes et sont calculées de la manière suivante, par exemple pour une valeur aberrante inférieure $x_{(1)}$:

$$W = \frac{n}{n-1} \frac{(\bar{x} - x_{(1)})^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2}.$$

Ces tests sont principalement utilisés pour tester des observations issues de la distribution normale et de la distribution exponentielle.

Carletti (1988) ainsi que Barnett et Lewis (1994) présentent une liste de tests de détection de valeurs aberrantes pour un échantillon issu d'une distribution normale en fonction du nombre de valeurs aberrantes à détecter. Les principaux auteurs de ces tests sont les suivants : Grubbs (1950, 1969), Dixon (1950), David *et al.* (1954), Murphy (1951), Tietjen et Moore (1972), Ferguson (1961), Shapiro *et al.* (1968), Royston (1982).

De même, pour les échantillons extraits d'une population non normale, ces mêmes auteurs présentent une liste de tests de détection de valeurs aberrantes. Les distributions concernées sont les suivantes : exponentielle, gamma, uniforme, Poisson et binomiale. Les principaux auteurs de ces tests sont : Likes (1966), Kabe (1970), Shapiro et Wilk (1972), Kimber (1988), Lewis et Fieller (1979), Chikkagoudar et Kunchur (1987). Une présentation exhaustive de tous les tests de discordance n'est pas l'objectif de ce travail.

Barnett et Lewis (1994) présentent également des tests de discordance spécifiques pour des distributions de Gumbel, Fréchet, Weibull, Pareto, Poisson et binomiale.

Il faut signaler que ces tests sont sujets à ce qu'on appelle l'*effet de masque*[16], qui consiste en l'incapacité d'une procédure statistique d'identifier une valeur aberrante en présence de plusieurs valeurs suspectes. Carletti (1988) signale que les tests de Dixon sont plus sensibles à cet effet de masque que les tests de Grubbs. Un autre danger concerne le *phénomène de débordement*[17] lié aux tests qui s'occupent de deux ou plusieurs valeurs aberrantes. Dans une telle procédure, appelée *procédure en blocs*, plusieurs valeurs aberrantes sont testées en bloc en une seule opération, les deux valeurs sont considérées comme aberrantes alors que l'une des deux ne l'est pas. Une approche différente consiste en l'examen des valeurs aberrantes de manière séquentielle en utilisant des tests sous forme hiérarchique. De tels tests ont été développés et sont présentés par Barnett et Lewis (1994).

Afin d'éviter l'effet de masque et les contraintes diverses liées à la majorité des tests de discordance (nombre de valeurs aberrantes à connaître *a priori*, position des valeurs aberrantes, normalité de la distribution), une variante du test de Grubbs classique est appliqué à un échantillon tronqué symétriquement afin d'obtenir une estimation plus robuste de la moyenne et de l'écart-type (Carletti, 1988). La troncature doit être suffisamment importante pour éliminer les données aberrantes mais de manière pas trop prononcée afin de conserver une bonne estimation des paramètres. Un bon compromis proposé semble correspondre à une troncature de 8% symétrique (4% à gauche, 4% à droite). Le test de Grubbs de détection de valeurs anormales réalisé sur un échantillon ordonné et tronqué symétriquement (4% de l'effectif de chaque côté) est plus intéressant en terme de puissance que le test classique de Grubbs.

Dans le cas de non-normalité des distributions, la troncature n'est pas toujours suffisante pour se rapprocher de l'hypothèse de normalité exigée par la méthode de Grubbs, une autre variante du test de Grubbs sur des échantillons tronqués et transformés par une fonction puissance (Box et Cox, 1964) est proposée par Carletti (1988). Pour ce test de Grubbs avec troncature et transformation, les résultats sont nettement moins puissants que le test sur l'échantillon tronqué non transformé.

Les transformations restent néanmoins utiles pour des distributions très éloignées de la distribution normale car, dans le cas de distributions non normales, le risque de mettre en évidence des valeurs aberrantes, alors qu'il n'y en a pas, est plus élevé pour un échantillon tronqué non transformé que lorsqu'il y a transformation. Ces transformations, de type Box et Cox, sont donc à appliquer lorsque la distribution est fortement dissymétrique et qu'elle n'est pas connue *a priori*. Par contre, lorsque la distribution est connue, une méthode de détection spécifique est à appliquer.

[16] En anglais : *masking effect*.
[17] En anglais : *swamping effect.*

Sachant que les distributions des données auxquelles nous avons à faire sont particulièrement dissymétriques, il ne nous semble pas intéressant de symétriser les distributions. Le risque de perdre beaucoup d'informations sur les données initiales n'est certainement pas négligeable. Dans notre cas, nous nous trouvons face à des distributions très dissymétriques que nous désirons identifier pour appliquer éventuellement des tests de discordance spécifiques à celles-ci.

c. Accommodation des valeurs aberrantes

Les procédures d'accommodation englobent des méthodes statistiques destinées à réaliser de l'inférence sur la population à partir de laquelle l'échantillon aléatoire a été obtenu. Les résultats acquis par l'intermédiaire de ces procédures ne sont pas sérieusement déformés par la présence des valeurs aberrantes ou par des contaminants. Lorsqu'on suspecte la présence de valeurs aberrantes suite à des erreurs d'exécution ou des mesures aléatoires et que l'objectif de l'étude correspond à l'estimation d'un paramètre du modèle initial, il est intéressant d'utiliser un estimateur qui n'est pas trop sensible à la présence de celles-ci. L'utilisation de la médiane de l'échantillon comme estimateur de position en est un exemple très simple. Rousseeuw et Bassett (1990) font également appel à la notion de *remédiane*. Pour calculer la remédiane, de base *b*, des médianes sont calculées à partir de *b* groupes d'observations, ensuite la médiane de ces médianes est recalculée et constitue la remédiane. Zhang et Zhang (1996) proposent le calcul d'une moyenne symétrique robuste en faisant appel à des troncatures et à la transformation de Box et Cox dans le cadre de bases de données environnementales. Notons que cette procédure est très semblable à celle proposée par Carletti (1988).

Les procédures d'accommodation permettent dès lors d'éviter de rejeter des valeurs aberrantes. Cette manière de travailler implique que les valeurs aberrantes en elles-mêmes ne sont plus le centre d'intérêt de l'étude, le but consiste alors à travailler correctement malgré leur présence. Ceci correspond exactement au concept de robustesse. Les techniques d'accommodation sont dites *robustes* face à la présence de valeurs aberrantes, cependant, le concept de *robustesse*, de grande importance dans le cadre général de l'inférence statistique, n'est pas spécifique à l'examen des valeurs aberrantes.

De nombreux travaux dans ce domaine ont commencé avec Glaisher (1872) et ont été poursuivis par divers auteurs, cités par Barnett et Lewis (1994). Ces travaux ont comme objectif de réduire le poids attaché aux valeurs extrêmes lors de l'estimation. Au cours des dernières décennies, des efforts considérables ont été réalisés pour obtenir des procédures statistiques qui fournissent une certaine protection contre divers types d'incertitude sur le mécanisme de génération des données. Ces procédures incluent les méthodes d'estimation ou de tests sur des statistiques descriptives calculées

à partir de la distribution sous-jacente. Elles comprennent également d'autres procédures plus générales d'inférence pour lesquelles les tests d'estimation retiennent les propriétés statistiques de tout un ensemble de distributions possibles (Rousseeuw et Leroy, 1987; Hampel *et al.*, 1986). Huber (1972; 1981) a proposé trois types d'estimateurs robustes appelés L-estimateurs, R-estimateurs et M-estimateurs.

Les méthodes robustes peuvent également répondre spécifiquement au problème de valeurs aberrantes lorsqu'il y a une contamination et dès lors un décalage par rapport à un modèle de probabilité initial. Il ne faut cependant pas négliger l'importance du modèle de base dans le cas de l'accommodation. Si des valeurs aberrantes sont détectées parce que le modèle initial ne reflète pas le degré approprié de variabilité, il est nécessaire de s'intéresser à des distributions plus étendues que la distribution normale, utilisée classiquement. L'omission de valeurs extrêmes pour se protéger contre les valeurs aberrantes est une manière robuste pour estimer des mesures de dispersion mais si le modèle de base n'est pas correctement choisi, la procédure encourage plutôt la sous-estimation, le but étant de réduire l'effet des valeurs extrêmes. Si d'un autre côté, une hypothèse alternative permet d'exprimer la contamination du modèle initial, l'estimation ou le test des paramètres du modèle initial peut être très intéressant et il est alors important d'utiliser des procédures robustes appropriées pour se protéger des composants de faible probabilité ou contre les valeurs décalées.

Les travaux récents qui, implicitement ou explicitement, tentent d'accommoder les valeurs aberrantes dans le processus d'inférence se divisent en deux tendances. La première tendance comprend les méthodes d'estimation qui protègent implicitement contre les valeurs aberrantes en plaçant moins d'importance sur les valeurs extrêmes que sur les autres observations de l'échantillon. Cet accent est une caractéristique de l'ensemble des méthodes robustes développées durant les 30 dernières années. La seconde tendance de l'étude sur la contamination par des valeurs aberrantes est spécifiquement liée à la robustesse lors de la présence de ces valeurs. Les méthodes d'estimations et les tests qui en découlent, portent un regard particulier sur la nature des modèles nécessaires à expliquer la présence des valeurs aberrantes. Ce domaine d'étude est en cours d'expansion et des techniques d'accommodations spécifiques sont développées actuellement.

1.4. Conclusions

En raison du développement rapide des moyens de collecte des données et du traitement informatique de l'information, le problème de la présence de valeurs aberrantes a pris une importance non négligeable durant les dernières décennies. Les observations contenues dans les bases de données doivent nécessairement faire l'objet d'une validation car l'apparition de valeurs aberrantes est inévitable en raison de la quantité des données traitées et des diverses sources d'erreurs lors de leur acquisition. Pour assurer des informations de haute qualité, une recherche de valeurs suspectes ou aberrantes doit être effectuée avant l'exploitation des bases de données. La présence de valeurs aberrantes peut conduire à des estimations biaisées de paramètres et, suite à la réalisation de tests statistiques, à une interprétation des résultats qui peut être très altérée. Il est donc naturel de rechercher les moyens d'interpréter ou de caractériser ces valeurs anormales et de mettre au point des méthodes pour les traiter, soit en les rejetant afin de restaurer les propriétés initiales des ensembles de données, soit en adoptant des méthodes qui diminuent leur impact au cours des analyses statistiques.

Les termes liés aux valeurs aberrantes ont été abondamment définis et ont montré l'importance de l'hypothèse d'un modèle de base pour les données traitées. La nature aléatoire ou déterministe d'une observation aberrante et les sources d'apparition de celles-ci ont été exposées en relation directe avec les différents objectifs possibles rencontrés lors de l'examen de valeurs aberrantes. En fonction de l'objectif à atteindre et de la nature de la valeur aberrante, le traitement des données est très différent. Les deux possibilités pour traiter les données anormales, dans le cadre d'un objectif donné, sont les tests de discordance et les procédures d'accommodation.

L'objectif que nous avons retenu dans le cadre de notre étude est le rejet des valeurs aberrantes et la manière de les détecter correspondrait dès lors à l'utilisation de tests de discordance dans l'hypothèse d'une distribution donnée, quelle que soit la nature de données. En effet, il nous est impossible de vérifier *a posteriori* s'il existe des erreurs d'encodage, d'enregistrement, etc. Cependant, lorsqu'il faut vérifier plusieurs valeurs qui pourraient être aberrantes et dont le nombre à tester n'est pas forcément connu *a priori*, il est nécessaire d'appliquer de multiples tests de discordance les uns à la suite des autres. En effet, comme nous l'avons présenté, les tests de discordance ne concernent en général qu'une seule valeur, parfois deux, trois, rarement plus. Lorsque plusieurs tests statistiques sont réalisés simultanément, il est bien connu que le risque global d'erreur de première espèce s'accroît, ce qui pose également le problème de la puissance des tests réalisés. De plus, parmi les tests proposés dans la littérature, il n'est pas possible de faire le choix d'un test *a priori* car il n'existe pas de base logique de comparaison. Il serait donc nécessaire de comparer les différents tests entre eux, ce qui n'est pas l'objectif principal de ce travail.

Pour cette étude, nous préférons donc utiliser une méthode plus simple qui consiste à comparer les valeurs à tester par rapport à des quantiles extrêmes, soit par exemple pour la partie droite, le quantile 0,999 et pour la partie gauche, le quantile 0,001. Ces quantiles sont déterminés à partir des distributions dont le modèle est choisi judicieusement, c'est-à-dire des distributions dissymétriques. Avec le problème de mélanges de distributions exposé précédemment, les paramètres des distributions sont estimés à partir des queues des distributions situées à droite ou à gauche.

Les observations à introduire ultérieurement dans la base de données doivent donc faire l'objet d'une comparaison par rapport aux quantiles extrêmes estimés à partir de paramètres spécifiques à un modèle de probabilité, à déterminer pour chacun des éléments étudiés. Les valeurs qui sont supérieures à ces limites seront alors considérées comme aberrantes.

Il est donc indispensable d'avoir une meilleure idée de la forme de la distribution des données issues d'échantillons de sols pour des éléments classiques tels que le pHKCl, le carbone, le potassium, le magnésium, le calcium. Une étude détaillée des distributions dissymétriques est nécessaire pour déterminer les paramètres à utiliser pour estimer les quantiles extrêmes et donc fixer les limites de rejet des valeurs aberrantes. L'étude des distributions dissymétriques est présentée au chapitre suivant.

L'étude des données multivariées n'est pas appliquée dans notre cas étant donné la dissymétrie des distributions des données et la nécessité de développer une technique de détection des valeurs aberrantes rapide et opérationnelle. Cependant, il pourrait être envisagé dans les perspectives de ce travail d'adapter les techniques multivariées au cas des distributions très dissymétriques.

Les données de type géochimiques présentent en plus des caractéristiques propres dont il faut tenir compte, étant donné l'existence des relations entre les éléments étudiés et une relative homogénéité spatiale entre des échantillons de sols issus de zones pédologiques similaires. Les considérations particulières à prendre en compte en fonction du caractère spatial des données sont développées au chapitre 3.

2. THEORIE SUR LES DISTRIBUTIONS A FORTE DISSYMETRIE

2.1. Introduction

Comme cité au début de ce travail, l'étude bibliographique des démarches classiques de détection de valeurs aberrantes a montré l'importance de l'hypothèse de normalité des distributions des populations-parents. Cette hypothèse n'est pas vérifiée pour des données de type géochimique obtenues lors de l'étude d'éléments d'échantillons de sols. Le nombre important de valeurs très élevées nous conduit vers l'étude des distributions dissymétriques et aux distributions qui les généralisent.

L'objectif de ce chapitre est d'identifier les distributions qui correspondent le mieux aux données observées. Ces distributions doivent permettre de répondre à des situations de mélanges, en s'attachant aux queues des distributions, tout en restant facilement applicables à la détection des valeurs aberrantes.

Des généralités sur les distributions à forte dissymétrie sont présentées au paragraphe 2.2. Une étude détaillée des distributions dissymétriques est ensuite réalisée au paragraphe 2.3 avec la distribution exponentielle (paragraphe 2.3.4), la distribution de Weibull (paragraphe 2.3.5), la distribution log-normale (paragraphe 2.3.6) et enfin la distribution de Pareto (paragraphe 2.3.7). La distribution de Burr est citée au paragraphe 2.3.8. Enfin, des conclusions concernant l'ensemble des distributions et sur la manière de traiter les valeurs aberrantes dans le cas de distributions dissymétriques clôturent ce chapitre (paragraphe 2.4).

2.2. Généralités sur les distributions à forte dissymétrie

2.2.1. Introduction

Les distributions à forte dissymétrie ont, entre autres, été étudiées à partir de l'examen des valeurs extrêmes qui consiste en la recherche et la caractérisation des distributions limites vers lesquelles les valeurs extrêmes convergent. Ces études, appelées de manière générale sous la dénomination de *théorie des valeurs extrêmes*[18] ont été développées pour l'estimation d'événements rares. Elle permet d'extrapoler le comportement de la queue de la distribution des données à partir des observations les plus élevées (Garrido, 2002), c'est-à-dire de calculer la probabilité qu'un événement se produise même si celui-ci n'a jamais eu lieu. Signalons qu'elle a été développée principalement pour traiter des valeurs extrêmes situées à droite de la distribution dissymétrique, soit les maxima et qu'elle a amené les auteurs à utiliser plus spécifiquement des fonctions dérivées de la fonction

[18] En anglais: *Extreme Value Theory*, acronyme : EVT.

de répartition d'une distribution ou de la fonction des quantiles (paragraphe 2.2.2).

L'étude des valeurs extrêmes a été initiée dans les années 1920 pour répondre à des problèmes de valeurs aberrantes rencontrés dans le domaine de l'astronautique (Johnson et Kotz, 1970). Les développements mathématiques ont débuté par la caractérisation de la distribution des valeurs extrêmes d'un échantillon univarié indépendamment et identiquement distribué (Dodd, 1923). Les résultats concernant le comportement asymptotique ont été présentés dans Fréchet (1927), Fisher et Tippett (1928), Gumbel (1935), Gnedenko (1943) et Gumbel (1958) ; ce dernier étant l'ouvrage de base pour la théorie des valeurs extrêmes. A partir des années 1930, de nombreuses applications pratiques ont vu le jour dans des domaines très diversifiés tels que les inondations, les crues des rivières, les tempêtes, les tremblements de terre, les données météorologiques, le temps de survie de micro-organismes ou le temps de survie face aux émissions radioactives (Johnson et Kotz, 1970 ; Kotz et Johnson, 1982). Dans toutes ces études, il est habituel d'ajuster les distributions des valeurs extrêmes et plus particulièrement la *distribution de Gumbel* aux maxima annuels. Cette distribution a également été utilisée pour étudier les phénomènes de corrosion ou le point de rupture de matériaux dans une optique de sécurité lors de la construction de bâtiments. Dans le domaine de la géologie et de la prospection minière, les distributions telles que les *distributions log-normales* et *log-hyperboliques* ont fait l'objet de nombreux travaux (Barndorff-Nielsen, 1977; Sichel, 1973). La notion de mélange de distributions log-normales a été introduite par ces auteurs avec des interprétations géologiques permettant de justifier l'adéquation de ces distributions. Caers *et al.* (1996) ont utilisé les *distributions de type Pareto* pour étudier la distribution de la taille de diamants et ont comparé les résultats obtenus à ceux de Sichel (1973). A l'origine, ces distributions de type Pareto ont été très développées dans le domaine économique et hydrologique et le sont actuellement dans le domaine des assurances. Leur utilisation permet d'alléger les traitements mathématiques rencontrés pour les autres distributions et de résoudre des problèmes pratiques.

Les distributions possédant une dissymétrie avec un étalement vers la droite font l'objet d'une généralisation par une distribution appelée *distribution généralisée des valeurs extrêmes* (paragraphe 2.2.3.a). Ces distributions peuvent être classées en trois classes de distributions en fonction de la valeur d'un paramètre (valeur nulle, négative ou positive), appelé *index des valeurs extrêmes* (paragraphe 2.2.3.b) (Hill, 1975; Pickands, 1975).

A l'heure actuelle, avec les progrès acquis dans le domaine informatique et grâce au nombre de données disponibles, la majorité des études dans le domaine des valeurs extrêmes se tournent vers une démarche qui consiste

plutôt à s'intéresser aux valeurs situées au-dessus de *valeurs seuil*[19] élevées plutôt que de considérer des valeurs maximales définies pour des périodes de temps fixées artificiellement (Caers *et al.*, 1996). De cette approche sont apparues les *distributions généralisées de Pareto* (paragraphe 2.2.4), étroitement liées aux distributions des valeurs extrêmes par l'intermédiaire de l'index des valeurs extrêmes.

Un inventaire des distributions dissymétriques a été proposé par Beirlant *et al.* (1996). Pour chacune des trois classes de distributions, les distributions sont classées en fonction de l'allure de la queue de la distribution, soit des queues de distributions les moins étendues aux queues de distributions les plus étalées (paragraphe 2.2.3.b). Parmi l'ensemble des distributions proposées par Beirlant *et al.* (1996), nous examinons, au paragraphe 2.3, certaines distributions conformément à l'ordre du classement établi par celui-ci, excepté pour la distribution exponentielle. Celle-ci sera présentée en premier (paragraphe 2.3.4) car elle est très connue et elle permet d'exposer plus facilement les aspects théoriques exposés pour les autres distributions ; la distribution exponentielle est plus dissymétrique que la distribution de Weibull et est présentée normalement après la distribution de Weibull. Notons que toutes les distributions proposées par Beirlant *et al.* (1996) ne sont pas présentées dans ce travail car peu d'applications pratiques en font mention dans la littérature.

2.2.2. Principales fonctions utilisées pour l'étude des distributions à forte dissymétrie

Essenwanger (1986) signale que lors de l'étude d'événements rares, il est intéressant de présenter le *risque d'occurrence* ou la *probabilité d'occurrence* de ces événements. C'est pour cette raison que, lors de la présentation des différentes distributions, nous exposons la manière de calculer ce risque. Celui-ci correspond à la probabilité que l'événement se produise et s'exprime de la manière suivante :

$$\text{probabilité d'occurrence} = 1 - F(x)\,,$$

$F(x)$ étant la fonction de répartition[20]. Notons que, dans le domaine biomédical, cette probabilité d'occurrence est appelée *fonction de survie*.

Une notion très fréquemment employée pour l'estimation de certains paramètres correspond à la *fonction des quantiles Q(p)*, qui fournit la plus petite valeur de x pour laquelle $F(x) \geq p$ avec $0<p<1$. Dans le cas d'une distribution théorique quelconque, elle correspond à l'inverse de la fonction de répartition et s'énonce de la manière suivante :

$$Q(p) = inv(x : F(x) \geq p)\,.$$

19 En anglais : *peak over threshold* – acronyme : POT

20 $F(x)=P(X \leq x)$

Lorsque la distribution n'est pas connue, la *fonction des quantiles empiriques* $\hat{Q}_n(p)$ est la fonction qui, pour une valeur donnée p ($0<p<1$), fournit la plus petite valeur, vers la gauche, pour laquelle une proportion p des données est rencontrée. Cette fonction correspond à une bonne approximation de la fonction des quantiles correspondante *Q(p)*. En pratique, pour une série de *n* observations, triées selon un ordre croissant (indiqué par l'astérisque), les quantiles observés $\hat{Q}_n(p)$ correspondent aux valeurs observées x_i^*, c'est-à-dire :

$$\hat{Q}_n(p)=x_i^*$$

$$\text{où } x_1^* \le x_2^* \le \ldots \le x_i^* \le \ldots \le x_n^* \text{ et } \frac{i-1}{n} \le p \le \frac{i}{n}.$$

A titre d'illustration, pour un ensemble de 10 échantillons de sols, dont la teneur en magnésium a été déterminée et exprimée en mg/100 g de terre sèche, on a obtenu les valeurs suivantes, qui ont été triées dans l'ordre croissant :

i	1	2	3	4	5	6	7	8	9	10
x	15,2	15,3	15,3	15,4	15,7	15,9	16,1	16,3	16,7	17,0
p	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00

Les quantiles empiriques pour diverses valeurs de *p* correspondent à :

$\hat{Q}$ (0,50) = 15,7 mg/100 g T.S. ;

$\hat{Q}$ (0,75) = 16,1 mg/100 g T.S. ;

$\hat{Q}$ (0,99) = 16,7 mg/100 g T.S.

Il est possible de vérifier le bon ajustement de la distribution des valeurs observées à une distribution théorique quelconque, en confrontant graphiquement les quantiles empiriques $\hat{Q}_n(p)$ aux valeurs théoriques fournies par la fonction des quantiles *Q(p)*, relative à cette distribution. L'inspection de ce graphique est présentée dans la littérature comme étant une technique qui permet de vérifier si l'hypothèse d'une distribution est correcte.

Pour la réalisation de tels graphiques, appelés *graphique des quantiles* ou graphiques des *quantiles-quantiles*, abrégés en *QQPlots*, Beirlant *et al.* (1996) proposent de considérer les seuils suivants :

$$p = \frac{1}{n+1}, \frac{2}{n+1}, 5\ , \frac{n-1}{n+1}, \frac{n}{n+1},$$

de telle sorte que :

$$\hat{Q}_n(p)=\hat{Q}_n(\frac{i}{n+1})=x_i^*, \qquad i=1, 2, 5 , n.$$

Le choix de telles valeurs de *p* permet de tenir compte de la correction de continuité nécessaire pour comparer une fonction discontinue $\hat{Q}_n$ avec une fonction théorique continue *Q*.

A partir du graphique des quantiles, il est également possible d'estimer, de manière empirique, la valeur des paramètres de la distribution étudiée.

Dans le cas de la distribution normale, la normalité des variables peut être vérifiée par l'existence d'une relation linéaire entre les quantiles *Q(p)* de la distribution normale $\mathrm{N}(\mu,\sigma^2)$ et les quantiles correspondants de la distribution normale réduite $\Phi^{-1}(p)$, définis par :

$$Q(p)=\mu+\sigma\,\Phi^{-1}(p),$$

où *p* est défini par :

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\Phi^{-1}(p)}\exp(-u^2/2)\,du\,.$$

Si les données suivent une distribution normale, les points de coordonnées :

$$\left(\Phi^{-1}(p),\hat{Q}_n(p)\right)$$

présentent une relation linéaire ; les valeurs $(\Phi^{-1}(\frac{i}{n+1}),\ x_i^*)$ étant reportées sur le graphique. Sous la condition d'une distribution normale, une droite est obtenue. La pente de celle-ci correspond à une approximation de l'écart-type σ de la distribution normale tandis que l'ordonnée à l'origine correspond à la valeur de μ.

2.2.3. Distribution généralisée des valeurs extrêmes et classes de distributions des valeurs extrêmes

a. Distribution généralisée des valeurs extrêmes

L'étude des valeurs extrêmes a fait l'objet de nombreux travaux (Essenwanger, 1986 ; Barnett et Turkman, 1993 ; Beirlant *et al.*, 1996 ; Rothenbuehler, 2002; Zivot, 2002) afin de déterminer les fonctions qui décrivent le mieux les queues de distributions. Suite à ces études, les auteurs ont proposé de généraliser l'ensemble des distributions dissymétriques au travers de la *Distribution Généralisée des Valeurs Extrêmes*[21] $G_\gamma(x)$, qui se présente sous la forme suivante :

$$G_\gamma(x)=\begin{cases} exp\left(-\left(1+\gamma x\right)^{-1/\gamma}\right) & \text{si } 1+\gamma x>0 \\ & \text{et } \gamma\neq 0 \\ exp\left(-exp(-x)\right) & \text{si } \gamma=0. \end{cases}$$

La distribution généralisée des valeurs extrêmes correspond à la distribution limite pour les valeurs les plus grandes d'échantillons aléatoires, triées dans l'ordre croissant (Beirlant et *al.*, 2004). En parallèle à l'étude de la distribution normale, ce théorème est l'équivalent du théorème limite central dans le cas particulier des valeurs extrêmes.

Le paramètre γ de la distribution des valeurs extrêmes est un nombre réel et $1/\gamma$ est appelé l'*index des valeurs extrêmes*[22], dont nous avons déjà parlé dans l'introduction. Des méthodes basées sur le maximum de vraisemblance ont été conçues pour estimer ce paramètre à partir de séries de *blocs de maxima*. Par exemple, dans le cas des études sur les inondations, les tempêtes, etc. qui sont liées à des séries temporelles, chaque bloc correspond à une année et la valeur maximale rencontrée au cours de cette année est utilisée pour l'ajustement. Les blocs comprennent alors les valeurs de maxima dérivés d'une seule valeur maximale par période de temps. Cette méthode est robuste et simple pour estimer les paramètres de la distribution généralisée des valeurs extrêmes mais présente certains désavantages exposés au paragraphe 2.2.4.

b. Distributions des valeurs extrêmes

En fonction de la valeur du paramètre γ, trois classes de distributions, dénommées *Distributions des Valeurs Extrêmes*[23], font partie de la distribution généralisée des valeurs extrêmes. Elles sont présentées au tableau I.1, de même que les distributions dissymétriques qu'elles généralisent. Notons ici que l'ordre de présentation de celles-ci correspond, par classe de distributions, aux distributions présentant les extrémités peu

[21] En anglais : *Generalized Extreme Value Distribution* – acronyme : GEV.
[22] En anglais : *Extreme Index Value* – acronyme : EIV.
[23] En anglais: *Extreme Value Distribution* – acronyme : EVD.

étendues (en haut du tableau) vers les extrémités les plus dissymétriques vers la droite (en bas du tableau) (Beirlant *et al.*, 1996).

- Dans le cas où γ=0, $G_\gamma(x)$ correspond à la *classe des distributions de type Gumbel* (ou Fisher-Tippett type I), notée couramment dans la littérature par $\Lambda(x)$. Pour ces distributions, tous les moments peuvent être calculés. Cette distribution correspond à une distribution limite pour les distributions présentant, à droite, des queues de distributions faiblement étendues (Beirlant *et al.*, 1996). Cette distribution limite est couramment appelée dans la littérature distribution de Gumbel. Font partie de cette classe, les distributions de Weibull, exponentielle, gamma, logistique, log-normale. Toutes les distributions appartenant à cette classe présentent une queue de distribution qui décroît rapidement de manière exponentielle. Les plus utilisées en pratique sont la distribution exponentielle, de Weibull et la distribution log-normale. Nous n'avons relevé aucun cas pratique dans la littérature où la distribution logistique est utilisée pour le traitement des valeurs extrêmes.

De même, la distribution gamma est citée le plus couramment dans la littérature car elle permet de faire le lien entre les distributions normales, χ^2, exponentielle et Poisson. En effet, cette distribution gamma présente trois paramètres (Johnson et Kotz, 1970 ; Beirlant *et al.*, 1996) et en fonction de la valeur qui leur est attribuée, l'une ou l'autre distribution citée est rencontrée. En pratique, très peu d'applications faisant directement référence à la distribution gamma sont rencontrées.

- Dans le cas où γ>0, $G_\gamma(x)$ correspond à la *classe des distributions de type Pareto* ou *Fréchet-Pareto* (ou Fisher-Tippett type II), notée $\Phi_\alpha(x)$, dans laquelle $\alpha = \frac{1}{\gamma}$ et est appelé, non plus index des valeurs extrêmes, mais *index de Pareto*. Une augmentation de la valeur de cet index a comme conséquence d'accroître la queue de la distribution dissymétrique vers la droite, c'est-à-dire que sa dispersion est plus grande. En effet, pour des valeurs de γ>0,5, la variance est infinie et pour γ>1, la moyenne n'existe pas (Beirlant *et al.*, 1996). C'est pour cette raison que les distributions de type Pareto sont souvent utilisées pour modéliser des queues de distributions très dissymétriques à droite.

Lorsqu'on ne considère que les queues des distributions, on peut traiter les distributions de type Pareto de manière identique quelle que soit la distribution prise en compte. Ceci permet d'éluder de nombreux problèmes mathématiques très fastidieux liés à la complexité des fonctions de distributions. Ces distributions sont donc très intéressantes dans notre cas, car, comme nous l'avons exposé précédemment, nous désirons traiter uniquement les queues des distributions afin d'éviter le problème des

mélanges des distributions dissymétriques. Des compléments théoriques relatifs à ces distributions sont présentés au paragraphe 2.3.7.

- Dans le cas où $\gamma<0$, $G_\gamma(x)$ correspond à la classe des distributions de type Weibull[24] (ou Fisher-Tippett type III), notée $\Psi_\alpha(x)$, dans laquelle $\alpha=-1/\gamma$. La queue de la distribution présente une limite qui est finie, tel est le cas pour la distribution uniforme. Cette classe de distributions n'entre pas dans le cadre de ce travail étant donné que, nous nous trouvons face à des distributions dissymétriques présentant des limites non finies.

Pour les distributions de type Gumbel ou de type Fréchet-Pareto, Essenwanger (1986) présente un test d'aptitude qui permet de déterminer le domaine de convergence le plus approprié pour les valeurs extrêmes (Van Montfort, 1970). D'autres références bibliographiques intéressantes traitent de ce sujet : Van Montfort et Otten (1978), Otten et Van Montfort (1980), Fraga Alves (1999) et Kysely et Huth (2001). Ce test présente le désavantage d'inclure l'ensemble des données observées et ne permet pas de tenir compte uniquement des queues de distributions.

Barnett et Lewis (1994) ont étudié les problèmes de valeurs aberrantes pour ces trois types de distributions. Dans le cas de la distribution de type Gumbel, une transformation de variable permet d'exprimer la distribution sous la forme d'une exponentielle et d'appliquer les tests proposés dans le cas de distributions exponentielles. D'autres tests spécifiques sont également proposés en fonction de la connaissance des paramètres de position et d'échelle de la distribution ; ces tests permettent de vérifier une, deux ou trois données anormales. Un nombre plus élevé d'observations suspectes peut être testé mais la performance des tests n'a pas été vérifiée par les auteurs.

Pour les distributions de type Pareto ou de type Weibull, les tests spécifiquement proposés pour les distributions de type Gumbel sont utilisés après l'application d'une transformation logarithmique aux données observées.

---

[24] Signalons que la dénomination *classe de distribution de type Weibull* ne nous semble pas appropriée car elle crée une confusion avec la distribution de Weibull pour laquelle $\gamma=0$ qui se trouve dans la classe des distributions de type Gumbel. Cette dénomination rend hommage aux recherches effectuées par l'ingénieur Suédois ainsi nommé.

Tableau I.1. Distributions des valeurs extrêmes et distributions dissymétriques en fonction de l'index des valeurs extrêmes et du degré d'étalement de la queue de la distribution vers la droite. Par classe de distributions des valeurs extrêmes, l'ordre de présentation des distributions correspond, de haut en bas, aux distributions présentant des queues peu étendues vers les queues les plus étalées vers la droite (Beirlant *et al.*, 1996, 1998).

Distribution généralisée des valeurs extrêmes	γ	Distributions des valeurs extrêmes	Notation	Expression mathématique	Conditions d'application	Distributions dissymétriques
	$\gamma=0$	Distributions des valeurs extrêmes de type Gumbel	$\Lambda(x) =$	$\exp(-\exp(-x))$	si x est un réel	Weibull Exponentielle Gamma Logistique Log-normale
$G_{\gamma,\beta}(y)$	$\gamma>0$	Distributions des valeurs extrêmes de type Fréchet-Pareto	$\Phi_\alpha(x) =$	$\exp(-x^{-\alpha})$	si $x>0$ et $\alpha>0$	Pareto Burr Log-gamma Loghyperbolique Log-logistique Fréchet
	$\gamma<0$	Distributions des valeurs extrêmes de type Weibull	$\Psi_\alpha(x) =$	$\exp(-(-x)^{\alpha})$	si $x\le 0$ et $\alpha>0$	*non approprié dans le cas de notre étude*

2.2.4. Distributions généralisées de Pareto

Selon divers auteurs, la distribution généralisée des valeurs extrêmes présente l'avantage d'être facile à mettre en place et à valider (Johnson et Kotz, 1970 ; Beirlant *et al.*, 1996 ; Madsen, 2001). Elle est également robuste quand peu de paramètres sont à estimer. Cependant, la distribution généralisée des valeurs extrêmes présente certains désavantages. En effet, elle exclut un grand nombre de données qui pourraient permettre de mieux comprendre le comportement des queues de distributions. Par exemple, dans le cas de données météorologiques, il est impossible de savoir si un maxima annuel est une accumulation de succession de maxima (le maxima correspond à une journée très chaude au cours d'une période où la température est particulièrement élevée) ou si c'est un événement isolé (une seule journée très chaude lors d'une période donnée). De plus, elle ne permet pas de mettre en évidence les relations entre les variables qui présentent des valeurs extrêmes. En effet, il peut y avoir des dépendances entre l'apparition d'un événement extrême pour une variable donnée, et l'apparition d'une autre observation extrême pour une autre variable (Madsen, 2001 ; Goegebeur *et al.*, 2002 ; Holt, 2002 ; Zivot, 2002).

Ces inconvénients ont poussé au développement des théories relatives aux distributions généralisées de Pareto qui permettent de considérer les valeurs extrêmes situées au-dessus de *valeurs seuil* élevées. Ceci signifie que ces distributions sont ajustées à toutes les valeurs les plus élevées d'une série d'observations plutôt que de considérer des valeurs maximales définies, par exemple, pour des périodes données (Caers *et al.*, 1996). Le nombre de valeurs extrêmes varie ainsi en fonction de la valeur du seuil fixée. Ces développements théoriques sont devenus possibles grâce à une plus grande facilité de calculs suite aux développements informatiques.

A l'heure actuelle, la majorité des études dans le domaine des valeurs extrêmes sont basées sur cette démarche appelée *méthode d'estimation au-dessus d'une valeur seuil*[25] ou les méthodes *POT*[26]. Cette manière de considérer des données au-dessus de valeurs seuil c'est-à-dire de prendre les *k* plus grandes observations d'un ensemble de *n* données, conduit, selon Caers *et al.* (1996), à un développement plus rigoureux et plus efficace de méthodes statistiques destinées à étudier les valeurs extrêmes (Pickands, 1975 ; Smith ,1987; 1989; Madsen, 2001 ; Zivot, 2002).

[25] Dans la littérature, la distribution généralisée des valeurs extrêmes et la distribution généralisée de Pareto sont parfois nommées, respectivement, par les termes de *distributions des maxima* (distributions of maxima) et de *distributions des queues* (distributions of tails).

[26] En anglais : *Peak Over Threshold* – acronyme : POT.

Pour des valeurs suffisamment élevées d'une valeur seuil *u*, il existe un paramètre d'échelle β, dépendant de *u*, telle que la distribution des valeurs supérieures à *u* est ajustée correctement par les *Distributions Généralisées de Pareto*[27] qui se présentent sous la forme suivante (tableau I.2).

Tableau I.2. Distribution généralisée de Pareto : expressions mathématiques en fonction de la valeur du paramètre γ (Beirlant *et al.*, 1996, 2004).

Distribution généralisée de Pareto	γ	Expression mathématique	Conditions d'application
$G_{\gamma,\beta}(x) =$	$\gamma=0$	$1-exp(-\frac{x}{\beta})$	$x \geq 0$, $\beta>0$
	$\gamma>0$	$1-\left(1+\gamma\frac{x}{\beta}\right)^{-1/\gamma}$	$x \geq 0$, $\beta>0$
	$\gamma<0$	$1-\left(1+\gamma\frac{x}{\beta}\right)^{-1/\gamma}$	$0 \leq x \leq -\frac{\beta}{\gamma}$, $\beta>0$.

Si le seuil est très élevé, le modèle peut être tout à fait acceptable, cependant, des valeurs de seuil trop élevées vont prendre en compte peu de valeurs pour l'ajustement tandis que des valeurs trop faibles vont entraîner un mauvais ajustement de la distribution des valeurs supérieures à *u* et l'ajustement sera biaisé. Il est donc fondamental de trouver une valeur seuil qui soit optimale pour réaliser l'ajustement des paramètres des distributions généralisées de Pareto.

Il existe un lien très étroit entre les distributions des valeurs extrêmes et les distributions généralisées de Pareto (Zivot, 2002). En effet, lorsque le paramètre γ est nul, comme pour les distributions des valeurs extrêmes, la distribution généralisée de Pareto présente une queue de distribution peu étendue à droite et la queue de la distribution généralisée de Pareto est de type Gumbel (pour $x \to \infty$). Lorsqu'il est positif, la distribution est très étendue à droite avec un index équivalent à $1/\gamma$ et la queue de la distribution est de type Pareto (pour $x \to \infty$).

Les principaux avantages des distributions généralisées de Pareto sont, d'une part, l'utilisation plus efficace des données car toutes les valeurs supérieures au seuil fixé sont prises en considération plutôt que la valeur maximale du bloc considéré, et, d'autre part, grâce aux valeurs seuils, il est possible d'étudier les relations entre les valeurs extrêmes de différentes variables, ce qui n'était pas le cas avec les distributions des maxima.

D'après Madsen (2001), les distributions généralisées de Pareto sont les distributions les plus intéressantes pour modéliser les queues de distributions très étendues. Elles nous semblent donc très intéressantes pour

[27] En anglais *: Generalized Pareto Distributions* – acronyme : GPD.

l'étude des queues des distributions. Notons la complexité mathématique de celles-ci, ce qui nous conduit finalement à envisager l'utilisation de distributions dissymétriques classiques de type Gumbel et de type Pareto.

2.3. Présentation de distributions dissymétriques

2.3.1. Introduction

L'objectif de ce chapitre est de présenter les distributions dissymétriques les plus appropriées à l'ajustement des données issues d'analyses d'échantillons de sols. Il faut bien signaler que nous recherchons les distributions dissymétriques qui s'ajustent le mieux aux extrémités de la distribution observée et non pas à l'ensemble de la distribution. Le choix de ces distributions a été réalisé en relation avec leur dissymétrie plus ou moins étalée vers la droite ou vers la gauche, leur facilité de mise en pratique et leur utilisation dans la recherche de valeurs aberrantes. A titre informatif, les tests de discordance spécifiques à chacune d'elles sont cités ; ceci permet de montrer la difficulté d'appliquer ces tests dans notre situation.

Pour mettre en parallèle les différentes distributions étudiées, nous faisons appel aux graphiques des quantiles qui permettent de comparer les quantiles observés aux quantiles théoriques de la distribution envisagée, et ce à partir de jeux de données issus de la base de données de *RéQuaSud* (paragraphe 2.3.2). Ces graphiques mettent en évidence la discordance par rapport au modèle choisi et principalement au niveau de la queue de la distribution. Leur courbure représente en effet une mesure de la déviation de la queue de la distribution par rapport au comportement supposé de celle-ci.

Afin de faciliter la compréhension de l'exposé théorique, un rappel succinct des principales méthodes d'ajustements des paramètres est exposé au paragraphe 2.3.3. Les principes liés à la manière de vérifier la qualité de l'ajustement sont également présentés.

En ce qui concerne les distributions présentant un index des valeurs extrêmes nul, nous présentons : la distribution exponentielle (paragraphe 2.3.4), la distribution de Weibull (paragraphe 2.3.5), la distribution log-normale (paragraphe 2.3.6). Pour les distributions présentant un index des valeurs extrêmes positif, nous étudions respectivement la distribution stricte de Pareto (paragraphe 2.3.7) et, très brièvement, la distribution de Burr (paragraphe 2.3.8). Les liens existants entre les différentes distributions sont présentés au paragraphe 2.3.9.

2.3.2. Description de deux jeux de données

Chacune des distributions étudiée est illustrée à l'aide de deux jeux de données issus de la chaîne SOLS de *RéQuaSud*, à partir des variables qui présentent les distributions les plus dissymétriques : le magnésium et le calcium. Pour le premier jeu de données, nous avons décidé de retenir les valeurs de magnésium d'une commune de la base de données dont la

distribution apparaît particulièrement dissymétrique et susceptible de contenir des valeurs aberrantes. Les valeurs de calcium d'une deuxième commune font l'objet du second jeu de données.

Afin d'éluder le problème de mélanges de distributions, nous allons étudier, d'une part, la partie droite de la distribution en tronquant 70% des données à gauche et, d'autre part, la partie gauche, en tronquant, 70% des données à droite. Ceci permettra de montrer comment les distributions étudiées permettent de traiter les queues de distributions droite ou gauche. Notons que le choix de troncature de 70% des données est tout à fait arbitraire et fera l'objet d'une étude bien précise au cours de l'application (deuxième partie).

Il est très important de signaler qu'aucune généralisation ne peut être réalisée à partir de ces deux jeux de données. Ceux-ci servent uniquement à montrer de manière générale les caractéristiques des distributions. Une étude approfondie sera réalisée lors de l'application. Chacune des communes de la base de données de la chaîne SOLS de *RéQuaSud* fera l'objet d'une analyse relative au type de distribution suivi par les données.

a. Premier jeu de données relatif à des données de magnésium

Le premier jeu de données contient 853 observations de magnésium dont la moyenne et l'écart-type sont respectivement de 13,6 mg/100g T.S. et 4,9 mg/100g T.S. Les valeurs des percentiles les plus intéressants sont présentées au tableau I.3.

La démarche classique, qui permet de bien mettre en évidence la dissymétrie, est de présenter un histogramme de fréquence des observations (figure I.6) mais celui-ci ne permet pas d'identifier de manière claire les observations extrêmes qui nous intéressent. De même, la fonction cumulative de fréquence ne permet pas de mettre en évidence ces valeurs extrêmes. Une représentation graphique des observations, triées par ordre croissant, nous semble plus appropriée pour montrer les observations anormalement élevées (figure I.7).

Tableau I.3. Percentiles pour les teneurs en magnésium relatives au premier jeu de données sélectionné.

Percentiles	Teneur en magnésium (mg/100g T.S.)
Minimum	4,9
p5	7,3
p10	8,5
p50	12,5
p90	19,2
p95	23,0
p99	29,0
Maximum	57,5

A première vue, la figure I.7 indique que la teneur maximale en magnésium (57,5 mg/100g T.S.) est une observation très isolée par rapport aux autres observations. On observe également que les 5 valeurs comprises entre 30,0 et 50,0 mg/100g T.S. s'écartent de l'ensemble des observations et sont susceptibles d'être aberrantes. Notons que la technique appliquée jusqu'à présent au sein du système de gestion de la base de données[28] a affecté le code de rejet uniquement à la dernière observation. Le graphique des quantiles normaux, présenté à la figure I.8, présente une relation non linéaire, ce qui indique clairement la non-normalité de la distribution des observations.

Pour étudier les parties droite et gauche de ce jeu de données, seules 30% des données sont prises en compte ce qui correspond à 256 observations ; 597 données étant ignorées.

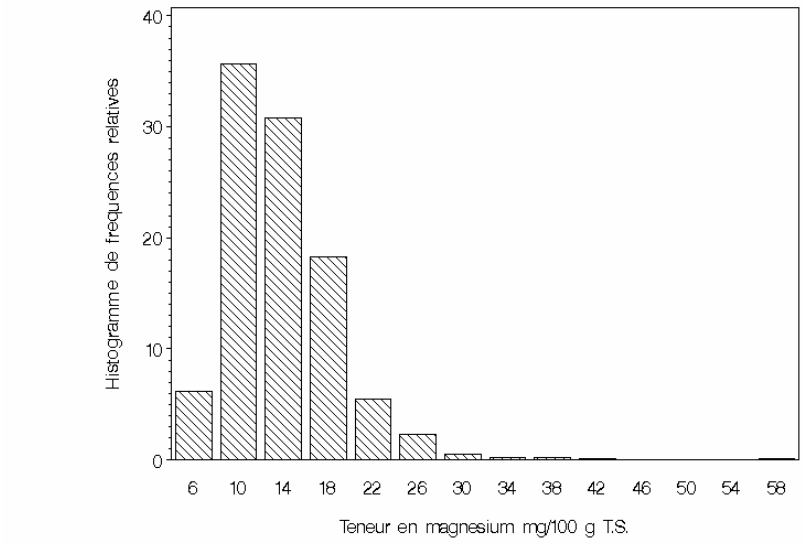


Figure I.6. Histogramme de fréquence relative de la teneur en magnésium (mg/100g TS) pour les observations du premier jeu de données sélectionné.

[28] Limites de rejet basées à partir d'un multiple de l'écart-type des sous-populations, cet écart-type est calculé par une méthode robuste basée sur les quantiles de la distribution, supposée normale.

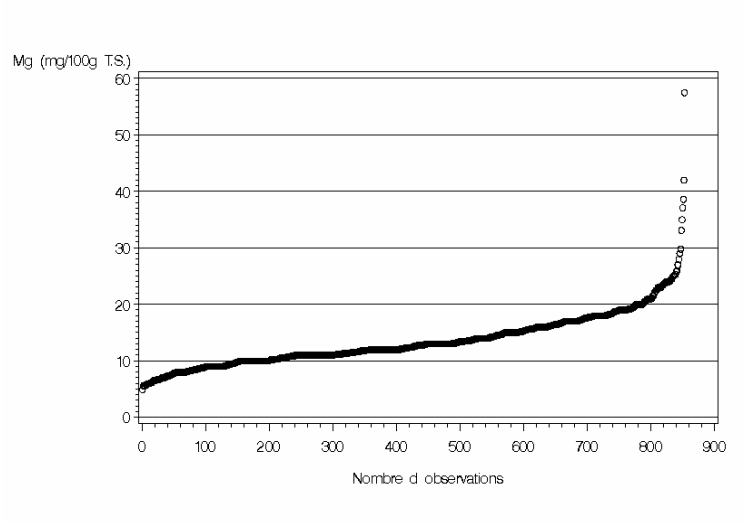


Figure I.7. Diagramme de dispersion des valeurs de magnésium en fonction du nombre d'observations (premier jeu de données).

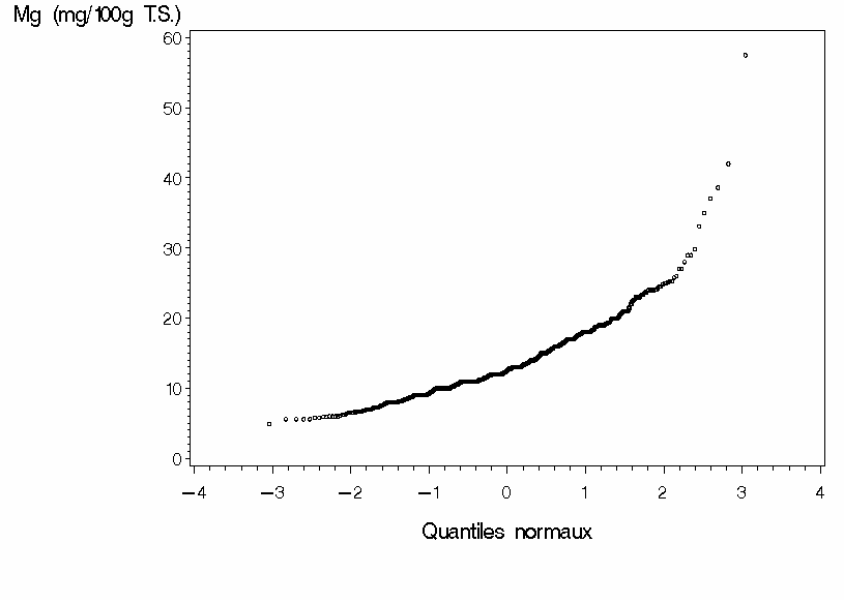


Figure I.8. Graphique des quantiles normaux pour les teneurs en magnésium du premier jeu de données sélectionné.

b. Deuxième jeu de données relatif à des données de calcium

Le deuxième jeu de données contient 1505 observations de calcium dont la moyenne et l'écart-type sont respectivement de 302,1 mg/100g T.S. et 192,8 mg/100g T.S. Les valeurs des percentiles les plus intéressants sont présentées au tableau I.4.

Tableau I.4. Percentiles pour les teneurs en calcium relatives au deuxième jeu de données sélectionné.

Percentiles	Teneur en calcium (mg/100g T.S.)
Minimum	19,0
p5	165,8
p10	185,7
p50	270,0
p90	425,0
p95	503,0
p99	753,3
maximum	3880,1

Comme pour le premier jeu de données, une représentation graphique (figure I.9) des observations, triées par ordre croissant est réalisée. Cette figure permet de montrer que les 7 valeurs supérieures à 1000,0 mg/100g T.S sont anormalement élevées et semblent être aberrantes.

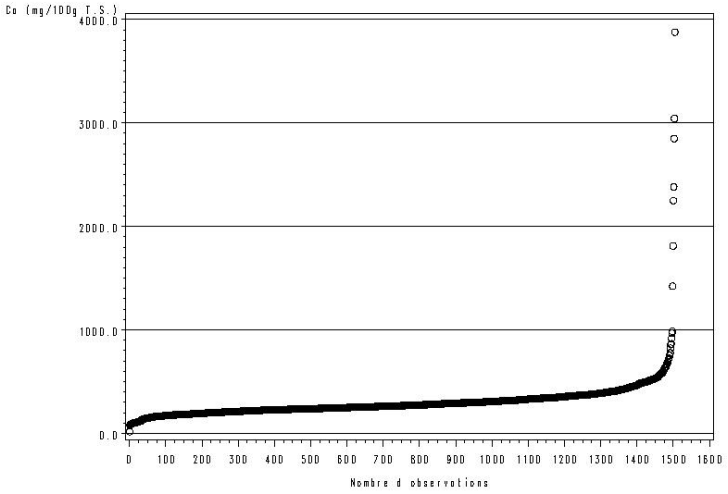


Figure I.9. Diagramme de dispersion des valeurs de calcium en fonction du nombre d'observations (deuxième jeu de données).

L'allure non linéaire du graphique des quantiles normaux, présenté à la figure I.10, indique clairement la non-normalité de la distribution des observations. A nouveau, on constate que les sept dernières observations présentent des valeurs anormalement élevées par rapport au reste des observations. Celles-ci sont rejetées par la technique utilisée actuellement au sein du système de gestion de la base de données de la chaîne SOLS. Pour l'étude des queues de cette distribution, 451 données sont prises en compte et 1054 observations sont ignorées.

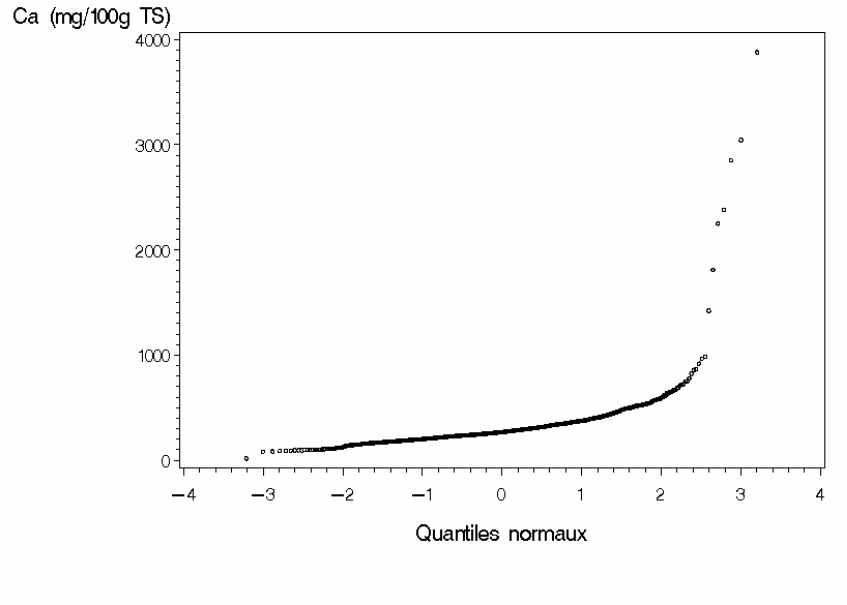


Figure I.10. Graphique des quantiles normaux pour les teneurs en calcium du deuxième jeu de données sélectionné.

2.3.3. Rappel théorique sur les méthodes d'estimation des paramètres

a. Principales méthodes d'estimation

Le processus d'estimation des paramètres des distributions fait appel à la notion d'*estimateur* du paramètre, qui correspond à toute fonction des valeurs observées susceptible de servir à estimer le paramètre. Pour être considéré comme un bon estimateur, celui-ci doit répondre à plusieurs critères de qualité. Il doit être non biaisé et présenter une variance minimale ou asymptotiquement minimale. Il faut également qu'il soit robuste c'est-à-dire peu sensible à la présence de valeurs aberrantes (Dagnelie, 1998a).

1° Diverses méthodes d'estimation des paramètres des distributions existent dans la littérature. La plus classique correspond à la *méthode du maximum de vraisemblance*[29] qui consiste à calculer la fonction de vraisemblance ; celle-ci correspond à la densité de probabilité, relative aux valeurs observées et exprimée en fonction du ou des paramètres à estimer. Les estimateurs du maximum de vraisemblance correspondent aux valeurs des paramètres qui rendent cette fonction maximum ; celui-ci est obtenu en annulant la dérivée de la fonction par rapport au paramètre ou en annulant la dérivée du logarithme (Essenwanger, 1986 ; Dagnelie, 1998a). L'avantage majeur de cette méthode est de fournir des estimateurs de variance (asymptotiquement) minimale dont la distribution d'échantillonnage est asymptotiquement normale. Un des inconvénients à prendre en compte est que ces estimateurs sont parfois biaisés.

D'autres méthodes d'estimations sont couramment utilisées. Sous des conditions de normalité, certaines d'entre elles fournissent des résultats identiques à la méthode du maximum de vraisemblance mais en général, les estimateurs obtenus ne présentent pas la même efficacité (Essenwanger, 1986 ; Dagnelie, 1998a).

2° Une de celles-ci correspond à la *méthode des moments* qui, pour estimer plusieurs paramètres *k,* a pour principe d'égaler les *k* premiers moments estimés de la population, exprimés en fonction des *k* paramètres, aux *k* premiers moments de l'échantillon. Les estimateurs obtenus ont la propriété de présenter des distributions asymptotiquement normales (Dagnelie, 1998a).

3° La *méthode des moindres carrés*[30] est appliquée majoritairement aux problèmes de régression linéaire et non linéaire pour l'estimation des coefficients. Elle est également utilisée dans les processus d'ajustement de courbes. Dans le cas de la régression linéaire, la somme des carrés des écarts entre les points observés et les points correspondants de la droite de régression sont minimisés en annulant les dérivées partielles de cette somme par rapport aux coefficients à estimer. Il suffit alors de résoudre le système d'équations obtenu (Dagnelie, 1998a). Dans le cas de distributions de

[29] En anglais : *maximum likelihood method.*
[30] En anglais *: least square method*.

fréquence, la même procédure est réalisée à partir des dérivées de la fonction de répartition de la distribution étudiée.

4° Dagnelie (1998a) cite aussi la méthode du χ^2 minimum qui s'applique à certains problèmes relatifs aux distributions de fréquence.

5° A partir du *graphique des quantiles*, il est également possible d'estimer, de manière empirique, la valeur des paramètres d'une distribution, après avoir vérifié le bon ajustement de la distribution des valeurs observées à une distribution théorique quelconque. Ceci est réalisé en confrontant graphiquement les quantiles empiriques $\hat{Q}_n(p)$ aux valeurs théoriques fournies par la fonction des quantiles *Q(p)* relative à cette distribution. L'estimation des paramètres est obtenue à partir de la droite de régression estimée sur base du graphique des quantiles. Les paramètres sont calculés de manière spécifique pour chacune des distributions, à partir des valeurs d'ordonnée à l'origine et de pente (paragraphes 2.3.4 à 2.3.7).

Des méthodes spécifiques à certaines distributions ont été développées. Pour les distributions qui nous intéressent, les détails théoriques relatifs à ces estimations sont exposés dans chacun des paragraphes présentant les graphiques des quantiles des distributions.

b. Vérification de la qualité de l'estimation des paramètres

Afin de vérifier la qualité de l'estimation des paramètres suite à l'application d'une méthode d'estimation présentée ci-dessus, une comparaison entre les valeurs observées et les valeurs estimées est réalisée à partir du paramètre appelé *root mean square error* (RMSE), calculé à partir de l'expression suivante :

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\hat{x}_i)^2}{n}},$$

où x_i correspond aux valeurs observées ou à $\hat{Q}_n(p)$ présenté au paragraphe 2.2.2. Pour le premier jeu de données (paragraphe 2.3.2.a), x_i se rapporte donc aux teneurs observées en magnésium ; $\hat{x}_i$ correspond aux valeurs estimées après ajustement en présence de valeurs extrêmes (dans le cas de la distribution exponentielle, la dernière formule du paragraphe 2.3.4.b présente la manière de calculer ces valeurs estimées). Les distributions qui présentent les valeurs de RMSE les plus faibles peuvent être considérées comme étant intéressantes pour la suite de notre étude.

Afin de vérifier de manière empirique la qualité des estimations des paramètres, la visualisation à l'aide des graphiques des quantiles, représentant les valeurs estimées et observées, permet de caractériser les divergences entres les valeurs estimées et observées.

2.3.4. Distribution exponentielle

a. Introduction

La fonction la plus connue de la classe des distributions de Gumbel est la distribution exponentielle, utilisée dans de très nombreuses études statistiques. Elle permet notamment de caractériser le délai d'apparition d'un événement aléatoire tel qu'un accident, la mort d'un individu, la défaillance d'un appareil, au cours d'un intervalle de temps (Johnson et Kotz, 1970 ; Dagnelie, 1998a). Cette distribution a été appliquée lors d'analyses de survie dans le domaine biomédical, le temps de survie étant représenté par une variable aléatoire exponentielle. De nombreuses études sur les distributions exponentielles concernent l'estimation du paramètre de celle-ci et le cas des distributions exponentielles tronquées à droite ou à gauche (Johnson et Kotz, 1970).

b. Aspects théoriques

La distribution exponentielle de paramètre λ présente une probabilité d'occurrence qui s'énonce de la manière suivante :

$$1 - F(x) = \exp(-\lambda x) \text{ où } x > 0 \text{ et } \lambda > 0 ,$$

et comme fonction de densité de probabilité :

$$f(x) = \lambda \exp(-\lambda x) .$$

Le paramètre λ correspond à un paramètre de forme. La figure I.11 présente les fonctions de densité de probabilité de cette distribution pour des valeurs du paramètre λ de 1, 2 et 5.

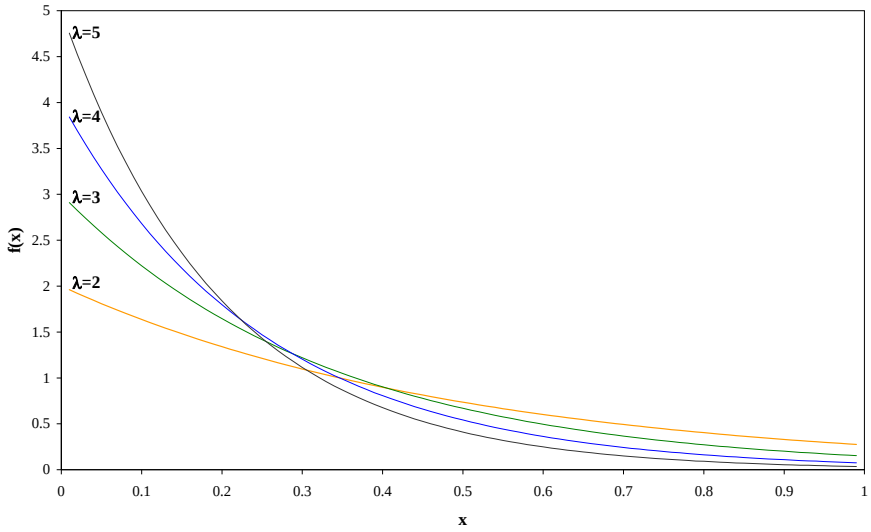


Figure I.11. Fonctions de densité de probabilité pour des distributions exponentielles de paramètres λ=2, λ=3, λ=4 et λ=5.

Pour cette distribution, la moyenne est égale à $\frac{1}{\lambda}$ et la variance à $\frac{1}{\lambda^2}$, l'écart-type est donc égal à la moyenne (Johnson et Kotz, 1970 ; Dagnelie, 1998a). La médiane est égale à $\frac{1}{\lambda}\log 2$ [31]. Les statistiques liées à cette fonction sont relativement simples par rapport aux autres distributions, ce qui l'a rendue très attirante au niveau pratique.

L'estimation du paramètre λ peut être réalisée notamment par le maximum de vraisemblance présentée par Johnson et Kotz (1970). Une méthode d'estimation robuste du paramètre λ, présentée par Brazauskas et Serfling (2001), est basée sur le calcul d'une médiane généralisée, applicable même pour des effectifs faibles mais dépendant de la taille des échantillons. Les auteurs donnent des conseils pour la sélection de l'un de ces estimateurs, en fonction du comportement vis-à-vis des valeurs aberrantes supérieures ou inférieures ou en fonction du niveau possible de contamination par des valeurs aberrantes. Jeevanand et Nair (1998) ont également proposé une méthode d'estimation des paramètres à partir des théories bayesiennes. Cette méthode a été développée à partir d'échantillons pour lequel le nombre de valeurs aberrantes est connu *a priori*.

L'estimation de λ peut également être réalisée de manière empirique par l'utilisation de la fonction des quantiles (Beirlant *et al.*, 1996) et la réalisation de graphiques des quantiles exponentiels. La fonction des quantiles de la distribution exponentielle s'énonce de la manière suivante :

$$Q(p) = -\frac{1}{\lambda}\log(1-p),$$

pour différentes valeurs de $0<p<1$.

La réalisation du graphique des quantiles exponentiels est basée sur l'existence d'une relation linéaire entre les quantiles de toute distribution exponentielle $Q_\lambda(p)$ et les quantiles correspondants de la distribution exponentielle de paramètre λ=1, notés $Q_1(p)$ (Beirlant *et al.*, 1996). On peut donc écrire :

$$Q_\lambda(p)=\frac{1}{\lambda}Q_1(p) \text{ avec } 0<p<1.$$

En pratique, pour réaliser le graphique des quantiles exponentiels d'un échantillon aléatoire et simple, la fonction des quantiles de la population inconnue est remplacée par les quantiles empiriques $\hat{Q}_n(p)$ qui correspondent aux observations triées dans l'ordre croissant x_i^*

[31] **Tout au long de ce travail, l'expression *log* correspond au logarithme népérien, présenté couramment sous la forme *ln*, *log*$_e$, *Log*.**

(paragraphe 2.2.2). Les quantiles empiriques sont ensuite reportés sur l'axe vertical du graphique tandis que les quantiles exponentiels théoriques, correspondants de la distribution exponentielle de paramètre λ=1, équivalents à $(-log(1-p))$, se trouvent sur l'axe horizontal. Le graphique des quantiles exponentiels est alors réalisé avec les points de coordonnées :

$$\left(-log(1-p), \hat{Q}_n(p)\right)$$

pour différentes valeurs de $0<p<1$, avec $p=\frac{i}{n+1}$.

Dans le cas du modèle exponentiel, ce graphique est linéaire. La droite obtenue passe par l'origine et la pente correspond à une approximation du paramètre $1/\lambda$ de la distribution exponentielle. Celle-ci peut être estimée par la méthode des moindres carrés en cherchant à minimiser :

$$\sum_{i=1}^{n}\left(x_i^* - (-b\log(1-\frac{i}{n+1}))\right)^2$$
$$= \sum_{i=1}^{n}\left(x_i^* + b\log(1-\frac{i}{n+1})\right)^2 .$$

L'estimation de la pente $\hat{b}$ est calculée par :

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i^* q_i}{\sum_{i=1}^{n} q_i^2}$$

où

$$q_i = -\log(1-\frac{i}{n+1}) \qquad \text{pour } i=1, 2, \ldots, n.$$

L'ajustement de la droite permet donc de vérifier la structure linéaire du diagramme de dispersion des points et de fournir une estimation du paramètre λ lorsque les conditions de linéarité sont remplies.

Un deuxième paramètre de position θ peut également être ajouté (Johnson et Kotz, 1970) à la distribution présentée ci-dessus et la probabilité d'occurrence se présente de la manière suivante :

$$1-F(x)=exp(-\lambda(x-\theta)) \text{ où } x>\theta \text{ et } \lambda>0 ,$$

avec comme fonction de densité :

$$f(x)=\lambda \exp(-\lambda(x-\theta)) .$$

L'estimation de λ et de θ peut également être réalisée en utilisant la fonction des quantiles et la réalisation de graphiques des quantiles exponentiels. L'ordonnée à l'origine de la droite, obtenue à partir du graphique des quantiles, correspond à l'estimation du paramètre θ; la pente correspondant à une approximation du paramètre $1/\lambda$ de la distribution exponentielle à deux paramètres.

Suite à l'estimation des paramètres par une méthode d'ajustement paramétrique, basée sur la fonction de répartition ou à partir du graphique des quantiles, les valeurs de x peuvent être estimées selon l'expression suivante :

$$\hat{x} = \hat{\theta} + (\frac{1}{\hat{\lambda}})(-\log(1 - \frac{i}{n+1})) \qquad \text{où } i=1, 2, \ldots, n.$$

c. Graphiques des quantiles pour la distribution exponentielle

- Quantiles exponentiels pour les données relatives au Mg

Afin d'avoir un nouvel aperçu de la distribution des valeurs de magnésium, le graphique des quantiles exponentiels a été réalisé pour les 853 valeurs triées dans l'ordre croissant x_i^*, avec les coordonnées $\left(-\log\left(1-\frac{i}{n+1}\right), x_i^*\right)$ (figure I.12). On observe que la partie gauche du graphique n'est pas linéaire tandis que la partie centrale est relativement linéaire. A droite, la valeur la plus extrême (57,5 mg/100g T.S.) décroche à nouveau par rapport aux autres données tandis que les observations situées entre 30 et 50 mg/100g T.S. s'écartent légèrement du reste des observations.

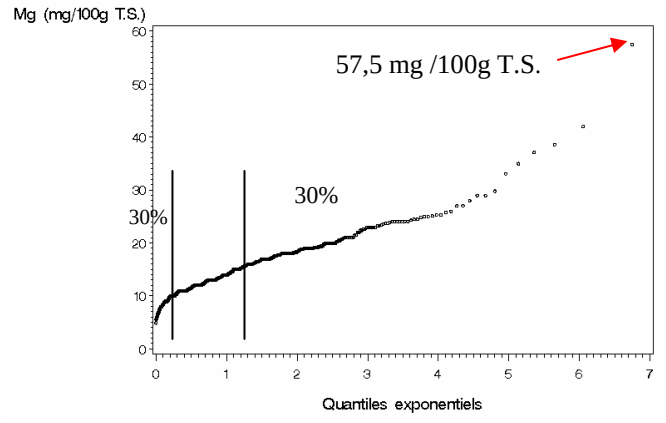


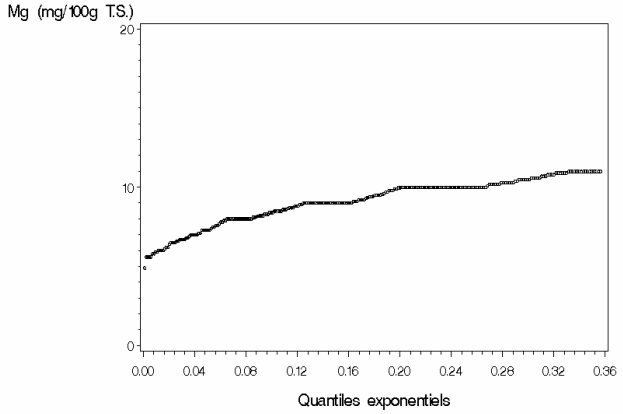
Figure I.12. Distribution exponentielle : graphique des quantiles exponentiels pour le jeu de données relatif au magnésium – ensemble des données (n=853).

La queue de distribution gauche (30% des données) s'étale jusqu'à une valeur de quantile exponentiel de 0,36 tandis que la queue de distribution commence à partir d'une valeur de quantile exponentiel de 1,20.

Les figures I.13(a) et (b) représentent les queues de distributions gauche et droite avec des échelles d'axes différentes par rapport au graphique de la figure I.12. Il est évident que la distribution exponentielle ne pourrait être

adaptée à la partie gauche des distributions (figure I.11), c'est à titre de comparaison par rapport aux autres distributions que nous présentons ces graphiques. La queue de distribution gauche présente une allure légèrement curviligne tandis que celle de droite est relativement rectiligne, les dernières observations s'écartant à nouveau des autres observations. La distribution exponentielle semblerait adaptée à la partie droite dans ce cas bien précis. Dans la suite du travail, la pertinence de l'utilisation de la distribution exponentielle ou d'une autre distribution présentée ci-après sera déterminée à partir de différents critères déterminés au niveau de la méthodologie (paragraphe 4.3).

(a)



(b)

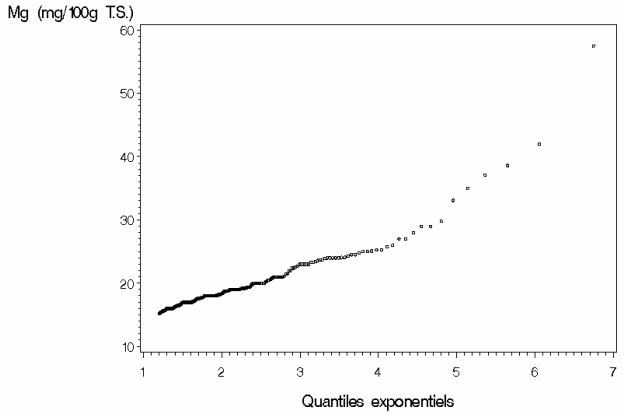


Figure I.13. Distribution exponentielle : graphique des quantiles exponentiels pour le jeu de données relatif au magnésium – (a) queue de distribution gauche (n_1=256) – (b) queue de distribution droite (n_2=256).

- Quantiles exponentiels pour les données relatives au Ca

Comme pour le premier jeu de données, le graphique des quantiles exponentiels a été réalisé pour les 1505 valeurs triées dans l'ordre croissant x_i^* (figure I.14). A partir de cette figure, on observe à nouveau que la partie gauche du graphique n'est pas linéaire tandis que les parties droite et centrale se présentent sous la forme d'une droite, excepté pour les sept observations extrêmes.

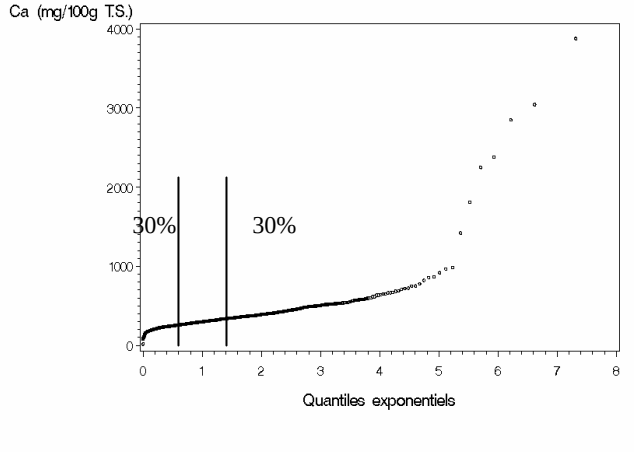
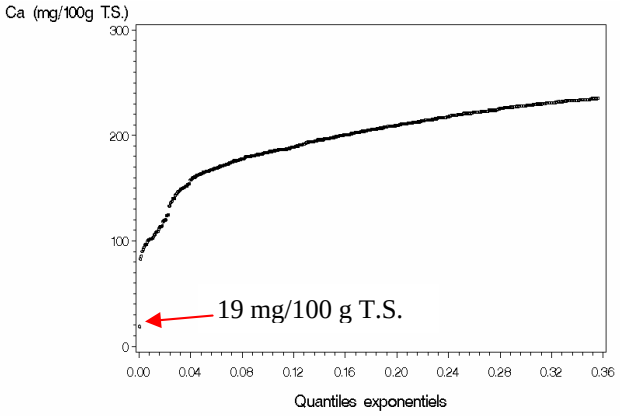


Figure I.14. Distribution exponentielle : graphique des quantiles exponentiels pour le jeu de données relatif au calcium – ensemble des données (n=1505).

Comme pour l'exemple précédent, les mêmes valeurs de quantiles exponentiels à gauche et à droite sont obtenues étant donné qu'on s'intéresse à une proportion équivalente d'observations (30%), que ce soit à gauche ou à droite.

Les figures I.15 (a) et (b) représentent les queues de distributions gauche et droite. Le graphique représentant la queue de distribution gauche montre plus clairement la valeur minimale de 19 mg/100 g T.S. L'allure générale de la queue de distribution gauche est curviligne, comme pour le magnésium, la distribution ne peut être adaptée à gauche. Pour la partie droite, hormis la présence des sept valeurs extrêmes, une relation globalement linéaire est observée. La distribution exponentielle semblerait adaptée à la partie droite pour ce jeu de données, comme dans le cas du magnésium.

(a)



(b)

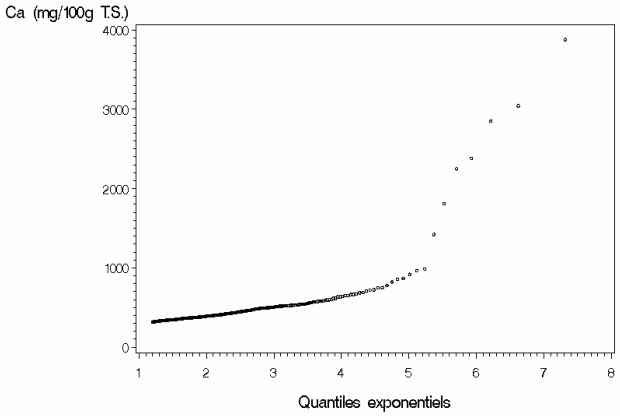


Figure I.15. Distribution exponentielle : graphique des quantiles exponentiels pour le jeu de données relatif au calcium – (a) : queue de distribution gauche (n_1=451) - (b) queue de distribution droite (n_2=451).

d. Tests de détection de valeurs aberrantes

En ce qui concerne la recherche de valeurs aberrantes, divers tests de discordance sont développés par Barnett et Lewis (1994). Ces tests sont les suivants : test de la valeur la plus élevée ou supérieure, test de la valeur la plus faible, test des deux valeurs les plus élevées ou les plus faibles, test de la valeur inférieure ou supérieure simultanément (quel que soit le paramètre de position ou celui-ci étant inconnu), etc. Dans notre cas, ces tests nous semblent limités dans leur utilisation car ils ne permettent de vérifier qu'une

ou deux valeurs à la fois. La méthode de détection de valeurs aberrantes que nous recherchons doit permettre de traiter facilement un nombre d'observations variable et plus élevé qu'une ou deux observations. Deux tests de détection de valeurs aberrantes d'un nombre *k* de valeurs aberrantes situées, soit à droite, soit à gauche de la distribution sont présentés par ces auteurs.

D'autres tests sont proposés mais concernent le problème d'accommodation des valeurs aberrantes ce qui ne répond pas à l'objectif de notre travail.

2.3.5. Distribution de Weibull

a. Introduction

La distribution de Weibull a été développée dans le domaine de la physique, pour la modélisation de la résistance à la rupture de matériaux (Johnson et Kotz, 1970). L'ajustement de la distribution exponentielle n'étant pas assez précise pour décrire ce type de données, une transformation puissance est appliquée aux variables observées et une distribution exponentielle est alors obtenue. La distribution de Weibull apporte ainsi une certaine flexibilité par rapport aux modèles obtenus à partir de la distribution exponentielle. Rappelons que la queue de la distribution de Weibull présente un étalement vers la droite plus faible que celle de la distribution exponentielle. Actuellement, cette distribution est utilisée dans le domaine biomédical (temps de survie), en actuariat et dans tout autre domaine où l'étude des valeurs extrêmes est essentielle. Dans le cas des analyses de survie, le temps de survie est représenté couramment par une variable aléatoire exponentielle ce qui n'est pas toujours idéal. La distribution exponentielle est alors judicieusement remplacée par la distribution de Weibull.

b. Aspects théoriques

La distribution de Weibull présente une probabilité d'occurrence qui s'énonce de la manière suivante :

$$1-F(x)=\exp(-\lambda x^{\tau}) \qquad \text{où } x>0 \text{ et } \tau>0\,,$$

avec comme fonction de densité :

$$f(x)=\lambda\tau\, x^{\tau-1}\exp(-\lambda x^{\tau})\,.$$

Il est possible de présenter la distribution de Weibull en faisant appel aux notions purement mathématiques sur les transformations de variables (Beirlant *et al.*, 1996). Pour cela, il est nécessaire de rappeler l'expression théorique de la probabilité d'occurrence et de la fonction de densité d'une variable aléatoire transformée :

$$1-F_Y(x)=1-F_X(y^{-1}(x))\,,\ \text{pour } x\in y(0,\infty)$$

et

$$f_Y(x)=f_X\left(y^{-1}(x)\right)\left|\ \frac{dy^{-1}(x)}{dx}\ \right|$$

où y^{-1} correspond à la fonction inverse.

Si les variables X suivent une distribution exponentielle, de paramètre $\lambda>0$:

$$\begin{cases} f_X(x)=\lambda \exp(-\lambda x) \\ 1-F_X(x)=\exp(-\lambda x)\,, \end{cases}$$

en appliquant la transformation $y(x)=x^{\frac{1}{\tau}}$, où τ est positif et à partir des formules présentées pour les variables aléatoires transformées, on obtient :

$$1-F_X(x)=\exp(-\lambda x)$$

$$1-F_Y(x)=1-F_X(y^{-1}(x))=\exp(-\lambda\, y^{-1}(x))\,.$$

Comme $y_{(x)}^{-1}=x^{\tau}$,

$$1-F_Y(x)=\exp\left(-\lambda x^{\tau}\right).$$

Pour la fonction de densité de probabilité, on trouve également :

$$f(x)=\lambda\tau\, x^{\tau-1}\exp(-\lambda x^{\tau})\,.$$

Cette transformation de puissance de paramètre τ, appliquée à des variables aléatoires qui suivent une distribution exponentielle de paramètre positif λ, mène à la distribution de Weibull de paramètre λ et τ. Le paramètre τ est appelé l'*index de Weibull*.

Les statistiques liées à cette fonction, présentées par Johnson et Kotz (1970), sont plus complexes que pour la distribution exponentielle et des tables de valeurs de moyennes, d'écart-types et des moments sont présentées en fonction des valeurs des paramètres τ et λ.

Les méthodes d'estimation des paramètres telles que la méthode du maximum de vraisemblance et la méthode des moments sont présentées par Johnson et Kotz (1970). Ces auteurs présentent également une méthode basée sur la médiane de la distribution.

A partir de la fonction de répartition et des paramètres estimés par l'une ou l'autre méthode d'ajustement (moindres carrés, maximum de vraisemblance, etc.), les valeurs estimées sont calculées de la manière suivante :

$$\hat{x}=\left(-\frac{1}{\hat{\lambda}}\quad \log(1-\frac{i}{n+1})\right)^{\frac{1}{\tau}}, \qquad \text{où } i=1, 2, \ldots, n.$$

Le graphique des quantiles permet également d'estimer les paramètres τ et λ de manière non-paramétrique (Beirlant *et al.*, 1996) lorsqu'une relation linéaire est obtenue.

La fonction des quantiles de la distribution de Weibull s'énonce comme suit :

$$Q(p)=\left(-\frac{1}{\lambda}\log(1-p)\right)^{1/\tau}.$$

En appliquant une transformation logarithmique à la fonction des quantiles, on obtient :

$$\log Q(p)=\frac{1}{\tau}\log(\frac{1}{\lambda})+\frac{1}{\tau}\log\left(-\log(1-p)\right) \text{ pour } 0<p<1.$$

Si les données suivent une distribution de Weibull, les points de coordonnées :

$$\left(\log\left(-\log(1-p)\right),\log\hat{Q}_n(p)\right)$$

présentent une relation linéaire de pente $1/\tau$, où $\hat{Q}_n(p)=\hat{Q}_n(\frac{i}{n+1})=x_i^* \ (i=1, 2, 5\ , n)$ et $p=\frac{1}{n+1}, \frac{2}{n+1}, 5\ , \frac{n\text{-}1}{n+1}, \frac{n}{n+1}$.

Le paramètre τ peut être considéré comme un paramètre d'étalement. En effet, plus τ est petit, plus la queue de la distribution est étendue vers la droite.

Le paramètre λ est estimé à partir de la valeur de l'ordonnée à l'origine qui est correspond à $\frac{1}{\tau}\log(\frac{1}{\lambda})$. Donc, si on définit a comme la valeur de l'ordonnée à l'origine de la droite des quantiles, le paramètre λ est équivalent à $\frac{1}{\exp(a\tau)}$.

A partir du graphique des quantiles et des paramètres estimés à partir des valeurs d'ordonnée à l'origine a et de la pente $1/\tau$, les valeurs estimées sont calculées de la manière suivante :

$$\hat{x} = \exp\left(a + \frac{1}{\tau}\log(- \log(1 - \frac{i}{n+1}))\right), \text{ pour } i=1, 2, \ldots, n.$$

Notons que pour les observations qui se distribuent suivant une distribution exponentielle, le graphique des quantiles de la distribution de Weibull se présente sous forme d'une droite de pente proche de 1.

Il est à noter qu'il n'existe pas de test disponible dans la littérature qui permette de vérifier si on peut faire l'hypothèse d'une distribution de Weibull. En général, les tests d'ajustement sont destinés à vérifier la normalité d'une distribution (par exemple, le test de Shapiro et Wilk) ou nécessitent de connaître avec précision les paramètres de la distribution théorique (paramètres de localisation, de forme et d'échelle) afin de comparer les fréquences observées aux fréquences attendues (test d'ajustement de Pearson, test de Kolmogorov et Smirnov, d'Anderson et Darling - Dagnelie 1998b). De plus, nous travaillons à partir des queues des

distributions. Les données que nous exploitons sont donc tronquées et à nouveau, aucune information pertinente n'a été rencontrée.

c. Graphiques des quantiles pour la distribution de Weibull

- Quantiles de la distribution de Weibull pour les données relatives au Mg

Le graphique des quantiles de la distribution de Weibull a été réalisé pour les 853 valeurs triées dans l'ordre croissant x_i^*, avec les coordonnées $\left(\log\left(-\log\left(1-\frac{i}{n+1}\right)\right),\log x_i^*\right)$ (figure I.16). Dans l'ensemble, le graphique des quantiles se présente sous la forme d'une courbe incurvée vers le bas avec une partie très étalée à gauche de la distribution.

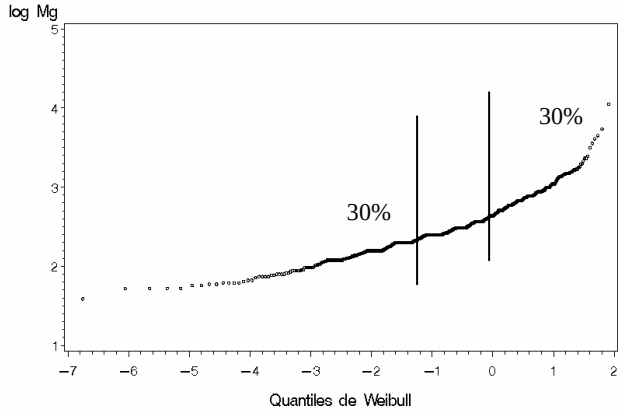


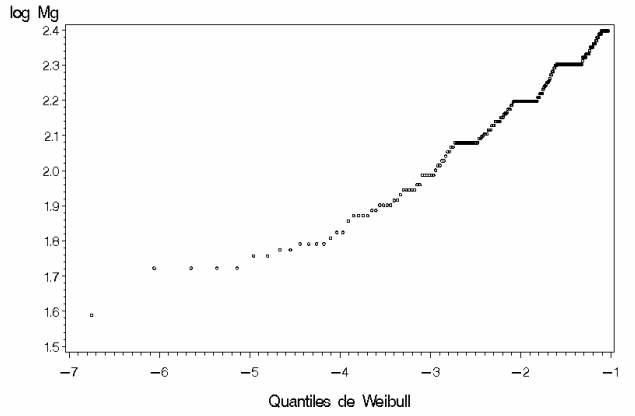
Figure I.16. Distribution de Weibull : graphique des quantiles de la distribution de Weibull pour le logarithme des valeurs de magnésium relatif au premier jeu de données – ensemble des données (n=853).

Pour les deux jeux de données, la queue de distribution gauche s'étale jusqu'à la valeur de quantiles de la distribution de Weibull de –1,03 tandis que la queue de distribution commence à partir d'une valeur de 0,19.

Les figures I.17 (a) et (b) représentent les queues de distributions gauche et droite. On observe que les queues de distribution gauche et droite présentent chacune une allure légèrement curviligne, la queue de distribution droite étant plus proche de la linéarité que la partie gauche. La distribution de Weibull semblerait donc adaptée pour la partie gauche. En ce qui concerne la queue de distribution droite, la distribution de Weibull semble moins adaptée que la distribution exponentielle.

Sur la figure I.17 (a), on observe des « plateaux » qui sont dus à la présence de plusieurs valeurs identiques qui, triées par ordre croissant et portés sur le graphique des quantiles se présentent sous une forme horizontale.

(a)



(b)

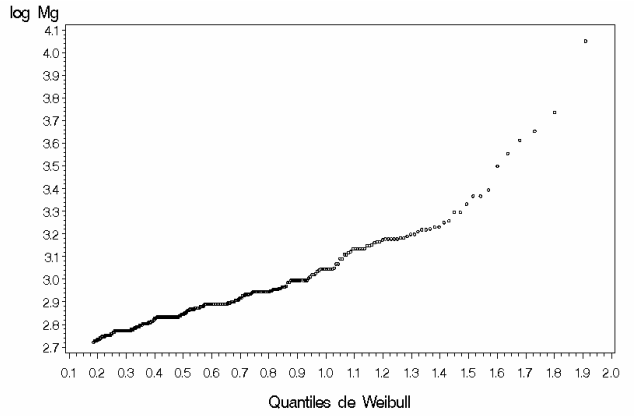


Figure I.17. Distribution de Weibull : graphique des quantiles de la distribution de Weibull pour le logarithme des valeurs de magnésium relatif au premier jeu de données – (a) queue de distribution gauche (n_1=256) - (b) queue de distribution droite (n_2=256).

- Quantiles de la distribution de Weibull pour les données relatives au Ca

Comme pour le premier jeu de données, le graphique des quantiles de la distribution de Weibull a été réalisé pour les 1505 valeurs triées dans l'ordre croissant x_i^* (figure I.18). A partir de cette figure, on observe à nouveau un étalement des observations de la partie gauche. La distribution de Weibull présenterait ainsi un effet de « zoom » ou d'étirement des données sur la

queue de distribution gauche des distributions. Ceci est très intéressant pour l'étude de cette partie des distributions et pourrait permettre de bien mettre en évidence les valeurs suspectes de la queue de distribution gauche. Ceci peut être mis en relation avec une étude très intéressante sur les performances de réseaux pour laquelle Lu et Sedransk (2002) proposent d'utiliser la distribution de Weibull pour modéliser la partie gauche des queues des distributions tandis que la partie droite est traitée à partir des distributions de type Pareto ou les distributions généralisées de Pareto. Rappelons qu'on ne peut généraliser à partir des deux exemples et que ces considérations seront approfondies dans la deuxième partie du travail.

Les figures I.19 (a) et (b) représentent les queues de distributions gauche et droite. Pour la figure de gauche, on observe à nouveau la valeur extrême qui s'écarte du lot des observations. Une relation légèrement de type sigmoïdale est rencontrée pour cette partie gauche. Dans le cas de la queue de distribution droite, excepté les 7 valeurs extrêmes, les valeurs ont tendance à s'accroître plus rapidement vers l'extrémité de la distribution.

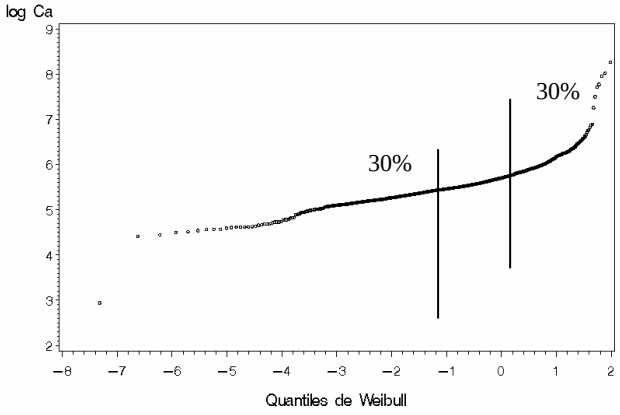
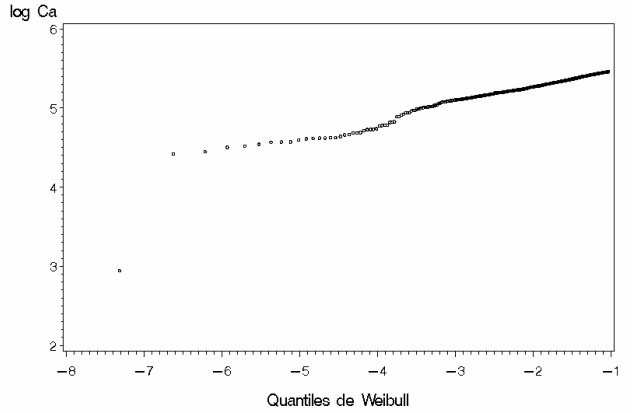


Figure I.18. Distribution de Weibull : graphique des quantiles de la distribution de Weibull pour le logarithme des valeurs de calcium relatif au deuxième jeu de données – ensemble des données (n=1505).

d. Tests de détection de valeurs aberrantes

Il est possible d'imaginer que les tests de détection de valeurs aberrantes, exposés pour la distribution exponentielle, peuvent être appliqués à des données présentant une distribution de Weibull. En effet, si la variable positive X est telle que $Y=X^{\tau}$ possède une distribution exponentielle de paramètre 1, alors X se distribue selon la loi de Weibull de paramètre τ. Si τ est connu, la variable transformée Y est utilisée et les techniques développées pour les distributions exponentielles sont alors facilement applicables.

(a)



(b)

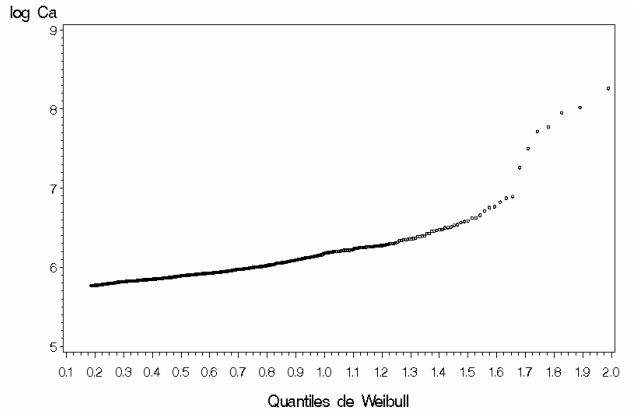


Figure I.19. Distribution de Weibull : graphique des quantiles de la distribution de Weibull pour le logarithme des valeurs de calcium relatif au deuxième jeu de données – (a) queue de distribution gauche (n_1=451) - (b) queue de distribution droite (n_2=451).

2.3.6. Distribution log-normale

a. Introduction

Une distribution classique et connue depuis très longtemps est la distribution log-normale citée dans les domaines les plus divers tels que l'économie, la sociologie, la psychologie, la pharmacologie, l'agronomie, l'entomologie, la géologie, etc. (Aitchison et Brown, 1969 ; Johnson et Kotz, 1970). Par définition, les variables qui suivent une distribution log-normale et qui sont transformées par la fonction logarithmique, sont distribuées de manière normale (Johnson et Kotz, 1970 ; Beirlant *et al.*, 1996 ; Dagnelie, 1998a). Cette caractéristique a bien évidemment rendu la distribution log-normale très attractive car les propriétés de la distribution normale sont aisément transposables à celle-ci.

Les études concernent, par exemple, la distribution de la dose critique pour des médicaments, la distribution de la taille de gisements de gaz et de pétrole (Houghton, 1988), la distribution de la taille de particules dans des agrégats d'origine naturelle tels que les poussières dans les zones industrielles ou le sable dans les problèmes d'érosion (Barndorff-Nielsen, 1977; Barndorff-Nielsen et Christiansen, 1988). La distribution log-normale a également permis d'étudier la durée de vie de produits manufacturés et a ainsi été utilisée dans le processus de gestion de la qualité au sein des entreprises. Les premières études sur les événements rares, tels que les inondations et l'écoulement des eaux, ont également fait appel aux distributions log-normales (Johnson et Kotz, 1970). Ce modèle a également été appliqué dans de nombreuses études géologiques (Barndorff-Nielsen, 1977) et plus particulièrement dans le domaine de la recherche de minerais et de métaux précieux (Aitchison et Brown, 1969). Cependant, la complexité géologique de beaucoup de minerais, tel que le diamant, a donné lieu à des modèles de distributions log-normales mélangées complexes dérivant directement de considérations sur la manière dont les dépôts se forment au cours du temps (Sichel, 1973).

b. Aspects théoriques

La probabilité d'occurrence de la distribution log-normale correspond à :

$$1-F(x)=\int_x^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma u}\exp\left(-\frac{1}{2\sigma^2}(\log u-\mu)^2\right)du \qquad x>0,\ \mu\in\Re\ ,\ \sigma>0.$$

où μ et σ^2 correspondent respectivement à la moyenne et à la variance des valeurs de x transformées par le logarithme (Johnson et Kotz, 1970 ; Beirlant *et al.*, 1996 ; Caers *et al.*, 1996).

La fonction de densité de la distribution log-normale s'énonce comme suit :

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{1}{2\sigma^2}(\log x-\mu)^2\right) \qquad \text{pour } x>0.$$

Cette fonction, qui présente deux paramètres, a été la plus utilisée dans les applications citées, cependant un troisième paramètre de position peut être ajouté à cette expression (Johnson et Kotz, 1970 ; Caers *et al.*, 1996).

L'estimation des paramètres classiques (moyenne, variance, etc.) peut être résolue très simplement en utilisant la transformation $Y=\log(X)$, cependant, de nombreux tableaux et graphiques ont été publiés afin de limiter les calculs (Johnson et Kotz, 1970). Pour ces mêmes paramètres, les relations entre, par exemple, les moyennes et les variances des variables log-normales X et les variables transformées par le logarithme sont présentées par Dagnelie (1998a).

La distribution log-normale peut être standardisée et la distribution *log-normale réduite* est obtenue. Dans ce cas, lorsque la variance tend vers zéro, la distribution log-normale réduite tend vers la distribution normale réduite tandis que lorsque la variance augmente, très rapidement, le comportement de la distribution s'éloigne de la loi normale. Pour différentes valeurs de variance, les auteurs présentent les valeurs correspondantes des coefficients de symétrie et d'aplatissement qui montrent que pour des variances élevées, la queue de la distribution est très étendue du côté droit (Johnson et Kotz, 1970).

La distribution log-normale possède des propriétés intéressantes grâce à l'application du théorème central limite. En effet, si une variable aléatoire X résulte d'un grand nombre d'accroissements successifs qui interviennent indépendamment les uns des autres selon un modèle multiplicatif et si ces accroissements sont tous du même ordre de grandeur, la variable X est approximativement log-normale. Ces conditions sont remplies dans le domaine biologique dans les cas de divisions cellulaires, de reproduction ou de croissance. C'est pour cette raison que la variable de poids d'organismes vivants correspondent à des variables approximativement log-normales contrairement à la variable de taille (Dagnelie, 1998a ; Johnson et Kotz, 1970). Il nous semble intéressant de citer que le produit de deux variables log-normales indépendantes est distribué de manière log-normale. Enfin, pour toute variable X log-normale, la fonction puissance du type : $Z=aX^b$ ($a>0$ et $b\neq 0$) possède également une distribution log-normale (Dagnelie, 1998a).

La fonction des quantiles de la distribution log-normale s'énonce de la manière suivante :

$$Q(p)=\exp(\mu+\sigma\Phi^{-1}(p)).$$

Pour le graphique des quantiles, une relation peut être facilement réalisée à partir du graphique des quantiles de la distribution normale. Le graphique réalisé à partir des points de coordonnées :

$$(\Phi^{-1}(\frac{i}{n+1}) , \log(x_i^*))$$

présente une relation linéaire dans le cas d'une distribution log-normale. La pente de celle-ci correspond à une approximation de paramètre σ de la distribution log-normale tandis que l'ordonnée à l'origine correspond à la valeur de μ.

c. Graphiques des quantiles pour la distribution log-normale

- Quantiles de la distribution log-normale pour les données relatives au Mg

Le graphique des quantiles normaux a été réalisé à partir des 853 données, triées dans l'ordre croissant et transformées par le logarithme (figure I.20). Du premier coup d'oeil, on observe que la distribution log-normale semble convenir aux données relatives au magnésium dans leur ensemble car une relation linéaire est obtenue pour les points de coordonnées.

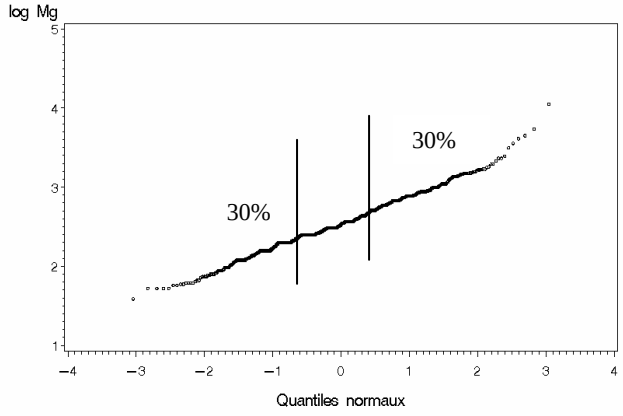
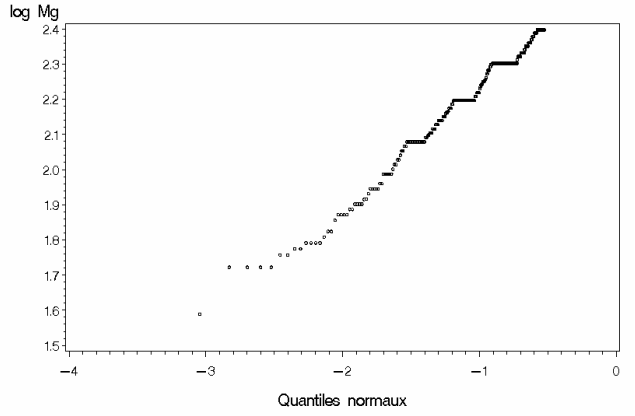


Figure I.20. Distribution log-normale : graphique des quantiles normaux pour le logarithme des valeurs de magnésium du premier jeu de données – ensemble des données (n=853).

Pour les deux jeux de données, la queue de distribution gauche s'étale jusqu'à une valeur de quantiles normal de –0,525 tandis que la queue de distribution commence à partir d'une valeur de quantile normal de 0,525. Les figures I.21 (a) et (b) représentent les queues de distributions gauche et droite. Pour les deux parties, la distribution log-normale semble convenir aux données car une relation linéaire est obtenue. Pour la partie gauche, la distribution log-normale semble plus adaptée aux données que la distribution exponentielle. L'effet de zoom rencontré avec la distribution de Weibull n'est pas aussi flagrant pour la distribution log-normale.

(a)



(b)

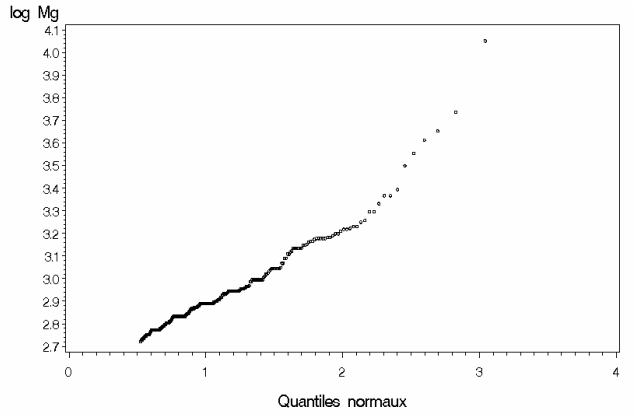


Figure I.21. Distribution log-normale : graphique des quantiles normaux pour le logarithme des valeurs de magnésium du premier jeu de données – (a) queue de distribution gauche (n_1=256) - (b) queue de distribution droite (n_2=256).

- Quantiles de la distribution log-normale pour les données relatives au Ca

Comme pour le premier jeu de données, le graphique des quantiles normaux a été réalisé pour les 1505 valeurs triées dans l'ordre croissant x_i^* (figure I.22). A partir de cette figure, on observe une tendance générale linéaire avec la présence d'un « saut » dans la partie gauche de la distribution.

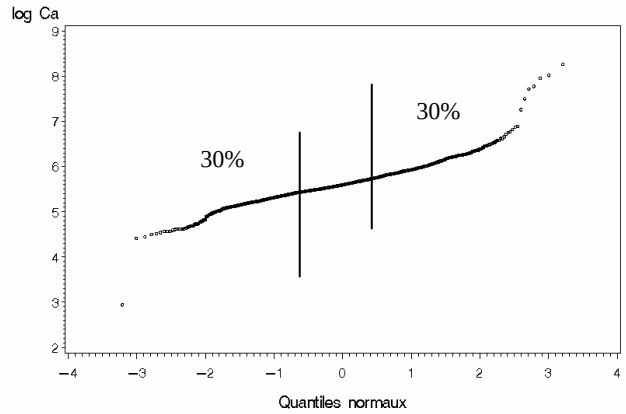
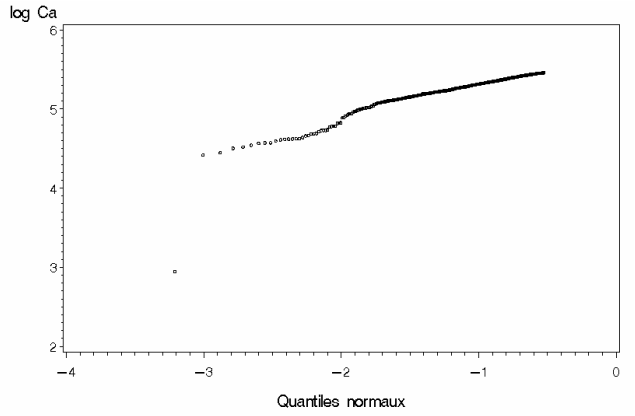


Figure I.22. Distribution log-normale : graphique des quantiles normaux pour le logarithme des valeurs de calcium du deuxième jeu de données - ensemble des données (n=1505).

Les figures I.23 (a) et (b) représentent les queues de distributions gauche et droite. La figure de gauche indique à nouveau une allure sigmoïdale déjà rencontrée dans le cas de la distribution de Weibull, ceci pourrait être dû à la présence d'un mélange de distribution. Pour la partie droite, on observe que les valeurs extrêmes (excepté les 7 valeurs suspectes) ont tendance à augmenter plus rapidement vers la queue de la distribution comme c'était le cas pour la distribution de Weibull.

(a)



(b)

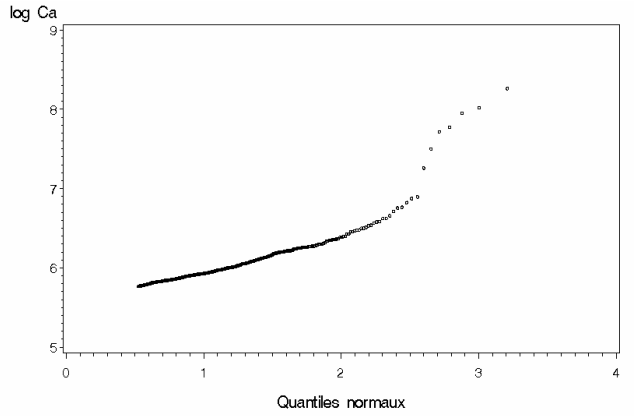


Figure I.23. Distribution log-normale : graphique des quantiles normaux pour le logarithme des valeurs de calcium du deuxième jeu de données – (a) queue de distribution gauche (n_1=451) - (b) queue de distribution droite (n_2=451).

d. Tests de détection des valeurs aberrantes

Par définition, les variables qui suivent une distribution log-normale transformées par la fonction logarithmique, sont distribuées de manière normale (Dagnelie, 1998a). Pour toutes les valeurs anormales d'un échantillon se distribuant selon une loi log-normale, les tests de discordance proposés dans Barnett et Lewis (1994) peuvent être appliqués sur les valeurs de l'échantillon, transformées par le logarithme.

2.3.7. Distribution de Pareto et de type Pareto

a. Introduction

La loi de Pareto trouve son origine en économie et établit que, dans une société donnée et pour une période donnée, la distribution du nombre de personnes en relation avec leur revenu a une très forte tendance à décroître dans leur partie supérieure, c'est-à-dire que cette distribution correspond à une fonction puissance décroissante de ce revenu. En d'autres mots, la queue supérieure de la distribution du logarithme de la variable tend vers une exponentielle (Barndorff-Nielsen, 1977).

Des références sont faites également par Mandelbrot (1963) qui discute de la loi de Pareto en relation avec les distributions de probabilité et qui, tout en traitant de questions économiques, indique des phénomènes parétiens en géologie, géographie, météorologie et en physique.

Dans un cas d'application de la distribution de la taille des diamants, les modèles de type Pareto sont très intéressants, la distribution log-hyperbolique faisant partie de ce type de distribution a été utilisée par Caers *et al.* (1996). D'autres modèles de distributions de type Pareto sont également applicables dans le domaine des métaux comme l'or, tels que l'ont mentionné Sichel *et al.* (1995).

Actuellement, la distribution est toujours utilisée en économie mais a pris un large développement théorique dans le domaine des assurances et fait actuellement son apparition dans le monde de la finance. De nombreuses études sur les performances des réseaux de télécommunication font également appel à cette distribution.

b. Aspects théoriques

La distribution de Pareto, de paramètre α, présente une probabilité d'occurrence qui s'énonce selon l'expression suivante :

$$1-F(x)=x^{-\alpha} \text{ ou } 1-F(x)=x^{-\frac{1}{\gamma}}, \text{ avec } x>1 \text{ et } \alpha=\frac{1}{\gamma}.$$

La fonction de densité correspond à :

$$f(x)=\alpha x^{-\alpha-1} \text{ pour } x>1.$$

Le paramètre α est positif et est appelé *index de Pareto*. Cette distribution de Pareto, à un paramètre, est aussi appelée *distribution stricte de Pareto*.

La **distribution stricte de Pareto** peut également être présentée en suivant les notions mathématiques sur les transformations de variables, comme exposées pour la distribution de Weibull (paragraphe 2.3.5). Pour obtenir la distribution stricte de Pareto, il faut considérer les variables X, qui suivent une distribution exponentielle de paramètre $\alpha>0$:

$$\begin{cases} f_X(x)=\alpha \exp(-\alpha x) \\ 1-F_X(x)=\exp(-\alpha x)\,. \end{cases}$$

En appliquant la transformation $y(x)=\exp(x)$ et à partir des formules présentées pour les variables aléatoires transformées :

$$1-F_X(x)=\exp(-\alpha x)$$

$$1-F_Y(x)=1-F_X(y^{-1}(x))=\exp(-\alpha y^{-1}(x))$$

comme $y_{(x)}^{-1}=\log(x)$, on obtient

$$1-F_Y(x)=\exp(-\alpha \log x)=\exp(\log x^{-\alpha})=x^{-\alpha} \text{ pour } x>1.$$

Une transformation exponentielle appliquée à des variables aléatoires qui suivent une distribution exponentielle de paramètre positif α, nous mène à la distribution stricte de Pareto de paramètre α. En d'autres termes, les variables aléatoires distribuées selon la distribution stricte de Pareto et transformées par la fonction logarithmique suivent une distribution exponentielle ayant comme paramètre α ou $1/\gamma$.

La fonction des quantiles de la distribution stricte de Pareto s'énonce de la manière suivante :

$$Q(p)=(1-p)^{-\frac{1}{\alpha}}=(1-p)^{-\gamma}$$

Dans le cas de la *distribution stricte de Pareto,* lorsqu'on établit la courbe des quantiles exponentiels théoriques, qui correspondent à $-\log(1-p)$, par rapport aux quantiles empiriques correspondants, transformés par la fonction logarithmique, une droite passant par l'origine et de pente γ est obtenue (Beirlant *et al.*, 1996).

Comme pour les autres distributions présentées, la distribution de Pareto peut présenter un deuxième paramètre de position κ et la probabilité d'occurrence s'exprime alors de la manière suivante (Johnson et Kotz, 1970) :

$$1-F(x)=\left(\frac{x}{\kappa}\right)^{-\alpha}, \text{ avec } \kappa>0.$$

De même, lorsqu'on établit la courbe des quantiles exponentiels théoriques par rapport aux quantiles empiriques correspondants, transformés par la fonction logarithmique, on observe une droite dont l'ordonnée à l'origine correspond au logarithme du paramètre κ, la pente étant égale à γ.

Suite à l'estimation des paramètres à partir de ce graphique des quantiles, les valeurs de x peuvent alors être estimées selon l'expression suivante :

$$\hat{x} = \exp\left((\log \hat{\kappa} + \hat{\gamma} q_i\right),$$

où q_i correspond aux quantiles exponentiels et s'exprime de la manière suivante :

$$q_i = -\log(1 - \frac{i}{n+1}) \qquad \text{pour } i=1, 2, \ldots, n.$$

Pour les distributions de type Pareto, la probabilité d'occurrence de celles-ci peut être décomposé de la manière suivante (Beirlant *et al*.,1996) :

$$1-F(x)=x^{-\frac{1}{\gamma}} l_F(x).$$

Ces **distributions de type Pareto** diffèrent les unes des autres par la fonction $l_F(x)$, appelée *fonction de variation lente*[32], qui satisfait à l'équation suivante :

$$\frac{l_F(\lambda x)}{l_F(x)} \to 1 \text{ lorsque } x\to\infty \text{ et pour tout } \lambda>0.$$

L'indice F de cette fonction fait référence à la distribution à laquelle on s'intéresse (Burr, Fréchet, etc.). Dans le cas de la distribution stricte de Pareto, $l_F(x)$ est équivalent à 1.

En ce qui concerne l'estimation du paramètre γ des distributions de type Pareto, le graphique des quantiles de Pareto, est linéaire pour les valeurs de x élevées, situées à droite de la distribution. Une illustration sera fournie dans le paragraphe suivant (figures I.24 et I.26). En effet, pour ces distributions de type Pareto, il est possible de vérifier visuellement, pour les valeurs de x élevées, l'hypothèse d'une distribution stricte de Pareto, en inspectant le graphique de dispersion des points ayant comme coordonnées :

$$\left(-\log(1-p), \log x_i^*\right) \qquad \text{pour } i=1, ..., n.$$

Si k représente le nombre de valeurs extrêmes à partir de laquelle la linéarité apparaît, une estimation de γ est obtenue par l'estimation de la pente du graphique des quantiles de Pareto à partir d'un point d'ancrage de coordonnées :

$$\left(-\log(\frac{k+1}{n+1}), \log x_{n-k}^*\right).$$

[32] En anglais : *slowly changing function.*

L'estimation de la pente est réalisée à partir de la méthode des moindres carrés pondérés. Les valeurs liées à la pondération sont largement développées par Csörgö *et al.* (1985), Deckers *et al.* (1989), Beirlant *et al.* (1996), Beirlant et Goegebeur (2000), Vandewalle (2004), Beirlant *et al.* (2004).

Une autre manière d'estimer l'index de Pareto peut être réalisée à partir de l'*estimateur de Hill*, noté $\hat{\gamma}_{k,n}^{H}$ ou $H_{k,n}$, en considérant k comme le nombre de valeurs extrêmes prises en compte lors de l'estimation et n, le nombre total d'observations. Cet estimateur s'énonce selon l'expression suivante (Hill, 1975) :

$$\hat{\gamma}_{k,n}^{H}=H_{k,n}=\frac{1}{k}\sum_{i=1}^{k}\log x_{n-i+1}^{*}-\log x_{n-k}^{*} .$$

L'index de Pareto α peut être estimé à partir de $1/\hat{\gamma}_{k,n}^{H}$ (Hill, 1975). Selon Beirlant *et al.* (1996), l'estimateur de Hill peut être interprété comme étant l'augmentation moyenne du logarithme de la variable X dans le graphique des quantiles de Pareto à partir du point d'ancrage défini ci-dessus.

La qualité de l'estimateur de Hill est très variable en fonction du nombre de valeurs extrêmes prises en compte lors du calcul de celui-ci. Il est donc difficile, en pratique, de déterminer le nombre de valeurs extrêmes à considérer afin d'optimiser l'estimation de γ. Pour résoudre ce problème, de nombreux auteurs ont utilisé un critère de décision basé sur la minimisation de l'erreur du carré moyen asymptotique[33] de l'estimateur de Hill, noté AMSE, lié à la variance de l'estimateur et au biais. Selon Beirlant *et al.* (1996) et à partir des formules de calculs d'AMSE, lorsque le nombre d'observations extrême k est élevé pour l'estimation de γ, la variance de l'estimateur est plus faible. Par contre, lorsque le nombre de valeurs extrêmes est faible, le biais diminue. Le compromis entre la variance et le biais doit mener à un choix optimal du nombre de données à utiliser pour l'estimateur de Hill et donc pour l'index de Pareto (Caers *et al.*, 1996 ; Beirlant *et al.* ,1996). Finalement, la valeur de k qui minimise la courbe d'AMSE estimée correspond au nombre optimal k_{opt} de valeurs extrêmes à prendre en compte.

De même, l'estimation de γ à partir de la pente de la partie droite du graphique des quantiles de la distribution de type Pareto, dépend du nombre de valeurs extrêmes prises en compte. Pour Beirlant *et al.* (1996), le problème de la détermination de la valeur optimale du nombre de valeurs k peut être considéré simplement comme un problème de régression qui permet de déterminer le point à partir duquel une estimation optimale linéaire est obtenue au niveau du graphique des quantiles exponentiels. Cette méthode a été largement développée par Beirlant *et al.* (2004).

[33] En anglais : *Asymptotic Mean Square Error* – acronyme : AMSE.

Dans le cadre de notre étude, il est important de signaler que les méthodes présentées ont été développées exclusivement pour les queues de distributions supérieures. Caers *et al.* (1996) indiquent que la partie gauche peut également être traitée de manière similaire en prenant en compte, non plus de la probabilité d'occurrence, mais de la fonction de répartition qui s'énonce de la manière suivante :

$$F(x)=x^{\xi}l(x) \qquad \text{quand } x\to 0,\ \xi>0,$$

et en inspectant les graphiques des quantiles de Pareto à gauche du point d'ancrage :

$$\left(\log\left(\frac{k+1}{n+1}\right),\log x^{*}_{k+1}\right).$$

D'autres possibilités pourraient néanmoins été appliquées pour traiter ces données. Une deuxième approche consisterait à s'intéresser à $1/X$ mais nous nous situons alors dans le cas des distributions des valeurs extrêmes de type Weibull (Goegebeur, 2003). Les méthodes de traitement des valeurs extrêmes développées pour ces distributions limites pourraient être utilisées.

Les techniques développées dans le cadre des distributions généralisées de Pareto pourraient également être utilisées en s'intéressant aux valeurs situées en dessous de valeurs seuils situées à gauche des distributions. Ces dernières distributions présentent l'avantage de tenir compte de l'ensemble des valeurs minimales situées en dessous d'une valeur fixée. Cependant, les théories liées à ces distributions sont complexes et très lourdes au niveau mathématique.

Concernant le cas de la partie gauche de la distribution, il est très étonnant d'observer qu'il existe peu d'intérêt pour cette portion de distribution. De très rares applications pratiques ont été rencontrées dans la littérature.

c. Graphiques des quantiles pour la distribution de Pareto

- Quantiles de la distribution de Pareto pour les données relatives au Mg

Le graphique des quantiles exponentiels a été réalisé pour les 853 valeurs observées transformées par le logarithme, triées dans l'ordre croissant x_i^{*}, de coordonnées $\left(-\log\left(1-\frac{i}{n+1}\right), \log x_i^{*}\right)$ (figure I.24).

On observe que la partie gauche du graphique n'est pas linéaire tandis que les parties centrale et droite sont relativement linéaires. A droite, la valeur la plus extrême (57,5 mg/100g T.S.) décroche à nouveau par rapport aux autres données. Les observations qui correspondent à celles situées entre 30 et 50 mg/100g T.S. s'écartent à peine du reste des observations.

Les figures I.25 (a) et (b) représentent les queues de distributions gauche et droite et permettent de montrer de manière très claire la non-linéarité de la

partie gauche et la tendance linéaire de la partie droite. Dans le cas de cet exemple, la distribution de Pareto semble donc convenir pour la partie droite de la distribution ce qui n'est pas le cas pour la partie gauche.

Notons que les valeurs des quantiles exponentiels sont identiques à ceux calculés lors de l'étude la distribution exponentielle.

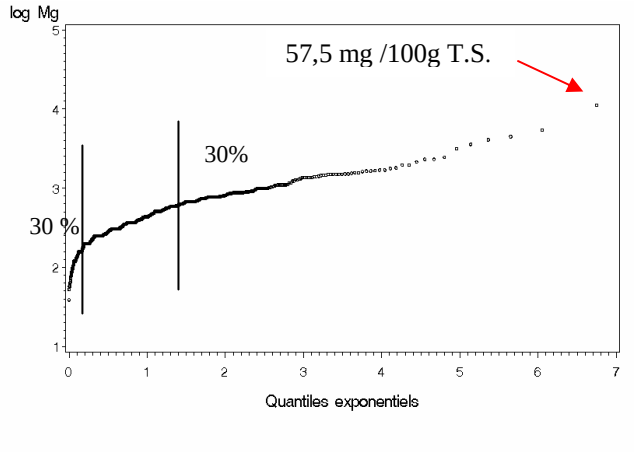
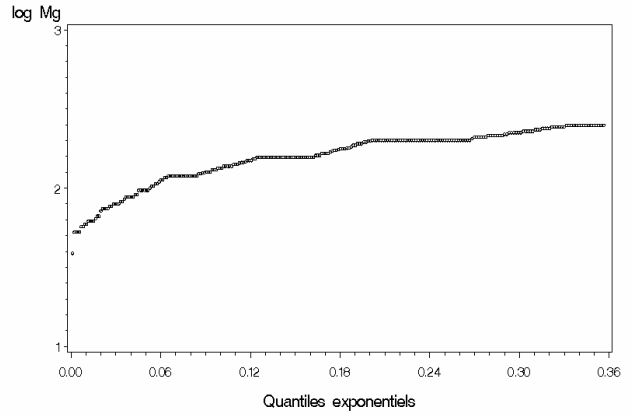


Figure I.24. Graphique des quantiles exponentiels pour le logarithme des valeurs de magnésium du premier jeu de données – ensemble des données (n=853).

(a)



(b)

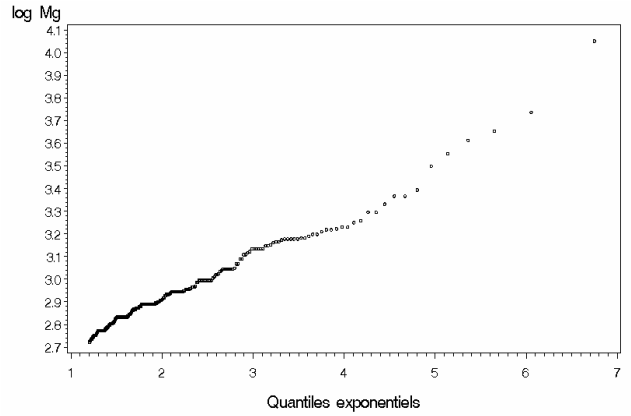


Figure I.25. Graphique des quantiles exponentiels pour le logarithme des données de magnésium du premier jeu de données – (a) queue de distribution gauche (n_1=256) - (b) queue de distribution droite (n_2=256).

- Quantiles de la distribution de Pareto pour les données relatives au Ca

Comme pour le premier jeu de données, le graphique des quantiles exponentiels a été réalisé pour les 1505 valeurs triées dans l'ordre croissant x_i^* (figure I.26). A partir de cette figure, on observe que la partie gauche du graphique n'est pas linéaire tandis que la partie droite se présente sous la forme d'une droite. Ceci indique donc un bon ajustement du modèle de type Pareto dans le cas bien précis de cet exemple. Les valeurs les plus extrêmes qui ne suivent pas la linéarité du quantile de la distribution de Pareto pourraient donc être considérées comme des valeurs aberrantes pour le modèle de type Pareto.

Les figures I.27 (a) et (b) représentent les queues de distributions gauche et droite avec des échelles d'axes différentes par rapport au graphique reprenant l'ensemble des données. Dans le cas de la partie droite, on observe que les quelques observations, situées avant les 7 valeurs qui décrochent totalement, suivent plus la tendance linéaire que pour les distributions précédentes.

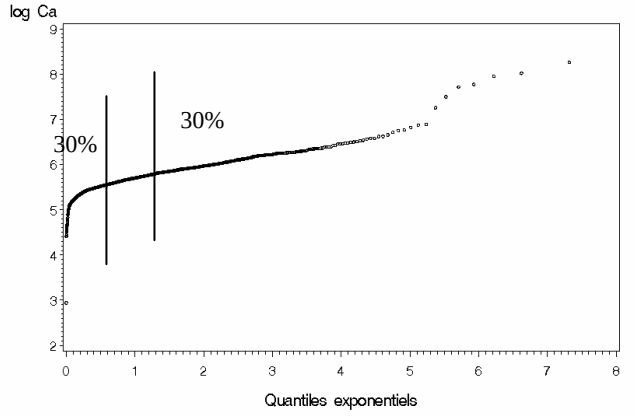
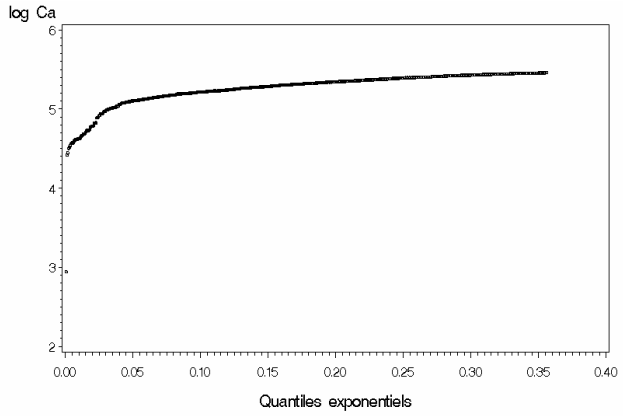


Figure I.26. Graphique des quantiles exponentiels pour le logarithme des données de calcium relative au deuxième jeu de données – ensemble des données (n=1505).

(a)



(b)

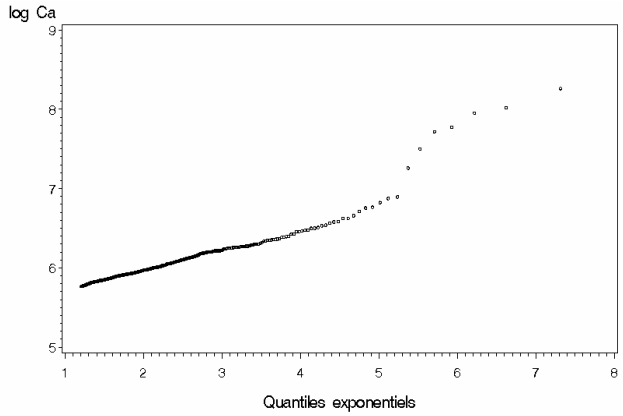


Figure I.27. Graphique des quantiles exponentiels pour le logarithme des données de calcium relative au deuxième jeu de données – (a) queue de distribution gauche (n_1=451) - (b) queue de distribution droite (n_2=451).

d. Tests de détection des valeurs aberrantes

En ce qui concerne la recherche de valeurs aberrantes par des tests de discordance, si une variable aléatoire X suit une distribution de type Pareto, d'index α et si la transformation $Y=\log X$ est appliquée, la fonction de répartion de Y est alors de la forme suivante (Barnett et Lewis, 1994) :

$$F(x)=1-\exp(-\alpha\, y) \qquad \text{pour } y\geq 1,$$

Y possède ainsi une distribution exponentielle de paramètre α.

Supposons que l'hypothèse de travail considère qu'un échantillon trié par ordre croissant est distribué selon la loi de Pareto $x_{(1)}, x_{(2)}, ..., x_{(n)}$ et contient une ou plusieurs valeurs aberrantes. Les valeurs transformées $\log x_{(1)}, \log x_{(2)}, ..., \log x_{(n)}$ sont alors également triées dans l'ordre croissant et les valeurs $y_{(1)}, y_{(2)}, ..., y_{(n)}$ sont issues d'une distribution Y exponentielle. Les valeurs $x_{(1)}, x_{(2)}, ..., x_{(n)}$ de l'échantillon peuvent faire l'objet de tests de discordance à partir des valeurs correspondantes de y qui concernent les distributions exponentielles. Les tests concernent alors les distributions exponentielles et sont présentés par Barnett et Lewis (1994).

Jeevanand et Nair (1993) ont élaboré, en suivant les théories bayesiennes, une méthode de prédiction des intervalles autour des statistiques d'ordre, à partir d'échantillons distribués selon la distribution de Pareto. Les observations qui ne sont pas incluses à l'intérieur de l'intervalle estimé sont considérées comme valeurs aberrantes. La méthode a cependant été développée dans le cas où une seule valeur aberrante, connue *a priori*, était présente. Ces mêmes auteurs ont également utilisé les méthodes bayesiennes pour estimer le paramètre de la distribution à partir d'un nombre de valeurs aberrantes connues *a priori* (Jeevanand et Nair, 1998).

2.3.8. Distribution de Burr

Beirlant *et al.* (1998) présentent également la distribution de Burr (Burr, 1942). Cette distribution dissymétrique est très exploitée dans le domaine de la théorie des valeurs extrêmes car elle présente une queue de distribution très étalée vers la droite. Elle correspond à l'une des distributions les plus importantes de la famille des distributions de type Pareto (Kotz et Johnson, 1982).

Beirlant *et al.* (1998) et Goegebeur *et al.* (2002 ; 2005) ont utilisé la distribution de Burr lors d'analyses multivariées et dans le cas de modèles de régression où une variable dépendante présente une queue de distribution très dissymétrique. Une application, basée sur le deuxième jeu de données de calcium de ce travail, a été réalisée et a donné lieu aux publications de Goegebeur *et al.* (2002; 2005). Ces auteurs ont utilisé la distribution de Burr dans le cas multivarié en prenant en compte la covariable pH pour l'estimation de l'index de Pareto. Nous ne présentons cependant pas cette distribution dans cette étude car nous nous limitons au cas univarié.

2.3.9. Liens entre les distributions

Dans ce paragraphe, nous présentons une synthèse des relations entre les distributions présentées dans les paragraphes précédents. Les caractéristiques principales de l'ensemble de ces distributions sont regroupées au sein du tableau I.5.

Tableau I.5. Tableau récapitulatif des distributions dissymétriques étudiées.

Distributions de type Gumbel	**Probabilité d'occurrence 1-F(x)**	**Index des valeurs extrêmes**	**Domaine**
Weibull	$\exp\left(-\lambda x^{\tau}\right)$	0	$x>0$, $\lambda>0$, $\tau>0$
Exponentielle	$\exp\left(-\lambda x\right)$	0	$x>0$, $\lambda>0$
Log-normale	$\int_{x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma u} \exp\left(-\frac{1}{2\sigma^2\left(\log(u)\right)\mu^2}\right) du$	0	$x>0$, $\mu\in\Re$, $\sigma>0$
Distributions de type Pareto		**Index de Pareto**	
Pareto	$x^{-\alpha}$	α	$x>1$, $\alpha>0$

Notons que de nombreux tests de discordance présentés dans la littérature font référence aux distributions classiques (normales, exponentielles, gamma) à partir desquelles la majorité des distributions présentées peuvent s'exprimer. Il est donc essentiel de bien maîtriser les relations entre chacune des distributions.

Comme expliqué au paragraphe sur les distributions des valeurs extrêmes, la distribution gamma est citée couramment dans la littérature car elle permet de faire le lien entre les distributions normales, exponentielle, χ^2, Poisson, etc. Il nous semble donc nécessaire de la présenter ici, afin d'aider à la compréhension des liens entre les distributions, même si aucune application pratique, basée sur la distribution gamma, n'a été rencontrée dans le cadre de la théorie sur les valeurs extrêmes.

La distribution gamma peut présenter trois paramètres (γ, β, α) et la fonction de densité de probabilité correspondante est de la forme suivante (Johnson et Kotz, 1970) :

$$f(x)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}(x-\gamma)^{\alpha-1}\exp(-(x-\gamma)/\beta)\,.$$

Les paramètres γ et β sont des paramètres d'échelle et α correspond à un paramètre de forme. Le paramètre γ est en général considéré comme étant nul ce qui conduit à la distribution gamma à deux paramètres (α, β). La forme standard de la distribution gamma correspond au cas où β=1 et la

fonction de densité de probabilité s'énonce de la manière suivante (Johnson et Kotz, 1970) :

$$f(x)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}\exp(-x)\,.$$

En ce qui concerne la **distribution exponentielle**, de nombreuses distributions sont liées à celle-ci (Johnson et Kotz, 1970) :

- En liaison avec la distribution de Weibull, si une variable X, est telle que la transformation $Y=X^{\tau}$ se présente sous la forme d'une distribution exponentielle de paramètre égal à 1, alors X suit une distribution de Weibull de paramètre τ.
- La fonction exponentielle est un cas particulier de la distribution gamma à deux paramètres et pour laquelle les paramètres β et α valent 1. La distribution de type exponentielle obtenue se présente sous la forme $f(x)=\exp(-x)$. De plus, la somme de variables aléatoires indépendantes distribuées de manière exponentielle présente une distribution gamma, celle-ci a dès lors été utilisée dans la théorie des comptages aléatoires et dans d'autres sujets associés au processus aléatoire dans le temps et en particulier le phénomène de précipitation en météorologie.
- La distribution exponentielle de paramètre β=2 correspond à la distribution χ^2 de paramètre k=2 (Dagnelie, 1998a).
- Des relations intéressantes entres les distributions exponentielles et les distributions de Poisson sont exposées par Johnson et Kotz (1970). Les auteurs comparent ces deux distributions de la manière suivante : les distributions exponentielles concernent la longueur des intervalles de temps séparant deux réalisations successives d'un même événement tandis que les distributions de Poisson sont relatives aux nombres de réalisations d'une même événement au cours d'un intervalle de temps de longueur donné (Dagnelie, 1998a). Les distributions exponentielles et de Poisson sont utilisées comme modèles classiques pour étudier les problèmes de congestion sur le réseau internet (Fischer *et al.*, 2001).

Concernant la **distribution de Weibull,** certaines distributions présentent également des relations avec celle-ci. Hormis la relation avec la distribution exponentielle présentée ci-dessus, on peut citer que, d'après Beirlant *et al.* (1998), la distribution de Burr, faisant partie des distributions de type Pareto peut être obtenue suite à un mélange lissé de distributions de Weibull avec une proportion correspondant à une distribution gamma à un paramètre.

Quant à la **distribution log-normale**, par définition, les variables qui suivent une distribution log-normale et qui sont transformées par la fonction logarithmique, sont distribuées de manière normale (Dagnelie, 1998a). D'autre part, Barndorff-Nielsen (1977) indique que la distribution log-normale correspond à la distribution limite de la distribution hyperbolique et que cette dernière distribution peut être générée par un mélange de distributions log-normales.

Notons que par définition, la **distribution gamma** est associée aux variables aléatoires distribuées de manière normale comme étant la distribution de la somme des carrés de variables normales réduites indépendantes ; c'est cette propriété qui a donné naissance à la distribution gamma en théorie statistique (Johnson et Kotz, 1970 ; Dagnelie, 1998a). Dans le cas des distributions théoriques à une dimension, si U_1, U_2, ... U_n, sont des variables normales réduites indépendantes, alors la distribution de $\sum_{j=1}^{\nu} U_j^2$ est une distribution gamma de paramètres $\alpha=\frac{\nu}{2}$, $\beta=2$ et $\gamma=0$; cette distribution a été appelée la *distribution* χ^2 à ν degrés de libertés. De ces propriétés, il découle que la distribution χ^2 de paramètres $k=1$ correspond au carré d'une distribution normale réduite et toute distribution χ^2 de paramètres k peut être considérée comme la somme des carrés de k variables normales réduites indépendantes. En se référant à la fonction de densité de la distribution gamma, lorsque le paramètre α tend vers l'infini la distribution gamma tend vers la distribution normale (Kotz et Johnson, 1982).

Enfin, selon Essenwanger (1986) et Jeevanand et Nair (1998), la **distribution de Pareto** correspond à un mélange de distributions exponentielles pour lesquelles le paramètre de forme λ se présente sous l'aspect d'une distribution gamma.

2.4. Conclusions

Dans ce chapitre, nous avons montré qu'en étudiant le comportement de la queue de la distribution des données à partir des observations les plus élevées, il est possible d'éluder le problème du mélange des distributions, en estimant les paramètres à partir des queues des distributions et en ne considérant que la partie droite ou gauche du mélange. Il est dès lors envisageable de déterminer des limites à partir desquelles des valeurs peuvent être jugées comme aberrantes.

Nous avons cherché à caractériser les principales distributions dissymétriques, en particulier pour la dissymétrie avec un étalement à droite en raison des nombreux travaux élaborés dans le domaine. Deux jeux de données ont été utilisés afin de caractériser les distributions candidates à la détection de valeurs aberrantes. Pour chacune des distributions dissymétriques présentées, une représentation graphique des données, à partir du graphique des quantiles de la distribution concernée, a été présentée afin de connaître l'allure de celles-ci.

Aucune conclusion quant à l'adéquation de l'une ou l'autre distribution pour la partie droite ou gauche ne peut être formulée ici car les exemples, qui ont été présentés, sont basés uniquement sur deux ensembles de données.

Néanmoins, il semble que, pour traiter la partie droite, les distributions de Pareto, exponentielle et éventuellement la log-normale sont intéressantes.

A partir des deux jeux de données, un effet de « zoom » au niveau de la queue de distribution gauche des distributions a été observé avec la distribution de Weibull. Il semble donc envisageable de traiter la partie gauche à partir de la distribution de Weibull.

Ceci peut être mis en relation avec une application liée à des distributions dissymétriques dan le domaine de la performances de réseaux pour laquelle Lu et Sedransk (2002) proposent de modéliser la partie droite des queues de distribution à partir des distributions de type Pareto ou les distributions généralisées de Pareto et la partie gauche avec la distribution de Weibull.

L'application pratique présentée dans la deuxième partie de ce travail nous permettra de proposer une méthodologie permettant de décider le type de distribution dissymétrique à appliquer dans notre étude pour les queues de distribution droite et gauche.

3. CLASSIFICATION AVEC CONTRAINTES SPATIALES

3.1. Introduction

Les techniques utilisées couramment en classification numérique offrent la possibilité de regrouper des individus ou des objets similaires à partir de caractéristiques déterminées. Lors de l'application d'une méthode de classification, il est nécessaire de déterminer les variables qui permettent de classer les observations de manière optimale. Dans le cas de notre étude, étant donné le mélange de distributions très dissymétriques, il n'est pas acceptable d'appliquer directement la classification numérique à partir des données brutes. Par contre, les quantiles estimés par entité communale, déterminés à partir des queues de distributions dissymétriques, semblent être les candidats adéquats pour regrouper des entités communales qui comprennent des observations plus ou moins élevées. En effet, il semble logique que deux communes similaires, présentent des quantiles extrêmes proches.

L'intégration de la contrainte de contiguïté permet de tenir compte, d'une part, des relations de voisinage entre les communes et, d'autre part, de la présence de différents types de sols.

La notion de contrainte spatiale n'est cependant pas exclusivement un problème de classification numérique ; elle peut être abordée à partir d'autres approches. En effet, parmi les méthodes liées à la géostatistique, certaines d'entre elles pourraient être utilisées. Par exemple, il est possible d'utiliser des variogrammes ou des corrélogrammes, qui permettent d'évaluer la dissimilitude entre deux observations afin de créer un indice de proximité entre les communes (Legendre et Legendre, 1998; Colinet, 2003). Nous n'envisageons cependant pas de recourir à ces méthodes dans ce travail.

D'autres techniques, relativement simples et plutôt intuitives, peuvent également être utilisées pour regrouper des zones similaires sans utiliser les techniques de classification numérique. Des indices de similitude peuvent, par exemple, être calculés en examinant la superficie de chaque type de sols par commune. Le classement des surfaces par ordre croissant et par type, pour l'ensemble des communes permet de déterminer les communes les plus représentatives d'un type de sols donné ; ce qui est possible lorsque le géoréférencement est réalisé. Par variable étudiée, le quantile extrême peut être estimé à partir de ces communes représentatives. Le quantile extrême correspond alors à la valeur limite du type de sols et peut être affecté à chacune des communes comprenant ce type de sols. Plusieurs types de sols étant présents dans une même commune, la valeur la plus élevée du quantile estimé peut être retenue et les observations qui présentent des valeurs plus élevées que le quantile estimé sont considérées comme aberrantes. Le choix du nombre de communes à prendre en compte pour l'estimation des quantiles est cependant un problème à résoudre. Dans ce cas de figure, la

contrainte est étudiée en terme quantitatif et permet de déterminer une mesure de similitude.

Comme la classification numérique permet de regrouper des objets semblables de manière relativement aisée, nous avons décidé de l'appliquer dans ce travail. L'objectif de ce chapitre est de présenter les techniques de classification numérique qui permettent de rassembler des entités géographiques similaires à partir de caractéristiques diverses. Plus particulièrement, dans le cas de notre étude, nous cherchons à constituer un référentiel par entité géographique déterminée, telle que des entités communales ou des groupements d'entités communales voisines. Ce référentiel correspondrait aux valeurs des quantiles extrêmes estimés pour chacune des variables, par entité géographique et pour les parties droite et gauche des distributions.

Nous présentons premièrement, de manière très générale, les principales méthodes de classification numérique classique (paragraphe 3.2). Ensuite, nous exposons de manière spécifique les principes sur la contiguïté spatiale (paragraphe 3.3). Des conclusions relatives à l'application de la classification dans notre travail sont présentées au paragraphe 3.4.

3.2. Aperçu des principales méthodes de classification numérique

3.2.1. Introduction

La classification numérique classique permet de constituer de manière objective des groupes d'individus semblables, à partir desquels un certain nombre de caractères ont été observés. Cette méthode est utilisée, la plupart du temps, au niveau de l'étape de l'exploration des données et permet, dans un premier temps, de regrouper les observations sur base de leur ressemblance ou similitude. Des sous-populations sont ainsi différenciées et des informations sur celles-ci peuvent être fournies.

Il existe une profusion d'indicateurs de similitudes et de méthodes de regroupements pouvant être combinées. En effet, quand une mesure de similitude est sélectionnée, basée sur une mesure de distance ou de densité, il faut choisir une méthode de regroupement adéquate afin de mettre en évidence la structure des groupes parmi les données. La diversité des méthodes et procédures est causée par la difficulté de déterminer les distances, de fixer les nombres de groupes et par l'impossibilité d'obtenir une classification optimale selon un critère global donné tel que la minimisation de la variation à l'intérieur des groupes. Par ailleurs, en pratique, il n'est potentiellement pas possible d'examiner toutes les classifications possibles et de retenir celle qui optimise le critère retenu.

Les méthodes de classification ont été classées en deux grandes catégories qui sont les méthodes hiérarchiques et les méthodes non hiérarchiques. Ces deux types de méthodes sont présentés respectivement aux paragraphes 3.2.2 et 3.2.3.

Enfin, les manières de visualiser les groupes obtenus et d'interpréter les résultats sont exposées au paragraphe 3.2.4. Les critères de choix d'une méthode de classification numérique classique sont présentés au paragraphe 3.2.5. A titre d'information, les procédures à appliquer par les logiciels les plus couramment utilisés sont exposées au paragraphe 3.2.6.

3.2.2. Méthodes hiérarchiques

a. Introduction

Au cours du temps, de nombreuses méthodes de classification se sont développées. Parmi celles-ci, la *méthode de classification hiérarchique agglomérative* est la plus connue. Cette méthode consiste à prendre comme point de départ, la partition de *n* objets en *n* classes d'un seul objet. A chaque étape ultérieure, deux classes sont fusionnées pour former une nouvelle classe. Par fusions successives, on passe de *n* groupes d'un objet à un seul groupe de *n* objets, et, à l'issue de la classification, des partitions à *n* groupes, *n*-1 groupes,…, 1 groupe sont obtenues (Palm, 1996). Par cette méthode, à toute étape de l'algorithme, l'affectation d'un objet à un groupe n'est jamais remise en cause.

Les méthodes hiérarchiques agglomératives se distinguent entre elles par la définition de la mesure de similitude entre un objet et un groupe d'objets ou entre deux groupes d'objets et par la stratégie d'agrégation choisie pour comparer les groupes.

Ces méthodes conviennent particulièrement bien lorsque des relations de dominance sont susceptibles d'être mises en évidence entre les observations.

b. Mesures de similitudes

Afin de comparer des observations, il est nécessaire de définir des *mesures de similitude* qui sont utilisées par différentes méthodes de classification numérique. Diverses possibilités existent pour la définition ou le calcul des mesures de similitude.

La *distance euclidienne* correspond à la mesure de ressemblance entre deux objets. C'est la distance la plus couramment utilisée pour les variables quantitatives. Cette distance peut être définie de diverses manières pour deux paires d'objets : distance minimale, maximale, moyenne. La distance euclidienne est l'indicateur le plus fréquemment utilisé en analyse spatiale car elle correspond le mieux à la notion de distance telle que nous l'appréhendons (Beghin, 1979 ; Chandon et Pinson, 1981). De plus, elle convient parfaitement pour les variables quantitatives.

La distance euclidienne correspond à un cas particulier d'une distance plus générale appelée *distance de Minskowski* à partir de laquelle différentes distances peuvent être obtenues (Beghin, 1979; Chandon et Pinson, 1981; Foguenne, 1994; Everitt *et al.*, 2001).

La distance euclidienne possède l'inconvénient de dépendre à la fois de l'unité de mesure et de la variance de chaque variable. De ce fait, la variable ayant la variance la plus élevée prendra une importance prépondérante dans la distance entre les objets. Afin de remédier à ce problème, la standardisation des variables est couramment appliquée et par cette transformation, l'influence exercée par chaque variable sur la mesure de distance est égalisée (Beghin, 1979 ; Chandon et Pinson, 1981). D'autres types de transformations sont également présentés par ces auteurs.

Un autre problème lié à l'utilisation de la distance euclidienne, malgré l'utilisation de transformations, est l'absence de prise en compte des corrélations qui peuvent exister entre les variables. Plus le nombre de variables augmente, plus la distance est élevée même si l'adjonction de variables supplémentaires n'apporte aucune information nouvelle qui permette de distinguer les deux objets (Dagnelie, 1975). Afin d'éviter ce problème, il peut être intéressant d'appliquer une analyse en composantes principales ou de recourir à la distance généralisée de Mahalanobis (Beghin, 1979 ; Lawson et Denison, 2002).

Pour des variables quantitatives résultant de comptages, la *distance du* χ^2 est plus adaptée. Ces distances sont d'autant plus élevées que les objets sont considérés comme différents et d'autant plus petites que les objets sont ressemblants (Palm, 1996).

Le *coefficient de corrélation* entre deux objets, calculée à partir des variables caractérisant ces deux objets, peut également être utilisé pour mesurer la ressemblance entre ceux-ci puisqu'il exprime le degré de similitude entre deux variables (Palm, 1996). Lors de la classification, à chaque nouvelle partition, le carré du rapport de corrélation c'est-à-dire *la valeur de* R^2 peut également être calculée et comparées. Cette valeur correspond également, suite à la décomposition des sommes des carrés des écarts variable par variable, au rapport entre la somme des carrés des écarts globale entre les groupes et la somme des carrés des écarts globale totale.

Selon Everitt *et al.* (2001), les différentes mesures de similitude peuvent, à partir d'un ensemble identique de données, mener à des solutions différentes lors d'un processus de classification numérique. Il n'est donc pas facile de déterminer les mesures de similitude à utiliser. Le choix de la mesure de similitude va être guidé par le type de données à analyser et par l'intuition de l'utilisateur.

c. Algorithmes d'agrégation

Différentes algorithmes d'agrégation ont été développés et sont largement présentés dans Palm (1996) et Everitt *et al.* (2001).

Parmi les méthodes hiérarchiques agglomératives, plusieurs stratrégies ont été développées et parmi celles-ci les plus couramment appliquées sont :

- l'*algorithme de Ward*, utilisé uniquement pour des données quantitatives ou mesures de distance, se base sur le principe de décomposition de la variation entre objets ;
- les méthodes du lien simple[34], du lien moyen et du lien complet qui se basent sur les mesures de distance entre groupes ou sur des mesures de similitude. Le principe de ces méthodes est de fusionner les groupes pour lesquels, respectivement, la distance minimale, moyenne et maximale est la plus faible ;
- la méthode des plus proches voisins[35] qui dérive en partie de la méthode du lien simple dans le but d'éviter l'effet de chaînage des groupes, qui est le principal inconvénient cette méthode (Palm, 1996)
- la méthode du centroïde qui est basée sur la distance entre les centres de gravité[36] des différents groupes.

A l'inverse de la méthode agglomérative, le principe de la *méthode hiérarchique divisive* est de partir de l'ensemble des objets considérés comme appartenant à un seul groupe. A chaque étape, l'algorithme divise un groupe pour en former deux et le processus s'arrête lorsqu'on obtient la partition à *n* groupes.

Un désavantage des méthodes hiérarchiques agglomératives et divisives est que lorsqu'un groupement a été réalisé, il est impossible par la suite d'améliorer ou corriger ce qui a été fait dans les étapes précédentes.

d. Règle d'arrêt pour la détermination du nombre de classes

Suite à l'application de méthodes de classification hiérarchique, il est nécessaire de définir le nombre de groupes à retenir. Les méthodes de détermination du nombre de classes, appelées *règles d'arrêt*, sont basées sur le calcul de critères mesurant la qualité des partitions des observations en un nombre différent de classes. Il existe dans la littérature un grand nombre de règles d'arrêt, particulièrement pour les méthodes hiérarchiques agglomératives, permettant de choisir de façon automatique le nombre de classes à retenir (Baamal, 1994).

---

[34] En anglais : *nearest neighbour, simple linkage*.
[35] En anglais : *nearest-neighbour clustering procedure.*
[36] En anglais : *centroïd linkage*.

Il est important de souligner que l'effectif et le nombre de variables de l'échantillon ont une influence très importante sur la performance des règles d'arrêt. La détermination du nombre de classes est d'autant plus aisée que l'on dispose de grands échantillons décrits par de nombreuses variables.

Une première catégorie de règles d'arrêt regroupe les règles appelées *non inférentielles*, dont le principe est de comparer les valeurs successives du critère de qualité lors de la classification et de choisir la solution la plus adéquate. La plupart de ces méthodes sont informelles et conduisent généralement à mettre en graphique, le critère d'optimisation par rapport au nombre de groupes. Les variations importantes du critère permettent de suggérer un nombre de groupes. Cette approche est cependant très subjective et le choix est réalisé en fonction des attentes de l'utilisateur.

Everitt *et al.* (2001) et Palm (1996) proposent de calculer de manière plus formelle, un *pseudo-F* à chaque étape ; la valeur maximale déterminant le nombre de groupes à prendre en compte. D'autres règles d'arrêt sont disponibles dans la littérature (Baamal, 1994).

Une seconde catégorie de règles d'arrêt, appelées *règles inférentielles* sont basées sur des propriétés asymptotiques et sont difficilement applicables sur des échantillons d'effectifs restreints. Sur le plan pratique, ces règles se révèlent être dans l'ensemble moins performantes que les règles non inférentielles (Baamal, 1994). L'une de ces règles d'arrêt concerne la règle d'arrêt, appelée *pseudo-*T^2, en liaison avec le test de Hotelling rencontré dans le cas multivarié (Palm, 1996).

Le choix d'une règle d'arrêt reste cependant un problème assez complexe et il semble qu'aucune règle ne soit meilleure qu'une autre dans toutes les situations.

3.2.3. Méthodes non hiérarchiques

a. Introduction

Contrairement aux méthodes hiérarchiques, certaines méthodes, appelées *non hiérarchiques*, fournissent directement une partition en un nombre de classes fixé *a priori*. L'objectif est de regrouper n objets en k classes de sorte que les objets d'une classe soient aussi semblables que possible et que les classes soient aussi différentes les unes des autres. Ces méthodes ignorent toute notion de hiérarchie entre les objets puisqu'elles considèrent une seule partition en k groupes à la fois.

Le principe est de répartir les n objets en un nombre de groupes spécifié par l'utilisateur ou déterminé par l'algorithme sur la base de paramètres introduits. Le transfert des objets d'un groupe à un autre vise à optimiser un critère prédéfini mesurant l'homogénéité des groupes, c'est-à-dire la qualité de la partition.

b. Critères d'optimisation

Plusieurs critères d'appréciation de la qualité d'une partition sont disponibles et sont souvent liés à l'équation de l'analyse de la variance multivariée à un critère (Palm, 1996). Les critères les plus souvent utilisés, dans le cas des données continues sont, par exemple, de rechercher, soit à minimiser la trace, soit le déterminant de la somme des carrés résiduelle, ou intra-groupes et donc, dans le même temps, de maximiser la somme des carrés des écarts inter-groupes.

Parmi les critères mentionnés, la minimisation de la trace de la somme des carrés résiduelle est la plus utilisée. Ce critère présente cependant le désavantage d'être dépendant de l'échelle. Ceci est de grande importance pratique, car il est alors nécessaire de faire appel à la standardisation des données pour contourner le problème. Ce critère peut être en quelque sorte être mis en parallèle avec la distance euclidienne présentée pour les méthodes hiérarchiques. Un deuxième problème rencontré avec ce critère est, à nouveau, la structure sphérique des groupes obtenus. Des exemples, présentés par Everitt *et al.* (2001) montrent la forme sphérique des groupes produits alors que la forme de ceux-ci est plutôt de type longitudinale. Ce critère présente également le désavantage de fournir des groupes comprenant un nombre quasi identique d'observations.

Le deuxième critère, qui consiste à considérer non plus la trace mais le déterminant de la somme des carrés résiduelle, est également très utilisé. Il présente le grand avantage de ne pas produire des groupes de forme sphérique. Cependant les groupes formés présentent des formes similaires, c'est-à-dire présentant tous la même orientation et le même degré elliptique. De plus, comme pour le premier critère, les groupes formés comprennent un nombre d'observations relativement égal. Les désavantages cités peuvent entraîner des problèmes lorsque les données se présentent de manière différente, tel est le cas lorsqu'on cherche à classer les observations d'un groupe comprenant de nombreuses observations et un autre petit groupe bien distinct. Enfin, d'autres critères relativement complexes sont disponibles dans la littérature (Everitt *et al.*, 2001).

c. Algorithmes d'optimisation

Comme pour les méthodes non hiérarchiques, il existe différents algorithmes permettant de déterminer la partition en k groupes. Ces algorithmes ont en commun de choisir ou de générer une partition initiale et de transférer ensuite les objets d'un groupe à l'autre, de manière à optimiser le critère retenu ; ces méthodes sont également appelées *méthodes de transfert* ou *de réallocation* (Chandon et Pinson, 1981). Elles se différencient par le choix de la configuration initiale, par le mode de calcul des nouveaux centres des groupes et enfin par les critères d'arrêt du transfert des objets.

Les principales méthodes rencontrées correspondent à la méthode des *k-means*, à l'analyse du mode (Cheng, 1995; Zeng et Starzyk, 2001; Comaniciu et Meer, 2002), à la classification floue (Van Meirvenne *et al.*, 1993; Derrig et Ostaszewsli, 1994; Höppner *et al.*, 2000; Hachama et Bohoua-Nasse, 2003; Smara *et al.*, 2003 ; Ping et Dobermann, 2003) et aux réseaux neuronaux (Immarco, 1992; Davalo et Naïm, 1993; Ripley, 1993, 1994; Smith, 1996; Sarle, 2001; Prévot, 2004). Seule la méthode des *k-means* est développée ci-dessous car elle permet de répondre facilement à la problématique rencontrée dans ce travail. Les autres méthodes nous conduisent vers des domaines d'étude très vastes qui nous mèneraient vers des perspectives trop éloignées de notre sujet ; c'est pourquoi nous ne cherchons pas à les exploiter ici.

- Méthode des *k-means*

Parmi ces méthodes non hiérarchiques, la *méthode des centres mobiles,* appelée très classiquement *méthode des k-means*, est la plus utilisée. Elle fait partie du sous-groupe des *méthodes des centroïdes* car les groupes sont représentés par leur centre de gravité et chaque observation est affectée au groupe dont le centre de gravité lui est le plus proche.

Les étapes essentielles de la classification par cet algorithme sont les suivantes. Dans un premier temps, k objets sont sélectionnés et les coordonnées constituent les centres provisoires. Ces objets sont choisis, soit sur la base de connaissances *a priori* sur les groupes, soit de manière purement arbitraire ; les k premiers objets ou k objets sont sélectionnés au hasard. Chacun des n objets est classé dans le groupe dont il est le plus proche du centre. Les k nouveaux centres de la partition sont ensuite recalculés et une nouvelle partition est effectuée en regroupant les objets dans les k groupes en fonction de leurs distances aux centres des groupes. Ainsi de suite jusqu'à ce que le critère de classification ne s'améliore plus (Palm, 1996). Selon Everitt *et al.* (2001), cette méthode utilise le critère de minimisation de la trace de la somme des carrés résiduels.

L'avantage de cette méthode est sa rapidité d'exécution mais elle présente cependant quelques inconvénients. Un premier inconvénient de la méthode est qu'il est possible obtenir des classes ne contenant aucun objet, c'est-à-dire une partition en moins de k classes. Un second inconvénient est que la partition finale dépend de la partition de départ et un optimum local pourrait alors être atteint (Everitt *et al.*, 2001). Avec un ensemble de données bien structurées, la convergence vers un optimum global peut être attendue. Cependant, les auteurs indiquent que lorsque la convergence est lente et que les groupes sont très hétérogènes pour des partitions initiales différentes, ceci indique que le nombre de groupes est mal choisi et que la classification obtenue n'est pas fiable. Il est donc très important de recommencer plusieurs fois la classification à partir de centres de classes différents afin d'obtenir les résultats les plus stables.

3.2.4. Visualisation des groupes et interprétation des résultats

Suite à l'application de la classification, différents groupes sont obtenus et il est intéressant d'en déterminer les principales caractéristiques. Le résultat des classifications avec les méthodes hiérarchiques est régulièrement représenté par des dendrogrammes qui schématisent les fusions successives permettant de passer de *n* classes à une seule classe qui regroupe tous les objets.

Pour les variables quantitatives, le calcul des moyennes et des écarts-types des différentes variables pour chacun des groupes peut être réalisé afin de localiser les groupes dans l'espace des variables. Ceci permet d'apprécier l'homogénéité des groupes formés. Le calcul des centres de gravité des groupes et des distances entre les centres de gravité permet de se faire une idée de la position relative des différents groupes dans l'espace des variables.

L'application de l'analyse en composantes principales avec la représentation des objets dans les plans factoriels permet également de visualiser les différents groupes. Ce type d'analyse permet de se retrouver dans un espace multifactoriel plus restreint et rend l'inspection visuelle plus accessible. La matrice des diagrammes de dispersion des premières composantes permet de visualiser facilement les résultats obtenus.

Une représentation très intéressante des groupes peut également être réalisée en représentant sous la forme d'une matrice, les graphiques de dispersion pour chacune des combinaisons des variables étudiées deux à deux ; les observations étant remplacées par l'identifiant du groupe obtenu. Cette représentation graphique permet une interprétation des résultats aisée et rapide (Everitt *et al.*, 2001).

Il faut cependant être conscient que, dans une certaine mesure, les choix réalisés par l'utilisateur conditionnent les résultats. Les procédures de classification sont essentiellement des techniques descriptives pour des données multivariées et les solutions fournies par la classification numérique devraient conduire à un réexamen de la matrice des données plutôt qu'à une simple acceptation des groupes obtenus (Everitt *et al.*, 2001).

3.2.5. Choix d'une méthode de classification numérique

Face à la diversité des méthodes de classification numérique apparaît la difficulté de choisir la mesure de similitude et la procédure d'agrégation la plus adéquate. Il est nécessaire de tenir compte des caractéristiques des observations initiales, de la structure des données et des spécificités propres à chaque méthode.

Différentes mesures de similitude calculées à partir d'un même ensemble de données mènent, dans la plupart des cas, à des solutions différentes quand

elles sont utilisées lors d'un processus de classification. Par conséquent, il est très utile d'apprécier les avantages et désavantages de chacune avant de réaliser toute étude. Malheureusement, malgré le nombre élevé d'études comparatives, il n'est pas possible de déterminer de manière absolue le type de mesure à utiliser. Le choix va être réalisé à partir du type de variables étudiées et en fonction de l'expérience de l'utilisateur. Everitt *et al.* (2001) conseillent cependant d'utiliser des coefficients relativement simples ce qui permet une interprétation plus aisée des résultats finaux.

Les méthodes hiérarchiques sont à la base de la majeure partie des études de classification. Elles sont particulièrement disponibles au niveau des logiciels et sont faciles à appliquer. Les choix que doit réaliser l'utilisateur correspondent à la mesure de similitude, la méthode de classification et le nombre de groupes à retenir à partir d'une règle d'arrêt déterminée. En pratique, le problème majeur est qu'aucune méthode bien précise ne peut être recommandée et qu'il est difficile de déterminer le nombre de groupes à garder.

Lorsqu'un nombre de groupes bien précis doit être déterminé, les méthodes non hiérarchiques sont disponibles. Les deux principales techniques classiques de partitionnement présentées, basées sur la minimisation de la trace ou du déterminant de la somme des carrés de l'erreur résiduelle présentent des avantages et des inconvénients qu'il faut prendre en compte lors de l'analyse des données. Il faut noter qu'étant donné la disponibilité, au niveau des logiciels, de la méthode des *k-means* avec minimisation de la trace, beaucoup d'applications sont réalisées à partir de cette méthode. Cependant la seconde méthode présente les avantages de ne pas être sensible au changement d'échelle dans les données observées et de ne pas produire des groupes de forme sphérique.

Les deux types de méthodes de classification hiérarchiques et non hiérarchiques peuvent être combinées en réalisant dans un premier temps le partitionnement afin de dégrossir le problème. Dans un deuxième temps, les méthodes hiérarchiques peuvent être appliquées sur les groupes obtenus.

Au vu de ces propos, il s'avère nécessaire de répéter des analyses sur une même matrice de données à l'aide de mesures de similarités et de techniques de classification différentes. Dans la majeure partie des cas, ces résultats présentent des groupes communs qui constituent l'ossature de la classification. Les divergences entre les résultats portent sur les parties les moins solides de la classification, celles qui requièrent une plus grande prudence quant à l'interprétation.

En dehors de toute contrainte spatiale, l'application de la classification numérique dans notre étude peut être basée sur la mesure de similitude du type de la distance euclidienne car les données, hormis celles liées à la signalétique (occupation du sol, types de culture, etc.), sont de nature quantitative (pHKCl, Mg, etc.). Etant donné le nombre important d'entités

communales à regrouper, l'application de la méthode des *k-means* permettrait de réaliser des groupes de taille relativement importante. Pour chacun de ces groupes, une méthode hiérarchique agglomérative pourrait éventuellement être ensuite utilisée pour obtenir des groupes d'entités présentant des caractéristiques identiques. Comme la contrainte spatiale n'est pas prise en compte, les groupes d'échantillons obtenus seraient alors morcelés et répartis sur l'ensemble du territoire ce qui évidemment n'est pas facilement exploitable au vu de l'objectif poursuivi.

3.2.6. Logiciels statistiques

Pour les techniques de classification hiérarchique basée sur les distances, les procédures à appliquer avec les logiciel SAS et MINITAB sont respectivement PROC CLUSTER et CLUO. La procédure à utiliser pour les techniques de classification hiérarchique basée sur les densités (méthodes des plus proches voisins) est également PROC CLUSTER avec le logiciel SAS ; le logiciel MINITAB n'offre pas cette possibilité. Pour la méthode des *k-means*, les procédures à appliquer sont pour les logiciels SAS et MINITAB respectivement PROC FASTCLUS et KMEAN.

3.3. Classification spatiale

3.3.1. Introduction

Comme nous l'avons cité précédemment, diverses techniques de classification, liées à des domaines spécifiques, ont été développées. Tel est le cas également pour les *techniques de classification avec contraintes*, où l'adhésion au groupe est déterminée partiellement par une information externe telle que la contrainte spatiale.

L'objectif de la classification spatiale est de trouver des régions, généralement dans l'espace à deux dimensions, dans lesquelles l'une ou l'autre caractéristique observée est plus ou moins dominante. La classification numérique appliquée au cas spatial permet de donner une image spatialement contrastée, tel est le cas par exemple, en agriculture, du point de vue de la localisation géographique d'activités agricoles (Lange, 1982).

Le but de ce type d'étude est d'analyser des données qui reflètent un processus spatial sous-jacent en un certain nombre d'endroits afin de déterminer la valeur du processus spatial à tous les endroits du domaine d'intérêt ; ceci en supposant que chaque mesure est observée de manière aléatoire. Cependant, il peut également être intéressant de déterminer les aires où le processus est, par exemple, supérieur à certaines limites prédéfinies ou situées au-dessus de la moyenne (Cressie, 1993).

Des données temporelles peuvent également faire l'objet de classification en se basant sur des observations rassemblées au cours du temps. Complémentairement, des données spatio-temporelles peuvent être

analysées afin de répondre à des questions liées conjointement à l'espace et au temps. De telles analyses sont réalisées dans le domaine de l'épidémiologie avec l'étude de l'évolution de maladies, dans le temps et dans l'espace (Barnett et Turkman, 1993).

La classification spatiale s'est développée de manière diverse au cours du temps. Ce type d'analyse a débuté dans les années 1970, principalement dans le domaine de l'écologie (Legendre et Legendre, 1984a; 1984b). Ensuite, durant les années 1980, un grand intérêt pour la classification spatiale a été suscité par l'analyse d'images, ce qui a finalement mené aux techniques de segmentation d'images et à la reconnaissance d'objet. La modélisation bayesienne hiérarchique des images a été développée dans ce contexte. Des connaissances préalables à propos de la structure spatiale de l'image sont, par cette théorie, incorporées à l'intérieur d'une distribution spatiale *a priori*, tandis que l'erreur de mesure est incluse dans l'erreur aléatoire (Lawson et Denison, 2002).

Durant les années 1980 à 1990, une autre forme de modélisation d'images a été développée qui met l'accent sur la localisation spécifique des caractéristiques de l'image. Ce procédé peut être décrit comme étant la technique de reconnaissance d'objets. Par cette approche, la localisation des caractéristiques locales des images est modélisée de manière spécifique. Une image dite *bruyante* est supposée avoir une distribution sous-jacente d'objets. L'objectif est de reconstruire le groupe d'objets sans le bruit de fond.

En parallèle aux grands développements de méthodologies nouvelles en analyse d'images, les théories ont également progressé dernièrement, pour les modèles bayesiens hiérarchiques, en particulier.

Les autres domaines qui ont vu des avancées considérables en classification spatiale concernent la climatologie, l'environnement, les sciences de la terre, la génétique et l'épidémiologie spatiale. Pour la climatologie, l'environnement et les sciences de la terre, les développements concernent principalement les méthodes géostatistiques en relation avec le krigeage (Lawson et Denison, 2002).

Durant la dernière décade, un énorme effort a été réalisé dans le domaine des données relatives à la santé humaine et principalement dans l'épidémiologie spatiale et les risques environnementaux (Barnett et Turkman, 1994).

3.3.2. Contiguïté spatiale

Dans le cas de classification d'unités spatiales, il peut se justifier de procéder à une classification en imposant au niveau de l'algorithme une contrainte de contiguïté spatiale de manière à obtenir des groupes d'unités géographiquement contiguës (Lange, 1982; Foguenne, 1994; Legendre et Legendre, 1998). Afin de garder le caractère non disjoint des groupes formés, le transfert d'un lieu, d'un groupe à un autre, n'est possible que si ce lieu est périphérique dans le groupe d'origine et il ne peut passer que dans un groupe contigu au groupe d'origine. Les données utilisées s'appuient sur un support géographique que l'on s'efforce de découper pour obtenir un ensemble de zones au contenu homogène. Une classification qui ne prendrait pas en compte la localisation géographique pourrait entraîner une cartographie en forme de mosaïque souvent confuse à interpréter. Lange (1982) par exemple a intégré une contrainte de contiguïté à la méthode des *k-means* dans le but d'identifier la répartition spatiale des activités agricoles en Belgique.

Les contraintes spatiales peuvent être appliquées dans le domaine géographique mais également dans le domaine de la génétique lors d'études de positionnement de parties de chromosomes, la classification devant respecter un certain ordre. Elles sont également utilisées dans le domaine de l'analyse d'images où un groupe, constitués de pixels, forme une aire continue et non fragmentée. Alors que la contrainte spatiale est bi-dimensionnelle, la contrainte à une dimension est rencontrée dans les études de stratigraphie, en archéologie ou géologie (Everitt *et al.*, 2001).

Legendre et Legendre (1998) présentent des applications où la classification avec contrainte a été appliquée, par exemple, afin de délimiter des régions écologiques ou de vérifier que des sites voisins sont similaires au niveau écologiques. Dans un autre contexte, ces applications ont également été réalisées pour identifier des régions de tendances politiques semblables, lors d'études psychologiques ou sociologiques, d'évolution du langage, etc. Des études à partir de données géologiques sont également présentées.

Afin de respecter la cohérence spatiale et de mettre en évidence des groupes de communes présentant les mêmes caractéristiques, si elles possèdent des types de sols majoritairement identiques, l'introduction d'une contrainte de contiguïté géographique dans l'algorithme de classification est une des méthodes les plus appropriées pour respecter strictement les conditions de voisinage. Cette matrice constitue donc un indice de similitude qui conduit à n'envisager la fusion de deux groupes ou communes que si ceux-ci sont voisins, c'est-à-dire s'ils présentent une frontière commune. Ceci implique l'utilisation d'une matrice de contiguïté, reflet de la position relative des communes les unes par rapport aux autres.

a. Matrice de contiguïté

Cette matrice de contiguïté, prise en compte par l'algorithme de classification, est booléenne, symétrique, formée d'éléments diagonaux nuls. La valeur unitaire indique que deux unités, correspondant à la ligne et à la colonne de la matrice, possèdent des frontières naturelles, tels est le cas pour des communes voisines. Dans le cas contraire, la valeur est égale à zéro (Foguenne, 1994).

Lorsque aucune distinction spatiale n'existe réellement (par exemple, une frontière administrative, etc.), des relations spatiales sont utilisées, tel est le cas par exemple dans le graphique de contiguïté appelé *diagramme de Voronoï*. Dans celui-ci, le plan, à l'intérieur duquel les observations doivent être classées, est subdivisé en régions polygonales de telle sorte que le points dans chaque région se trouvent proches des points candidats par rapport à tout autre point. Les points, dont les régions sont voisines, sont dits contigus (Everitt *et al.*, 2001).

Lorsque le nombre de points est particulièrement élevé, le champ spatial, dans lequel se trouvent les observations, peut être divisé en unités utilisant des grilles régulières, par exemple dans le processus d'analyse d'images.

Legendre et Legendre (1998) citent une méthode basée sur la géostatistique pour créer une matrice de similitude en utilisant un variogramme uni ou multivarié comme fonction de pondération spatiale avant de réaliser la classification.

Lorsqu'une matrice de contiguïté appropriée a été définie, les méthodes classiques de classification peuvent être appliquées en modifiant l'algorithme de manière appropriée afin de garantir le regroupement des points qui sont contigus.

b. Utilisation de la matrice de contiguïté lors du processus de classification

Legendre et Legendre (1998) présentent les étapes à suivre pour intégrer la matrice de contiguïté au processus de classification numérique.

1. La matrice de similitude est calculée pour l'ensemble des observations, utilisant toute l'information de type non géographique.
2. La matrice de contiguïté est ensuite créée de la manière la plus adéquate, en utilisant les valeurs 0 ou 1, en fonction du voisinage ou non des communes.
3. Le produit entre les deux matrices est réalisé et est appelé produit de Hadamard. Ce produit correspond au produit élément par élément et la matrice qui en découle contient des valeurs de similitude dans les cellules où la matrice de contiguïté contient la valeur 1 et 0 ailleurs, comme présenté à la figure I.28.

4. Les valeurs de similitude les plus grandes de la matrice résultant de l'étape 3 détermine les paires d'objets ou de groupes à regrouper. Le vecteur des groupes est alors modifié en donnant le même nom de groupe à tous les membres d'un même groupe.
5. La matrice de similitude est recalculée.
6. La matrice de contiguïté est également révisée en fonction en fonction du voisinage des nouveaux groupes formés. A nouveau, l'étape 3 est appliquée et réitérée jusqu'à l'obtention d'un seul groupe comprenant l'ensemble des objets.

Par rapport aux méthodes classiques de classification numérique, la modification se situe donc lors de l'agrégation des observations ou des groupes. A chaque itération, le couple des éléments les plus proches est recherché et la condition de voisinage est testée. Si les éléments sont contigus, ils sont fusionnés et la matrice est modifiée de la même façon que lors du processus de classification numérique classique. La matrice de contiguïté doit également être adaptée. La ligne et la colonne correspondant au premier élément du groupe formé et contenant les éléments de contiguïté entre le groupe formé et les autres éléments sont modifiées de manière à ce qu'elles intègrent les éléments de contiguïté des deux prédécesseurs immédiats. Les lignes et les colonnes correspondant au deuxième élément du groupe sont annulées. Lorsque les éléments les plus proches ne sont pas contigus, la condition de contiguïté est testée sur les couples suivants, constitués des éléments les plus proches. Le processus se poursuit jusqu'à formation d'un seul groupe contenant tous les objets (Lange, 1982).

Legendre (1987) a montré que les résultats de la classification avec contraintes spatiales est relativement stable quelle que soit la mesure de similitude utilisée. Afin d'évaluer la qualité de la classification obtenue, Lange (1982) a comparé les résultats obtenus, d'une part, au moyen des algorithmes avec et sans contrainte de contiguïté et, d'autre part, avec et sans exécution d'une procédure de réallocation. La contrainte de contiguïté a, en effet, le grand avantage d'offrir une clarté supérieure au niveau cartographique. Cependant, l'inconvénient rencontré est une plus grande hétérogénéité à l'intérieur des groupes et une moindre discrimination entre les groupes. L'introduction de la contrainte en début d'algorithme permet d'atténuer ce problème.

Pour les méthodes non hiérarchiques, l'ajout d'une contrainte de contiguïté lors de la classification spatiale présente également le désavantage de retarder le processus de convergence et par conséquent augmente le nombre d'itérations.

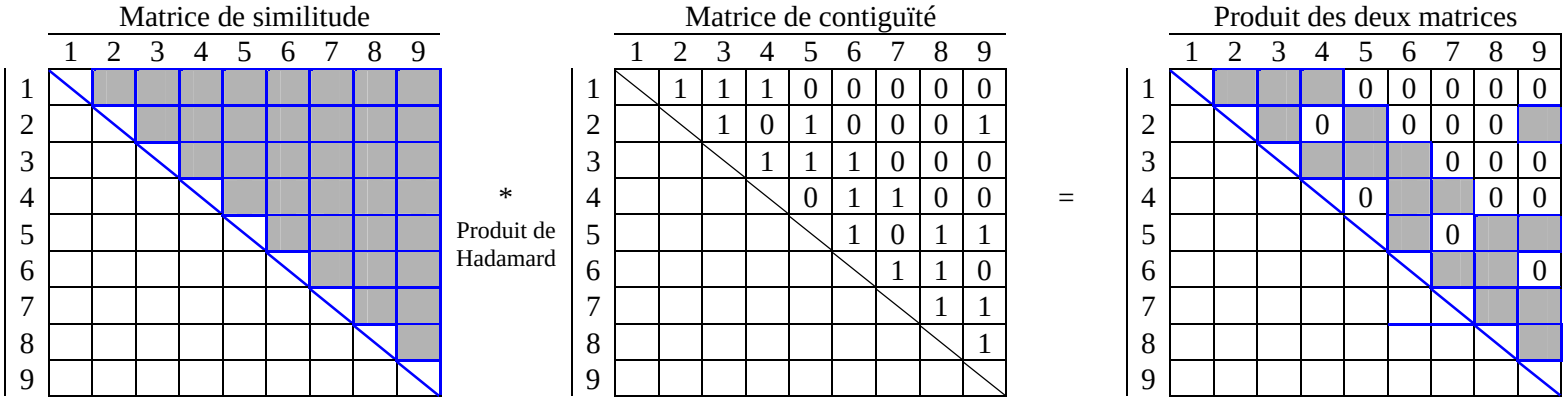


Figure I.28. Procédure de classification sous contraintes spatiales : produit entre la matrice de similitude pour l'ensemble des observations et la matrice de contiguïté (Legendre et Legendre, 1998).

Everitt *et al.* (2001) présentent des exemples d'application de classification avec contraintes de contiguïté par la méthode hiérarchique. Cependant, les méthodes modifiées ne gardent pas forcément les propriétés des méthodes parents telles que le fait d'éviter les inversions. En effet, la contrainte de contiguïté peut mener à des inversions dans le processus de classification, c'est-à-dire à une augmentation de la proximité entre les deux groupes agrégés au cours d'une étape par rapport à la proximité entre les groupes réunis à l'itération précédente.

Selon les auteurs, la méthode du lien complet ne donne pas d'inversions. Murtagh (1995) s'est également penché sur le problème des inversions et a montré que la méthode du lien simple permet d'éviter ce problème également. Les méthodes du centroïde et de Ward peuvent produire des inversions mais elles ont l'avantage de fournir des représentations de zones géographiques plus naturelles ou typiques.

Legendre et Legendre (1998) citent que les résultats, c'est-à-dire les groupes, obtenus par la classification avec contraintes sont moins variables que lorsque aucune contrainte n'est intégrée ; les résultats sans contraintes étant très différents en fonction des méthodes utilisées. En effet, la contrainte spatiale réduit le nombre de solutions possibles et force les différents algorithmes de classification à converger vers des groupes similaires plus grands.

3.3.3. Logiciels statistiques

Au niveau des logiciels disponibles, peu d'entre eux proposent la prise en compte de la contrainte de contiguïté spatiale. Foguenne (1994) a utilisé le logiciel MINITAB (version 9.01) et a programmé une macro qui réalise la classification hiérarchique ascendante avec la mesure du lien moyen.

Everitt *et al.* (2001) citent le progiciel R[37] (version 4.0) destiné aux analyses multidimensionnelles et spatiales. Ce progiciel a été initié suite à des études de type écologique (Legendre et Legendre, 1998). Dans ce progiciel, le programme BIOGEO applique la méthode hiérarchique agglomérative avec différentes mesures de similitude en prenant en compte la contrainte de contiguïté spatiale. La mesure de similitude utilisée est déterminée par l'utilisateur, (par exemple, 0 pour le lien simple et 1 pour le lien complet). Les résultats sont présentés sous forme de cartes. Comme la classification est réalisée à partir d'une matrice de similarité, qui est dans la plupart du temps multivariée, cette méthode est présentée comme une forme de cartographie multivarié (Birks et Gordon, 1985; Casgrain, 2004). La méthode des *k-means* est applicable à partir du programme K-MEANS qui permet de réaliser la classification non hiérarchique sous contrainte spatiale. Ce progiciel est disponible uniquement sous MacOS et serait en cours de développement sous Windows (Casgrain, 2004).

[37] Ce progiciel R est différent de l'environnement de programmation statistique R (Brostaux, 2002).

3.4. Conclusions

Le regroupement d'entités communales voisines, similaires au niveau pédologique, est tout à fait réalisable par l'application de la classification numérique à partir d'une matrice de similitude basée sur les quantiles extrêmes estimés à partir des queues des distributions dissymétriques et par la prise en compte d'une matrice de contiguïté, construite de manière adéquate.

Comme le nombre de communes de la Région wallonne est relativement élevé, il n'est pas concevable de réaliser une classification de type hiérarchique. Par contre, la méthode non hiérarchique des *k-means* offre la possibilité de réaliser rapidement des groupes de taille relativement importante et permet la réallocation des observations à chaque étape de classification.

L'intégration de la contrainte de contiguïté spatiale sous la forme d'une matrice de similitude créée de façon appropriée, offre la possibilité de tenir compte simultanément des relations de voisinage entre les observations et de la présence de différents types de sols. L'inclusion de cette contrainte dans l'algorithme de classification permet de créer des zones non morcelées et de représenter plus facilement les résultats.

Diverses manières de construire cette matrice de similitude sont possibles. Par exemple, il est envisageable de calculer le produit de la matrice de contiguïté des communes et de la matrice de présence/absence des types de sols par commune, ou même de faire intervenir la surface des types de sols, etc. Cependant, malgré qu'il ne soit pas facile de déterminer la mesure de similitude ou la procédure d'agrégation la plus adéquate, il est conseillé d'utiliser des procédures simples afin de faciliter l'interprétation des résultats.

Concrètement, la méthode des *k-means* est appliquée dans ce travail à partir des quantiles extrêmes estimés par entité communale en utilisant comme matrice de contiguïté, une matrice *communes-types de sols* basée sur le voisinage des communes et la présence ou non de types de sols en commun. Nous proposons de tester le progiciel R qui permet de réaliser la classification sous contraintes spatiales.

Par la suite, des quantiles extrêmes sont estimés par groupement de communes en considérant la distribution choisie initialement. Les quantiles estimés servent alors à fixer les limites au-dessus desquelles les valeurs extrêmes sont consid érées comme aberrantes.

II. DEUXIEME PARTIE – METHODOLOGIE ET APPLICATIONS

4. DEMARCHE METHODOLOGIQUE

4.1. Introduction

Suite aux considérations théoriques exposées au cours de la première partie de ce travail et sur base des différentes pistes proposées, nous présentons, dans cette deuxième partie, la démarche suivie pour la recherche d'une procédure de détection de valeurs aberrantes afin de répondre à l'objectif du travail. Celle-ci consiste à appliquer successivement différentes méthodes d'analyse (choix d'une distribution, étude de la troncature, etc.) au niveau des queues des distributions. Le but du chapitre 4 est de présenter le « matériel » (paragraphe 4.2), la méthodologie développée pour aboutir à la procédure finale (paragraphe 4.3) ainsi que la manière de valider la méthode de détection de valeurs aberrantes (paragraphe 4.4).

Au terme de cette partie, un tableau récapitulatif présente, de manière synthétique, la succession des différentes méthodes appliquées lors de la procédure d'établissement de limites de détection de valeurs aberrantes (tableau II.9).

4.2. Matériel

Cette procédure de détection de valeurs aberrantes est développée à partir d'un sous-ensemble de la base de données de *RéQuaSud*, limité aux terres de culture des entités communales de la région agricole du Condroz. Les données disponibles sont relatives aux échantillons de sol récoltés de 1994 à 2003.

La région agricole du Condroz[1] (Anonyme, 1951; 1975) a été sélectionnée pour sa diversité en terme de types de sols au sein de chaque entité communale et donc la difficulté de détecter des valeurs aberrantes liée aux mélanges de distributions. A titre illustratif, la figure de l'annexe 1 présente les différentes régions agricoles de la région wallonne. Notons que certaines entités communales contiennent de nombreux types de sols différents, par exemple, les entités de Gerpinnes ou Walcourt en contiennent 19. D'autres régions agricoles, telles que la Région limoneuse ou l'Ardenne, présentent moins de types de sols différents par entité.

Comme cité ci-dessus, ce sous-ensemble de données comprend uniquement les terres de culture. En effet, les échantillons de sols relatifs aux prairies permanentes ou temporaires ne sont pas retenus dans cette étude car des valeurs tout à fait incohérentes peuvent apparaître à cause de la présence d'échantillons de sol prélevés là où se trouvaient antérieurement des excréments du bétail.

[1] Les limites des régions agricoles ont été inscrites dans la loi. Quatre arrêtés royaux les concernent : 1° : 24 février 1951, 2° : 15 juillet 1953, 3° : 8 mars 1968, 4° : 15 février 1974.

Dans la région agricole du Condroz, nous observons que 67 communes[2] chevauchent ou font partie intégrante de la région agricole du Condroz. Dans ce travail, nous avons pris en considération les entités communales dont la moitié au moins de la surface totale se trouve en région condruzienne.

Le sous-ensemble comprend donc 28.809 échantillons de sols pour 92 communes[3] différentes du Condroz, ce qui correspond finalement à 44 entités communales. Le tableau de l'annexe 2 présente le nombre de données par commune[2]. Ce nombre d'observations varie donc de 0 à 2230, ce qui n'est pas sans poser problème lors des ajustements, étant donné le nombre insuffisant d'observations pour certaines communes. Nous en parlons dans la suite de ce travail.

Il faut signaler que pour 2 entités communales (Farciennes et Saint-Nicolas), aucun échantillon de sol n'a été prélevé car elles sont situées dans des zones non agricoles (zone urbaine) et ne font dès lors l'objet d'aucune observation.

Lors de cette étude, les variables (pHKCL, Ca, etc.) sont étudiées indépendamment les unes des autres. En effet, le caractère dissymétrique des données, la présence de mélanges de distributions et la difficulté de mise en oeuvre des études multivariées nous ont poussés à étudier les variables d'une manière univariée.

Parmi les variables, seule une partie d'entre elles fait l'objet d'une étude approfondie. Les variables sélectionnées sont celles qui paraissent les plus intéressantes par rapport aux objectifs du travail. Elles sont choisies en fonction de certains critères qui ne sont pas forcément remplis en même temps ; ces critères sont les suivants.

1° Pour le premier critère, dans la mesure du possible, les variables doivent avoir un caractère dissymétrique. De plus, il est intéressant de retenir des variables qui présentent des degrés de dissymétrie différents. Le tableau II.1 présente les coefficients de symétrie b_1 pour les variables de la base de données de *RéQuaSud* calculés pour l'ensemble des entités communales. Pour rappel, b_1 est nul lorsque la distribution est symétrique, il est positif lorsque la dissymétrie est gauche[4] et négatif lorsque la dissymétrie est droite[5]. On observe à partir de ce tableau que la variable pHKCl est légèrement dissymétrique à droite. Les variables C, P, K et Mg sont

[2] Anciennes communes avant la fusion des communes de 1975.

[3] Dans la base de données SOLS de *RéQuaSud*, les laboratoires référencent les échantillons de sol à partir du nom de la localité et de son code postal (anciennes communes). Ces échantillons de sols se rapportent à 92 localités différentes comprises dans les 44 entités communales sélectionnées (communes après fusion). Signalons que chaque entité communale, de code INS donné, comprend plusieurs *localités* présentant soit le même code postal que l'entité communale, soit un code postal différent.

[4] La queue de la distribution est alors très étalée vers la droite.

[5] Inversement, la queue de la distribution est alors très étalée vers la gauche.

dissymétriques à gauche tandis que le Ca est particulièrement dissymétrique à gauche ; les histogrammes de fréquence relative et les graphiques normaux de l'annexe 3 illustrent cette tendance pour les variables pHKCl, C et Ca qui représentent les différents types de dissymétrie rencontrés.

Tableau II.1. Coefficients de symétrie pour les variables de la base de données de RéQuaSud.

Variables	b_1
pHKCl	-0.62
C	2,73
P	2,13
K	2,55
Mg	3,31
Ca	13,03

2° Le deuxième critère concerne le niveau des variables qui doit être très contrasté par rapport aux régions voisines. En effet, au niveau de la méthodologie suivie pour la mise en place de la procédure de fixation de limites, une phase d'évaluation du taux de détection de valeurs aberrantes d'autres régions agricoles est prévue.

Celle-ci consiste à comparer des observations issues d'entités communales ne faisant pas partie du Condroz aux quantiles extrêmes estimés considérés comme valeurs limites. Si les observations de ces régions limitrophes sont d'un ordre de grandeur relativement proche de celles rencontrées dans la région du Condroz, il est difficile de les détecter, quelle que soit la méthode mise en oeuvre. Par contre, si les observations sont d'un ordre de grandeur relativement différent, ces observations peuvent alors être mises en évidence.

Les observations issues d'autres régions agricoles correspondent, comme nous l'avons présenté dans la première partie de ce travail (paragraphe 1.2.2), à des contaminants qui, par définition, sont des observations issues d'une autre distribution. Rappelons que les contaminants qui ne sont pas détectables sont gênants car ils polluent la base de données et ils modifient la valeur moyenne fournie pour une entité communale.

Les contaminants peuvent provenir d'entités communales voisines du Condroz appartenant par exemple à la région agricole de la Famenne ou d'entités communales plus éloignées issues de l'Ardenne. Nous proposons d'appeler les observations issues d'entités communales voisines à la région condruzienne les *contaminants proches* tandis que les observations issues de l'Ardenne sont considérées comme des *contaminants éloignés*.

Les *contaminants proches* peuvent provenir d'une mauvaise attribution du code postal lorsque le lieu de la parcelle échantillonnée est différent du lieu où réside l'exploitant (cas typique lorsque l'exploitation se trouve sur deux régions différentes ou sur deux communes différentes). Le code postal de

l'exploitant est encodé alors que c'est celui de la commune où se situe la parcelle qui doit être retenu. Les *contaminants éloignés* apparaissent plus vraisemblablement lors d'une erreur d'encodage du code postal. Pour les deux types de contaminants, la distinction est donc liée au code postal.

Les contaminants sont considérés comme des valeurs aberrantes lorsqu'ils sont identifiés lors de l'application d'un test de détection adapté au modèle de probabilité, c'est-à-dire qu'ils apparaissent douteux dans le contexte d'un modèle de probabilité connu.

Les statistiques descriptives du tableau de l'annexe 4, extrait de Colinet *et al.* (2005), nous permettent d'évaluer les différences de niveaux de teneurs des variables C et Ca d'une région agricole à l'autre. A partir de ces tableaux, on observe que la Famenne et l'Ardenne constituent des régions agricoles pour lesquelles les variables sont relativement contrastées par rapport à la région condruzienne. Pour C, ces différences sont observées pour la partie droite de la distribution tandis que pour Ca, c'est la partie gauche qui est concernée.

Dès lors, en combinant les deux critères de dissymétrie et de contraste entre régions, nous avons choisi d'étudier, de manière approfondie, les variables qui correspondent au carbone organique total (C - g/100g de terre sèche) et au calcium échangeable (Ca - mg/100g de terre sèche).

Concernant les queues de distributions droite et gauche, nous étudions ces deux parties séparément tout en suivant la même démarche, comme présenté dans la partie bibliographique.

La méthode est donc mise en oeuvre à partir des observations des terres de culture de 92 communes, soit 44 entités communales, du Condroz, pour les parties droite et gauche des distributions et pour les variables C et Ca.

4.3. Méthodologie

4.3.1. Introduction

La procédure d'élaboration de limites pour la méthode de détection de valeurs aberrantes est composée d'étapes successives pour lesquelles, dans un premier temps, la contrainte spatiale n'est pas prise en compte (paragraphes 4.3.2 à 4.3.9) tandis qu'ensuite elle est intégrée lors de la classification spatiale (paragraphe 4.3.10). La description des démarches suivies est identique pour les deux variables étudiées. Le tableau récapitulatif II.9 permet de suivre aisément la succession des différentes méthodes appliquées pour l'établissement des limites de détection de valeurs aberrantes ; les références aux différents paragraphes y étant indiquées.

4.3.2. Choix des niveaux de troncature et détermination du nombre minimal d'observations pour les ajustements[6]

Comme cité dans la partie bibliographique, l'ajustement des distributions et l'estimation des paramètres dépendent du nombre d'observations prises en compte dans la queue de la distribution droite ou gauche. Il est donc nécessaire de tester différents niveaux de troncature. Par définition, dans ce travail, nous exprimons le niveau de troncature à droite ou à gauche comme étant le pourcentage d'observations qui se trouvent à droite ou à gauche. Lors de l'étude, nous prenons la queue de distribution à droite ou à gauche qui correspond à un niveau de troncature représentant *x* % des données à droite ou à gauche.

Les différents niveaux sont déterminés à partir des critères suivants.

1° Le choix des différents niveaux de troncature est défini par rapport à la problématique des mélanges et au nombre d'échantillons de sols qui pourraient appartenir à un mélange. En effet, on peut évaluer qu'il est possible d'avoir, au niveau de la queue de la distribution, au maximum, 30% des données situées à droite ou à gauche qui soient issues d'un mélange. Cependant, d'une entité communale à l'autre, le mélange de distribution est différent et peut conduire à des taux de troncature différents tels que 5, 10, 20%. Notons que Goegebeur *et al.* (2002) et Beirlant *et al.* (2004) ont montré, à partir de l'exemple relatif au calcium (deuxième jeu de données présenté dans la première partie du travail), qu'en prenant 402 observations (sur 1.505=26,7%), le paramètre de la distribution de Pareto (estimateur de Hill) est optimal tout en étant stable pour un nombre de valeurs extrêmes *k* variant de 250 à 500 (17% à 33%).

[6] Voir point 1 du tableau II.9.

Il faut cependant tenir compte du concept suivant. Au fur et à mesure qu'on se rapproche de la queue de la distribution (soit 5 à 10%), on se rapproche d'une partie de la distribution non contaminée par le mélange, à droite ou à gauche. Le biais lié au mélange de distributions est donc moins important, cependant, moins de données sont prises en compte pour réaliser l'ajustement et l'estimation des paramètres est moins précise.

Par contre, plus la troncature est augmentée (soit 20 à 30%), plus le biais lié au mélange de populations est important car un trop grand nombre de valeurs centrales sont prises en compte. Un compromis entre ces deux notions d'effectif et de biais doit donc être recherché.

Ceci avait été déjà exprimé en d'autres termes au paragraphe 2.3.7.b à propos de la qualité de l'estimateur de Hill et du choix du nombre d'observations à prendre en compte.

2° Les niveaux de troncature à sélectionner se justifient également par rapport à l'effectif minimum nécessaire pour ajuster une distribution. En effet, il est impossible d'ajuster une distribution à partir de deux ou trois données. Nous avons pris l'option de déterminer cet effectif minimum d'un point de vue pratique, en réalisant des tableaux reprenant les effectifs des communes pour différents niveaux de troncature, comme présenté dans l'annexe 2. Ces effectifs sont à considérer tant pour la partie droite que pour la partie gauche. Au vu du manque de données rencontré pour certaines communes, nous proposons dès lors de regrouper *a priori* les observations d'entités communales limitrophes. Ce regroupement est réalisé en deux temps. Premièrement, les communes (avant fusion) présentant le même code INS sont rassemblées ; le code INS correspondant à une entité communale[7]. Ceci permet de regrouper très facilement des communes (avant fusion) voisines présentant des effectifs faibles. La notion de « communes avant fusion » à l'intérieur des entités communales n'est donc pas prise en compte lors de la procédure de fixation des limites de détection des valeurs aberrantes.

Ensuite, le regroupement est réalisé sur base des effectifs et de la contiguïté déterminée à partir d'une matrice de contiguïté (paragraphe 4.3.3) qui permet de définir la similitude spatiale entre les entités communales. L'agrégation est réalisée sur base du niveau de troncature le plus contraignant, c'est-à-dire pour lequel le nombre de données est le plus faible. A partir du tableau de l'annexe 2, on observe qu'un niveau de troncature plus faible que 10% n'est pas réaliste car il nécessiterait de regrouper trop d'entités communales *a priori*.

[7] Les localités portant un nom différent de l'entité communale sont donc directement regroupées par entité communale sur base du code INS.

Dans un premier temps, nous proposons donc d'étudier les niveaux de troncature de 10, 20 et 30% même si, pour certaines entités communales et pour les niveaux de troncature de 10% ou 20%, le nombre d'observations est encore trop faible pour réaliser un ajustement et nécessite un regroupement *a priori*.

4.3.3. Agrégation *a priori* des entités communales à faible nombre d'observations[8]

Parmi les 44 entités communales citées, certaines d'entre elles présentent un nombre d'observations relativement faible ou même nul suite à la troncature, ce qui ne nous donne pas la possibilité de réaliser l'ajustement des données. Il est donc nécessaire de regrouper les entités communales à faible nombre d'observations pour le niveau de troncature le plus contraignant, c'est-à-dire 10%.

Pour ce faire, nous avons créé une matrice de contiguïté qui va permettre de regrouper les entités communales voisines les plus similaires sur base de la présence de types de sols identiques. Cette matrice correspond à une combinaison de la *matrice communes*[9] et de la *matrice types de sols*[10] (paragraphe 3.4) présentées ci-dessous. Elle prend donc en compte le voisinage des entités communales et la présence ou non des types de sols dans la commune.

Création de la matrice *communes*

Une matrice de dimensions *44x44* reprend l'ensemble des entités communales sélectionnées. Pour les entités communales voisines, le chiffre 1 est noté tandis que le chiffre 0 indique les entités communales non voisines. Le tableau II.2 présente une partie de la matrice de contiguïté pour six entités communales situées dans la région condruzienne. Cette matrice est de dimension *6*x*6*.

Tableau II.2. Exemple de matrice de contiguïté pour six entités communales de la Région wallonne (*matrice communes*).

	Hamois	Havelange	Assesse	Gesves	Namur	Ohey
Hamois	1					
Havelange	1	1				
Assesse	1	0	1			
Gesves	1	1	1	1		
Namur	0	0	1	1	1	
Ohey	0	1	0	1	0	1

[8] Voir point 2 du tableau II.9.
[9] Cette matrice est de dimension *pxp*, *p* correspondant au nombre de communes prises en compte.
[10] Nous appelons dans ce travail *matrice types de sols*, la matrice de contiguïté basée uniquement sur les types de sols. Cette matrice est également de dimension *pxp*.

Création de la matrice *types de sols*

Par entité communale, le pourcentage de surface de chaque type de sols est calculé de la manière suivante. Les surfaces de chaque type de sols des terres de culture du Condroz sont estimées pour chaque entité communale étudiée. La surface relative de chaque type de sols est calculée par rapport à la surface totale des terres de cultures. En pratique, pour qu'un type de sols soit pris en considération, il faut qu'il présente une surface minimale dans l'entité communale (par exemple un minimum de 5% de la surface totale de l'entité communale).

Par entité communale voisine, nous proposons de calculer les indices de similarité de la manière suivante. Le pourcentage de type de sols en commun est calculé en sommant le pourcentage de surface relative pour les types de sols que les entités présentent en commun.

L'exemple du tableau II.3 permet d'illustrer ce calcul. Les entités communales 1 et 2 présentent un indice de similarité de 25%. En effet, pour le type de sols T1, le pourcentage minimal de surface relative en commun est de 5% et pour le type T2, il est de 20%. Pour les types de sols que les deux entités communales ont en commun, la surface relative totale est donc de 25%. Les pourcentages de surface relative liés aux types de sols T3 et T4 ne sont pas pris en compte étant donné que ces deux types de sols ne sont pas communs aux deux entités.

Les entités communales 1 et 3 présentent, elles, un indice de similarité de 35% étant donné que pour le type T1, le pourcentage minimal de surface relative commune est de 5% et pour T2, il est de 30%. La somme des deux fait bien 35%. Pour les entités communales 2 et 3, l'indice de similarité est de 30%.

Pour les six entités communales présentées ci-dessus, la matrice type de sols est présentée au tableau II.4.

Tableau II.3. Exemple de calcul de l'indice de similarité entre les entités communales pour la création de la matrice *types de sols*.

	Types de sols	% de surface relati

Tableau II.4. Exemple de matrice de contiguïté pour six entités communales de la Région wallonne (*matrice types de sols*).

	Hamois	Havelange	Assesse	Gesves	Namur	Ohey
Hamois	1					
Havelange	23,46	1				
Assesse	31,56	80,46	1			
Gesves	39,41	60,14	65,67	1		
Namur	80,00	30,54	32,43	51,28	1	
Ohey	67,85	40,34	49,77	53,98	62,43	1

Création de la matrice *communes x types de sols*

Afin d'obtenir la matrice *communes x types de sols*, le produit des matrices communes et types de sols est réalisé (produit de Hadamard - voir chapitre 3, figure I.28). Le tableau II.5 présente les résultats obtenus pour les 6 entités communales.

Tableau II.5. Exemple de matrice de contiguïté pour six entités communales de la Région wallonne (*matrice communes x types de sols*).

	Hamois	Havelange	Assesse	Gesves	Namur	Ohey
Hamois	1					
Havelange	23,46	1				
Assesse	31,56	0	1			
Gesves	39,41	60,14	65,67	1		
Namur	0	0	32,43	51,28	1	
Ohey	0	40,34	0	53,98	0	1

Regroupement *a priori* des entités communales

Le tableau II.6 présente la liste des entités communales qui font l'objet d'un regroupement en raison de leur faible nombre d'observations. Ce regroupement a été réalisé, sur base de la matrice de contiguïté, à partir des coefficients les plus élevés des entités communales voisines à celles-ci.

Plusieurs entités communales ne sont pas regroupées même si leur nombre d'observations est encore faible. L'entité communale de Lobbes ne comprend que 10 observations pour le niveau de troncature de 10% mais elle ne présente aucune entité communale voisine ; nous jugeons ce nombre suffisant pour réaliser les ajustements. Ensuite, les entités communales de Charleroi (regroupé avec Châtelet et Montignies-le-Tilleul) et d'Aiseau-Presles (regroupé avec Farciennes) possèdent 17 observations. Enfin les entités communales de Flémalle et de Sambreville possèdent respectivement 15 et 18 observations.

En partant des 44 entités communales, nous obtenons donc 39 entités communales regroupées *a priori*. Le tableau de l'annexe 5 présente le nombre d'observations par entité communale et par niveau de troncature.

Tableau II.6. Liste des entités communales regroupées *a priori*.

4.3.4. Méthodes d'estimation des paramètres des distributions[11]

Lorsque le nombre minimum d'observations est atteint suite à l'agrégation des entités communales présentant un faible nombre d'observations, l'estimation des paramètres est réalisée. Les paramètres des distributions de Weibull, exponentielle, log-normale et de Pareto sont ajustés en examinant séparément la partie droite et la partie gauche des distributions pour les trois niveaux de troncature ; ceci est réalisé pour chaque commune individuelle mais aussi par groupe de communes et de manière globale comme cela sera détaillé au paragraphe 4.3.6. Afin de comparer les résultats à ce qui se fait couramment, les paramètres sont également ajustés pour la distribution normale.

Les paramètres des distributions sont d'abord estimés à partir des graphiques des quantiles (QQplot), que nous appelons dans la suite du travail, la ***méthode des graphiques des quantiles***. Ceci est réalisé en confrontant graphiquement les quantiles empiriques $\hat{Q}_n(p)$ aux valeurs théoriques fournies par la fonction des quantiles *Q(p)* relative à chaque distribution (paragraphe 2.3.3). L'estimation des paramètres est réalisée suite à l'ajustement de la droite de régression linéaire obtenue à partir du graphique des quantiles. Les paramètres sont calculés à partir des valeurs de l'ordonnée à l'origine et de la pente des droites de régression ainsi obtenues ; ceci est réalisé de manière spécifique pour chacune des distributions (paragraphes 2.3.4 à 2.3.7).

[11] Voir points 3, 10 et 13 du tableau II.9.

Une deuxième estimation plus robuste est ensuite réalisée par un ajustement non linéaire (procédure NLIN de SAS version 9.13) en recherchant la relation entre les $\frac{i}{n+1}$ (qui correspondent au *y*) et les valeurs de la fonction de répartition de la distribution étudiée. Cependant, cette procédure nécessite initialement la désignation d'un ordre de grandeur des paramètres à estimer ; les paramètres obtenus à partir de la méthode des graphiques des quantiles sont ainsi utilisés.

L'ajustement non linéaire est basé sur la méthode des moindres carrés ; le choix de cette méthode est fondé sur la facilité d'utilisation de celle-ci[12]. En effet, il est plus aisé de la mettre en pratique que la méthode du maximum de vraisemblance ou la méthode des moments.

Concernant la méthode des graphiques des quantiles, les graphiques présentés ci-dessous (figures II.1 et II.2) permettent de mettre en évidence la faiblesse de cette méthode d'ajustement qui est très sensible à la présence de valeurs aberrantes ; elle n'est donc pas robuste. Sur ces figures, les valeurs observées sont représentées en bleu tandis que les valeurs estimées à partir des valeurs des paramètres ajustés (par les deux méthodes) sont de couleur noire. Dans notre cas, l'estimation des paramètres par la méthode du graphique des quantiles a donc un intérêt uniquement pour proposer les valeurs des paramètres initiaux nécessaires pour lancer la procédure d'estimation non linéaire.

4.3.5. Evaluation de la qualité des paramètres estimés[13]

Pour les différents niveaux de troncature, la distribution la plus appropriée est choisie, entre autre, en fonction de la qualité d'estimation des paramètres ajustés. Nous proposons d'évaluer cette qualité en procédant de la manière suivante pour chaque commune.

Les valeurs de RMSE sont calculées par distribution et pour chaque niveau de troncature à partir du graphique correspondant à la distribution étudiée. Notons que la valeur de RMSE est liée à l'ordre de grandeur des unités de mesure de la variable étudiée.

Des cartes de la région condruzienne sont réalisées en représentant par différentes couleurs la distribution qui présente le RMSE le plus faible. Cette représentation graphique permet de visualiser la répartition des distributions au sein du Condroz et d'établir s'il existe une certaine homogénéité.

---

[12] L'utilisation d'un facteur de pondération tels que la variance des quantiles ou la valeur des quantiles améliorerait également l'estimation des paramètres.

[13] Voir point 4 du tableau II.9.

Des tableaux qui permettent de distinguer quelles sont les distributions les plus intéressantes en terme de RMSE le plus faible sont présentés. D'une part, par entité communale et par distribution, le nombre de niveaux de troncature (1, 2 ou 3) pour lesquelles le RMSE est le plus faible est exposé. D'autre part, le nombre d'entités communales pour lesquelles les distributions se situent en première position, en deuxième position, etc. (la première position correspondant au RMSE le plus faible) est comptabilité par distribution et un pourcentage global est calculé.

Afin de vérifier si le niveau de troncature influence la valeur de RMSE, des graphiques, représentant en abscisse le niveau de troncature et en ordonnée la médiane du RMSE, sont effectués par distribution. **Idéalement, la distribution la plus intéressante est celle qui présente le RMSE le plus faible et le plus stable quel que soit le niveau de troncature.**

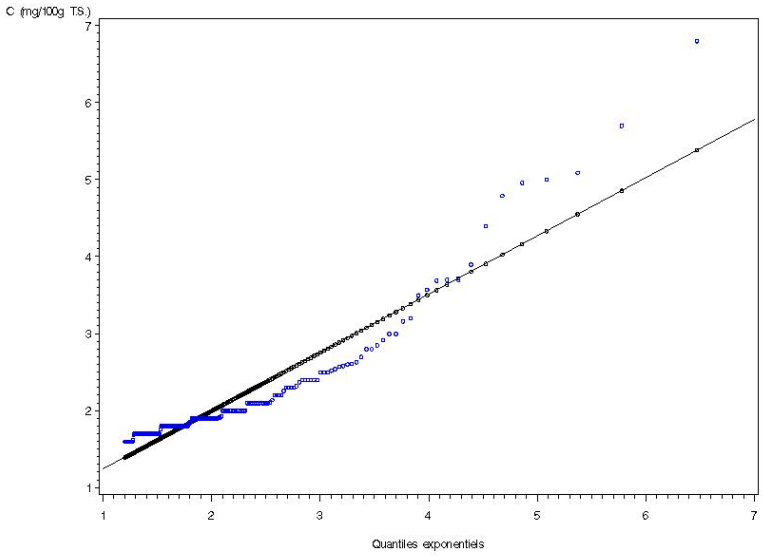


Figure II.1. Ajustement par la méthode des graphiques des quantiles (QQplot) pour l'élément carbone dans le cas de la distribution exponentielle et un niveau de troncature de 30% (entité communale de code INS 92003) – *en noir : valeurs observées, en bleu : valeurs estimées.*

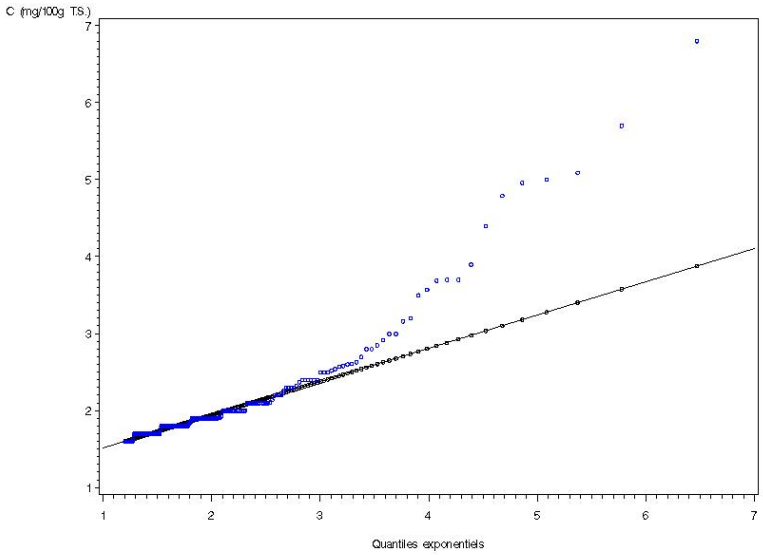


Figure II.2. Ajustement non linéaire basé sur les moindres carrés pour l'élément carbone dans le cas de la distribution exponentielle et un niveau de troncature de 30% (entité communale de code INS 92003) – *en noir : valeurs observées x, en bleu : valeurs estimées* $\hat{x}$.

4.3.6. Estimation des valeurs limites[14]

Par entité communale, les quantiles extrêmes 0,999, pour la partie droite, et 0,001 pour la partie gauche sont estimés par distribution dissymétrique et par niveau de troncature à partir des paramètres estimés dans l'étape exposée ci-dessus. Ces valeurs de quantiles ont été choisies car elles permettent de valider les résultats par rapport à ceux obtenus classiquement par RéQuaSud. De plus, afin de permettre d'évaluer la qualité d'estimation des valeurs limites par les différentes distributions et niveaux de troncature, les quantiles extrêmes 0,99, pour la partie droite, et 0,01 pour la partie gauche sont estimés ; ceci sera exposé au paragraphe 4.3.7.

Afin de vérifier si les deux niveaux de quantiles estimés dépendent du niveau de troncature, un graphique, représentant en abscisse les niveaux de troncature et en ordonnée la médiane des quantiles, est réalisé. Ceci permet de vérifier la sensibilité de la distribution à l'effet de la troncature quel que soit l'ordre de grandeur de la variable étudiée. **La distribution la plus intéressante est celle qui présente des quantiles estimés stables par rapport au niveau de troncature.**

Une deuxième estimation de valeurs limites est réalisée suite à la classification spatiale, lorsque les groupes d'entités communales sont définis lors du regroupement *a posteriori*. Les paramètres des distributions sélectionnées à droite et à gauche et pour le niveau de troncature choisi sont estimés par un nouvel ajustement et les quantiles 0,999 (droite) et 0,001 (gauche) sont estimés par groupe d'entités communales.

Une dernière estimation de valeurs limites (quantiles 0,999 et 0,001) est obtenue à partir de l'ensemble des données de la région agricole du Condroz sans tenir compte des entités communales. Ces quantiles estimés de manière globale vont permettre de comparer les résultats obtenus par entité communale, par groupe d'entités communales et de les valider par rapport à ceux de RéQuaSud.

4.3.7. Evaluation de la qualité d'estimation des valeurs limites[15]

Quantiles estimés (0,99-0,01) en fonction des quantiles observés

Afin de comparer les différentes distributions et les trois niveaux de troncature, nous avons jugé intéressant de confronter les quantiles estimés par rapport aux quantiles observés correspondants.

Une discordance importante entre les quantiles estimés et les quantiles observés correspond à la présence de données suspectes. En effet, idéalement, les quantiles estimés devraient correspondre aux quantiles observés, c'est-à-dire que le biais doit être le plus faible possible. Le

[14] Voir points 5, 11 et 15 du tableau II.9.
[15] Voir point 6 du tableau II.9.

graphique représentant en abscisse les quantiles observés et en ordonnées les quantiles estimés devrait alors se présenter sous la forme d'une droite de pente 1 passant par l'origine.

Cependant, dans notre travail, nous savons que des valeurs aberrantes se trouvent dans le sous-ensemble de données initial et que pour certaines entités communales, il y aura toujours un biais quelle que soit le distribution prise en compte. Si les quantiles estimés étaient quasiment semblables aux quantiles observés, il n'y aurait aucun intérêt à travailler à partir de ceux-ci ; les valeurs observées pourraient alors servir de valeurs limites. Etant donné la présence de valeurs aberrantes, l'incertitude sur les valeurs les plus élevées est trop importante et il n'est pas raisonnable de travailler à partir de celles-ci. Ici, l'intérêt de la démarche est d'observer le comportement des différentes distributions et niveaux de troncature à partir des mêmes observations en terme de biais. L'objectif étant de mettre en évidence la distribution qui décrit le mieux les données, c'est-à-dire pour laquelle le biais est le plus faible.

Comme le quantile 0,99 est très rarement observé à partir des observations, il est nécessaire d'estimer la valeur du quantile observé 0,99, noté x_{obs99}.Ceci est réalisé par interpolation linéaire de la manière suivante :

$$x_{obs99} = x_{obs-1} + \left(\frac{x_{obs+1} - x_{obs-1}}{y_{+1} - y_{-1}} \right) (0{,}99 - y_{-1}),$$

où y_{+1} et y_{-1} correspondent aux quantiles observés qui précèdent et qui suivent le quantile 0,99 et x_{obs+1} et x_{obs-1} sont les valeurs observées correspondantes de la variable estimée.

Par distribution et niveau de troncature, des graphiques représentant en abscisse les quantiles observés et en ordonnée les quantiles estimés permettent de vérifier si les distributions étudiées surestiment ou sous-estiment les valeurs des quantiles. Afin d'évaluer la dispersion des quantiles estimés par rapport aux quantiles observés, l'écart-type *s* entre les quantiles estimés et les quantiles observés pour le niveau 0,99 par distribution et niveau de troncature est calculé de la manière suivante :

$$s = \sqrt{\frac{\sum \left(\hat{Q}_{99} - Q_{obs99} \right)^2}{n}},$$

n étant le nombre de séries d'observations traitées.

D'autres graphiques présentant, par niveau de troncature, la médiane des différences entre quantiles estimés et quantiles observés permet d'identifier la distribution et le niveau de troncature qui estime le mieux les quantiles. Ce graphique permet également de vérifier facilement si la distribution a tendance à surestimer ou à sous-estimer les quantiles. **La distribution la plus intéressante est celle qui présente le biais le plus faible et le plus stable par rapport au niveau de troncature.**

Les quantiles observés 0,999 et 0,001 ne peuvent être calculés car le nombre d'observations par entité communale n'est pas assez élevé. Le calcul du biais entre les quantiles estimés et observés n'est donc pas possible pour ce niveau.

Variabilité des quantiles estimés

Pour évaluer la variabilité des valeurs limites, nous nous basons sur l'intervalle de confiance autour des quantiles estimés[16]. Cet intervalle est obtenu en calculant la différence entre la limite supérieure et la limite inférieure des quantiles estimés. Les quantiles supérieurs et inférieurs sont obtenus par la procédure NLIN de SAS et en travaillant sur base de l'intervalle moyen afin de tenir compte du nombre différent d'observations d'une entité communale à l'autre pour l'estimation.

Des graphiques représentant en abscisse les niveaux de troncature et en ordonnée, la médiane de l'intervalle des quantiles par distribution sont réalisés afin d'évaluer le niveau de variabilité des quantiles et l'influence du niveau de troncature. **La distribution la plus intéressante est celle dont la variabilité des quantiles est la plus faible et la plus stable par rapport au niveau de troncature.**

Afin de connaître la relation entre le nombre d'observations par entité communale et la variabilité des quantiles estimés, des graphiques sont réalisés avec en abscisse le nombre d'observations et en ordonnée l'intervalle entre les limites supérieures et inférieures autour des quantiles 0,99 et 0,999, pour l'ensemble des distributions et pour les trois niveaux de troncature. Ce type de graphique permet de déterminer le nombre d'observations nécessaires pour une estimation fiable des quantiles extrêmes.

[16] Afin d'évaluer la variabilité des quantiles estimés, il aurait été idéal de calculer l'écart-type des quantiles, ceci n'est cependant pas facile à réaliser au niveau pratique.

4.3.8. Sélection de la distribution et du niveau de troncature pour la détection des valeurs aberrantes[17]

En fonction des résultats obtenus, d'une part, en terme de qualité d'estimation des paramètres (paragraphe 4.3.5) et, d'autre part, en terme de qualité d'estimation des valeurs limites (paragraphe 4.3.7), la distribution et le niveau de troncature les plus adéquats sont sélectionnés. La stabilité des quantiles estimés par rapport au niveau de troncature est également prise en compte (paragraphe 4.3.6).

Les critères de sélection sont indiqués en gras dans les paragraphes cités. La distribution et le niveau de troncature retenus répondent le mieux possible à ces critères.

4.3.9. Etude des propriétés de la distribution et du niveau de troncature sélectionné[18]

Nous cherchons ensuite à caractériser la distribution et le niveau de troncature sélectionné par la capacité de détection de valeurs aberrantes qui se traduit par un pourcentage de détection et un rapport d'efficacité de détection.

Evaluation du taux de valeurs aberrantes d'origine

A partir du sous-ensemble de données sur lesquelles les ajustements ont été réalisés, nous dénombrons, par entité communale, les observations supérieures (partie droite) ou inférieures (partie gauche) aux valeurs limites pour chaque distribution et niveau de troncature.

Ces observations correspondent à des valeurs aberrantes dites *valeurs aberrantes d'origine* car elles se trouvent initialement dans la base de données de *RéQuaSud*. Pour les quantiles 0,99 ou 0,01, le pourcentage de valeurs aberrantes d'origine est de l'ordre de 1%, pour les quantiles 0,999 et 0,001, il est de 0.1%.

A titre d'illustration, soit x_1, x_2,..., x_n, les observations ordonnées dans l'ordre croissant d'une entité communale du Condroz pour un niveau de troncature donné (tableau II.7). Si on considère que le quantile estimé 0,99 ($\hat{Q}_{99}$) est la valeur limite de détection et si $\hat{Q}_{99}$ correspond à l'observation x_{n-2}, alors les observations x_{n-1} et x_n sont considérées comme des valeurs aberrantes d'origine.

Ces valeurs aberrantes correspondent à des valeurs extrêmes ou à des erreurs produites lors du processus d'acquisition des données ou à des

[17] Voir point 7 du tableau II.11.
[18] Voir points 8, 12 et 16 du tableau II.9.

contaminants issus d'entités communales voisines ou éloignées qui se révèlent être aberrants. Rappelons ici que l'estimation des quantiles se fait en présence de ces valeurs aberrantes car on ne peut supprimer des valeurs qui présentent un sens pour les laboratoires ayant fourni l'information. Nous fixons donc les limites de détection à partir d'un ensemble de données qui comprend des valeurs aberrantes ; la méthode mise en place devant être suffisamment robuste pour permettre ces ajustements.

Tableau II.7. Observations d'une entité communale du Condroz et limite de détection des valeurs aberrantes.

	Observations	Dénomination
	x_1	
	x_2	
	…	
$\hat{Q}_{99}$	$\boldsymbol{x_{n-2}}$	
	$\boldsymbol{x_{n-1}}$	**valeurs aberrantes**
	$\boldsymbol{x_n}$	**d'origine**

Evaluation du taux de détection de valeurs aberrantes issues d'autres régions agricoles

L'évaluation du taux de détection de valeurs aberrantes issues d'autres régions agricoles est ensuite réalisée pour la distribution et le niveau de troncature retenu.

Nous réalisons cette opération en comparant aux valeurs limites estimées à partir du sous-ensemble de données, des observations réelles et non pas simulées car il est très difficile de donner un sens pratique à ce type de données. En introduisant des données réelles, il est plus facile d'évaluer la qualité de la méthode proposée et d'interpréter concrètement les résultats obtenus.

Les observations à comparer proviennent d'autres populations bien distinctes qui correspondent aux régions agricoles de la Famenne et de l'Ardenne. En effet, les sols de ces régions agricoles présentent des caractéristiques physico-chimiques différentes de la région du Condroz. De plus, elles permettent de prendre en compte le problème des contaminants proches ou éloignés, présenté lors de l'introduction.

A nouveau, afin d'illustrer cette phase d'évaluation du taux de détection de valeurs aberrantes, nous reprenons l'exemple du tableau II.7. Soit y_1, y_2,..., y_n, des contaminants correspondants aux observations ordonnées dans l'ordre croissant d'une entité communale ne faisant pas partie du Condroz (tableau II.8). Lorsque ces contaminants sont comparés aux données de l'entité communale concernée (tableau II.7), les contaminants de valeur supérieure à la valeur de x_{n-2} ($=\hat{Q}_{99}$), sont considérés comme des valeurs aberrantes et sont donc détectables. On parle donc de contaminants

détectés comme des valeurs aberrantes. Par contre, les contaminants $y_1, y_2, \ldots, y_{n-3}$ dont la valeur est inférieure à x_{n-2} ne sont pas identifiés comme des valeurs aberrantes.

Tableau II.8. Comparaison de contaminants au sein des données d'une entité communale du Condroz et détection des valeurs aberrantes.

Observations	Contaminants d'autres régions agricoles	Dénomination
x_1		
x_2		
…	y_1	**contaminants** non détectables
…	y_2	
…	…	
$\boldsymbol{x_{n-2}}$	$\boldsymbol{y_{n-3}}$	
	$\boldsymbol{y_{n-2}}$	contaminants considérés
	$\boldsymbol{y_{n-1}}$	comme des
	$\boldsymbol{y_n}$	**valeurs aberrantes**

Pour les observations de la Famenne de type *contaminant proche*, toutes les observations issues de la base de données SOLS de *RéQuaSud* sont comparées, d'une part, dans les 11 communes du Condroz limitrophes à la Famenne[19] et, d'autre part, dans les autres communes du Condroz non limitrophes à cette région agricole. Cette distinction en classes d'entités communales est réalisée car les communes limitrophes sont beaucoup plus contaminées par des observations de la Famenne et les ajustements réalisés sont « perturbés ». La méthode de détection risque d'être moins performante pour ces dernières communes.

Pour les contaminants éloignés de l'Ardenne, toutes les observations de la base de données SOLS sont également comparées aux valeurs limites des deux classes d'entités communales : communes limitrophes à la Famenne et autres communes.

La figure II.3 présente de manière synthétique la réalisation de cette étape.

- Pour le C

Que ce soit pour la Famenne ou pour l'Ardenne, les teneurs en C sont plus élevées que celles observées en Condroz (annexe 4). Les contaminants sont donc à détecter au niveau de la partie droite de la distribution.

Concernant l'évaluation de la performance dans la partie gauche de cette variable, ces deux régions agricoles possèdent des valeurs plus élevées que

[19] Ces communes du Condroz sont : Beaumont, Walcourt, Florennes, Onhaye, Dinant, Ciney, Hamois, Havelange, Clavier, Ouffet, Anthisnes.

le Condroz. L'étude de la capacité de détection des valeurs aberrantes n'est donc pas possible pour cette partie.

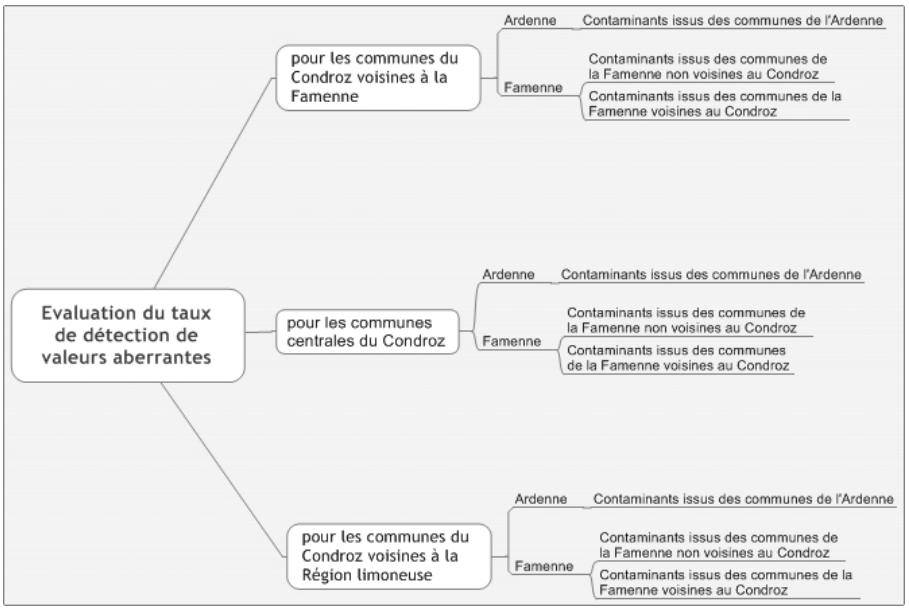


Figure II.3. Représentation de l'étude relative à l'évaluation du taux de détection de valeurs aberrantes issues d'autres régions agricoles, réalisée par distribution et par niveau de troncature pour les deux éléments étudiées.

- Pour le Ca

Le processus de contamination réalisé pour le C est réalisé de la même manière pour le Ca. Cependant, les observations d'Ardenne et de Famenne permettent d'évaluer la méthode de détection au niveau de la partie gauche de la distribution car les teneurs y sont plus faibles que dans la région condruzienne (annexe 4). Pour la partie droite, seules les observations issues de la Région Jurassique sont plus élevées et auraient pu être utilisées comme contaminants mais le nombre de données n'est pas assez élevé et ne nous permet pas de les exploiter.

Chaque fois que des données sont comparées suivant le processus décrit ci-dessus, elles sont testées par rapport aux valeurs limites estimées précédemment.

Comme aucun critère d'évaluation de la qualité de la détection n'a été rencontré dans la littérature, des tableaux présentant le pourcentage de détection des contaminants comme étant des valeurs aberrantes sont réalisés en distinguant les catégories d'entités communales (communes du Condroz voisines à la Famenne, communes centrales du Condroz, communes du Condroz voisines à la Région limoneuse) et en fonction de l'origine des contaminants (issus des entités communales de l'Ardenne, de la Famenne

non voisines au Condroz, de la Famenne voisines au Condroz). De tels tableaux ont été couramment réalisés par Zhang *et al.* (1998).

Les taux de détection de valeurs aberrantes obtenus permettent de les comparer à ceux calculés de manière globale (sans tenir compte des entités communales) et surtout par rapport à ceux de RéQuaSud.

Calcul du rapport d'efficacité[20]

Pour comparer les méthodes de détection de valeurs anormales, Carletti (1988) utilise le *rapport d'efficacité* qui est défini comme étant le rapport du pourcentage de détection de valeurs anormales dans les échantillons avec et sans introduction de valeurs anormales. Les méthodes sont jugées d'autant meilleures qu'elles signalent un rapport d'efficacité élevé.

Dans notre cas, où nous considérons la région agricole du Condroz, le rapport d'efficacité correspond dès lors au rapport entre le pourcentage de contaminants issus d'autres régions agricoles que celle considérée, détectés comme valeurs aberrantes, et le pourcentage de valeurs aberrantes d'origine identifiées dans le sous-ensemble de données de la région agricole du Condroz.

Plusieurs rapports d'efficacité sont calculés dans ce travail dans le but de déterminer quelle entité géographique il faut prendre en compte (entité communale, groupe d'entités communales, ensemble de la région agricoles du Condroz).

Dans le cas de la classification spatiale, ce rapport permet de déterminer les regroupements qui permettent de détecter le plus de valeurs aberrantes issues d'autres régions agricoles mais qui n'identifient pas trop de valeurs aberrantes d'origine (c'est-à-dire les observations du sous-ensemble de données à partir duquel les limites ont été estimées). Le rapport d'efficacité est alors calculé pour chaque regroupement effectué.

4.3.10. Classification spatiale et regroupement *a posteriori* des entités communales[21]

a. Présentation générale de la démarche

Comme présenté au début de ce travail, la contrainte spatiale liée au voisinage des entités communales et à la présence de différents types de sols dans les communes doit être incluse dans notre démarche afin de construire un système cohérent de détection de valeurs aberrantes. La méthode de classification spatiale des entités communales permet de tenir compte de cette contrainte spatiale (paragraphe 3).

Nous appliquons la classification avec contrainte de contiguïté spatiale afin de regrouper des entités communales présentant des quantiles extrêmes

[20] Voir points 8, 12 et 16 du tableau II.9.
[21] Voir point 9 du tableau II.9.

similaires mais également un effectif suffisant pour estimer les quantiles extrêmes de manière fiable suite au regroupement ; cet effectif étant déterminé à partir du graphique représentant en abscisse le nombre d'observations et en ordonnée l'intervalle entre les limites supérieures et inférieures autour des quantiles 0,99 et 0,999 (paragraphe 4.3.7).

La méthode de classification spatiale proposée dans la partie bibliographique, correspond à la méthode de classification non hiérarchique des *k-means* avec contrainte, appliquée à l'aide du logiciel R (Casgrain, 2004). Elle est appliquée séparément pour le carbone et pour le calcium, c'est-à-dire en ne considérant qu'une seule variable à la fois. La variable liée au voisinage des communes est évidemment prise en compte par le logiciel lors de chaque création de groupes d'entités communales.

Cette méthode permet de réaliser le groupement non hiérarchique par minimisation de la variance intragroupe. Pour cette méthode de partitionnement, le nombre *k* de groupes à former est précisé au préalable. Nous utilisons le terme *regroupement* comme étant l'étape de classification menant à la constitution des *k* groupes distincts.

Suite aux regroupements obtenus, le calcul du rapport d'efficacité (Carletti, 1988) permet de comparer les résultats obtenus et le critère de choix du nombre de groupes à prendre en considération est basé sur le rapport d'efficacité le plus élevé.

b. Regroupement *a posteriori* des entités communales sur base de la contrainte de contiguïté

Lors de l'agrégation des entités communales *a priori*, une matrice de contiguïté prenant en compte le voisinage des entités communales et la présence des types de sols (matrice communes-types de sols) est utilisée afin de regrouper les entités de la manière la plus pertinente. La prise en compte de la contrainte spatiale *a priori* a permis, dans un premier temps, de constituer des groupes formés d'entités communales voisines relativement homogènes en terme de types de sols (39 entités communales).

Lors de l'agrégation *a posteriori* des entités communales, nous estimons que la prise en compte de l'information types de sols n'est plus nécessaire.

En effet, l'utilisation des quantiles extrêmes, estimés à partir des queues de distribution, est destinée à éluder le problème de mélanges. Les queues de distributions correspondent dès lors à un ou plusieurs types de sols bien particuliers et non pas à un ensemble homogène de types de sols au niveau de l'entité étudiée. Il n'y aurait donc finalement qu'un ou plusieurs types concernés par le regroupement. Dès lors, l'information type de sols n'est plus vraiment nécessaire car les entités qui présentent des queues de distributions similaires, c'est-à-dire les entités avec les mêmes populations extrêmes (de types de sols) à droite ou à gauche, sont regroupées. La matrice de contiguïté utilisée correspond dès lors à la matrice basée sur le

voisinage des entités communales constituée des indices 0 et 1 (paragraphe 4.3.3).

La matrice de similarité est composée des quantiles extrêmes estimés précédemment pour la distribution dissymétrique et le niveau de troncature sélectionnés à droite et à gauche.

Les résultats obtenus à partir du logiciel R sont présentés sous la forme d'une liste de codes des entités communales retenues dans chaque groupe.

Le nombre total de regroupements réalisés est limité aux groupes qui présentent un effectif suffisant. Le premier regroupement réalisé permet de distinguer 2 groupes, ensuite 3 groupes, etc. jusqu'au nombre de groupes qui permet de rencontrer l'effectif minimum.

4.4. Validation de la méthode de détection

L'évaluation de la méthode proposée est finalement réalisée en comparant le rapport d'efficacité obtenu d'une part, pour la nouvelle procédure de fixation de limites (39 entités communales, *x* groupes d'entités communales et ensemble de la région agricole du Condroz) et, d'autre part, pour la méthode appliquée actuellement au sein de *RéQuaSud*.

Les valeurs limites de RéQuaSud ont été calculées, sur base d'une distribution normale, en estimant l'écart-type de manière robuste sur base des quartiles, séparément à droite et à gauche. Les valeurs limites sont équivalentes à quatre fois cette valeur d'écart-type distinctement pour la partie droite et pour la partie gauche ; ce qui correspond finalement aux quantiles extrêmes de 0,999 et 0,001 également à droite et à gauche.

Le rapport d'efficacité étant obtenu à partir des mêmes observations que celles utilisées précédemment. Cette dernière est basée sur une distribution normale dont les paramètres sont définis pour la partie droite et gauche séparément ; de plus, la contrainte spatiale n'est pas prise en compte. La comparaison du rapport d'efficacité permet de montrer l'intérêt de la nouvelle procédure de fixation des limites.

Ultérieurement, lors de l'introduction de nouvelles observations en routine, dans la base de données, les valeurs les plus élevées (queues droites des distributions) ou les plus faibles (queues gauches des distributions) sont comparées aux valeurs limites déterminées par cette procédure.

Tableau II.9.

5. ETUDE DE LA PARTIE DROITE DES DISTRIBUTIONS (ELEMENT CARBONE)

5.1. Ajustements et évaluation de la qualité des paramètres estimés (par entités communales regroupées *a priori*)

5.1.1. Ajustements par distribution et niveaux de troncature

Pour l'élément carbone, les ajustements ont été réalisés les 5 distributions et les trois niveaux de troncature pour chacune des 39 communes[22]. Les valeurs de RMSE ont été calculées.

Des cartes de la région condruzienne ont été réalisées en indiquant, par différentes couleurs, la distribution qui présente le RMSE le plus faible par commune. Ces cartes ont été réalisées pour les trois niveaux de troncature (figures II.4 à II.6). A partir de ces cartes, on observe des zones géographiques relativement homogènes, principalement pour les distributions exponentielles et de Pareto et pour le niveau de troncature de 30%. Pour les troncatures de 20 et 30%, les entités communales pour lesquelles les distributions normales et de Weibull sont observées concernent des communes de faible effectif. Pour le niveau de troncature de 10%, on observe que la zone intérieure au Condroz est homogène tandis que la zone voisine à d'autres régions agricoles est plus hétérogène ; ce qui serait peut-être lié à la contamination. Cependant, pour la troncature de 10%, le nombre d'observations utilisées lors de l'ajustement est plus faible et l'estimation des paramètres pourrait dès lors être moins robuste. Pour le niveau de troncature de 30%, le nombre plus important de données a tendance à uniformiser les distributions.

Lors d'une étude préliminaire, nous avons pu vérifier que le nombre d'observations intervenant dans l'estimation des paramètres n'a pas d'influence sur le calcul du RMSE quelle que soit la distribution prise en compte.

[22] Dans la suite de ce travail, les termes « communes » et « entités communales » sont utilisés sans distinction pour exprimer les communes après fusion.

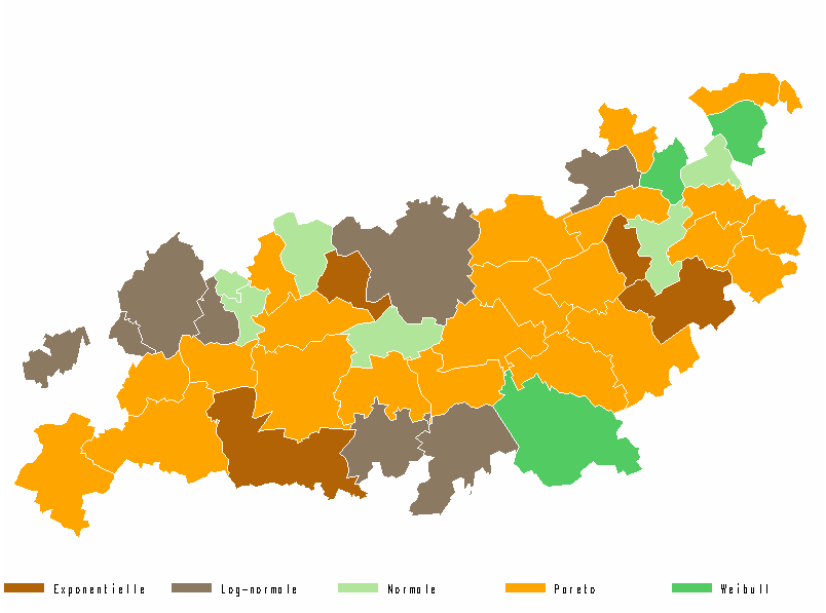


Figure II.4. Elément carbone (partie droite) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 10% et pour la région agricole du Condroz.

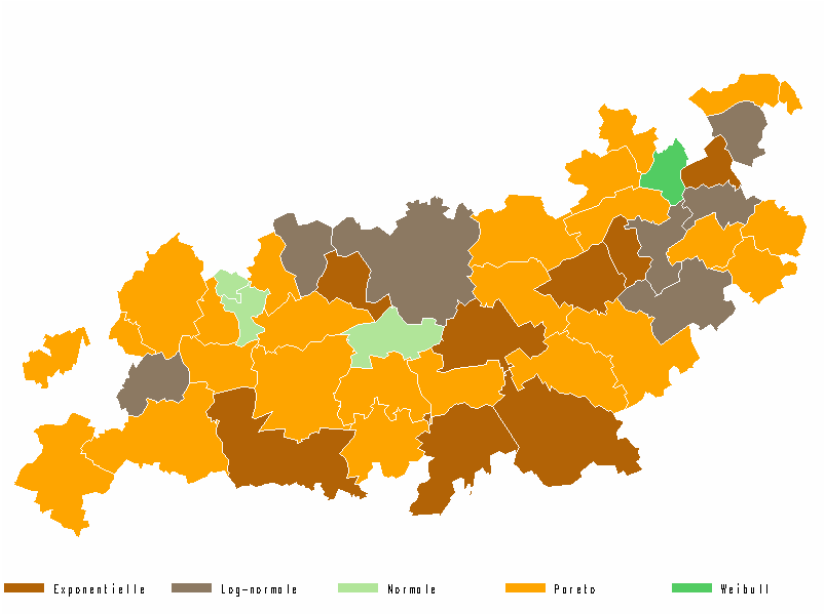


Figure II.5. Elément carbone (partie droite) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 20% et pour la région agricole du Condroz.

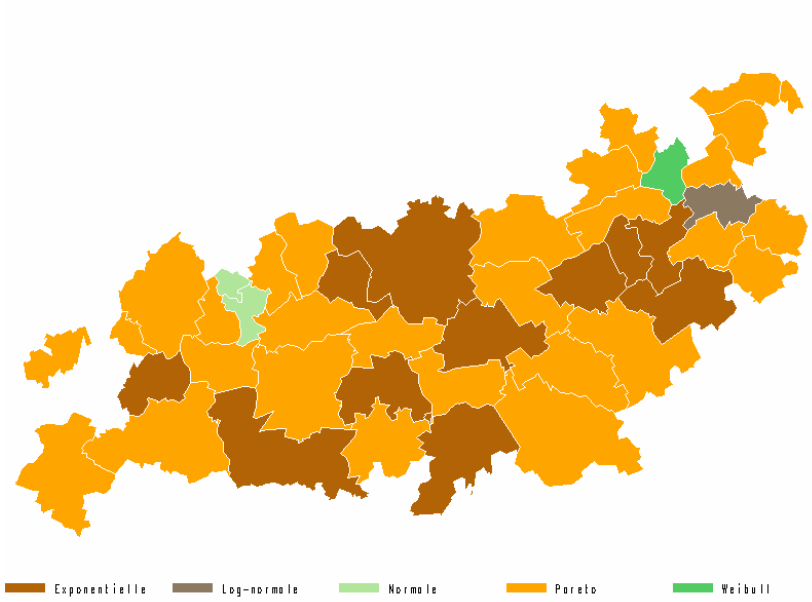


Figure II.6. Elément carbone (partie droite) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 30% et pour la région agricole du Condroz.

Le tableau II.10 présente le nombre d'entités communales pour lesquelles les distributions se situent en première position en terme de RMSE, en deuxième position, etc. ; la première position correspondant au RMSE le plus faible.

La distribution de Pareto se situe en première position dans 56.4% des cas, suivie de la distribution exponentielle dans 54.7%. Ensuite, les distributions log-normale, de Weibull et normale se situent en 3ème, 4ème et 5ème position. On peut observer que lorsque la distribution de Pareto n'est pas en première position, elle se trouve principalement en 5ème position. Par contre, lorsque la distribution exponentielle ne se trouve pas en deuxième position, elle se trouve en première position. Cette dernière serait donc plus 'stable' que la distribution de Pareto.

Tableau II.10. Elément carbone (partie droite) : comptage de la position (rang) des différentes distributions par niveau de troncature en fonction de la valeur de RMSE.

Troncature	Rang	normale	Weibull	log-normale	exponentielle	Pareto
10%						
	1er	5	3	7	3	21
	2ème	3	5	13	18	0
	3ème	0	8	15	13	3
	4ème	11	19	4	4	1
	5ème	20	4	0	1	14
20%						
	1er	2	1	7	8	22
	2ème	1	2	12	23	1
	3ème	0	11	20	4	4
	4ème	15	18	0	4	2
	5ème	21	7	0	0	11
30%						
	1er	1	1	1	12	24
	2ème	1	1	13	23	1
	3ème	2	7	25	2	3
	4ème	16	19	0	2	2
	5ème	19	11	0	0	9
Global						
(en %)	1er	6.8	4.3	12.8	19.7	**56.4**
	2ème	4.3	6.8	32.5	**54.7**	1.7
	3ème	1.7	22.2	**51.3**	16.2	8.5
	4ème	35.9	**47.9**	3.4	8.5	4.3
	5ème	**51.3**	18.8	0.0	0.9	29.1

5.1.2. Etude de l'influence de la troncature sur le RMSE

La représentation des valeurs de RMSE de chacune des entités communales par rapport aux trois niveaux de troncature pour chacune des distributions ne permet pas de distinguer clairement l'effet de la troncature car les observations sont masqués les unes par rapport aux autres par niveau de troncature. Afin d'améliorer la présentation des résultats, nous avons calculé la médiane des RMSE par niveau de troncature et par distribution (figure II.7).

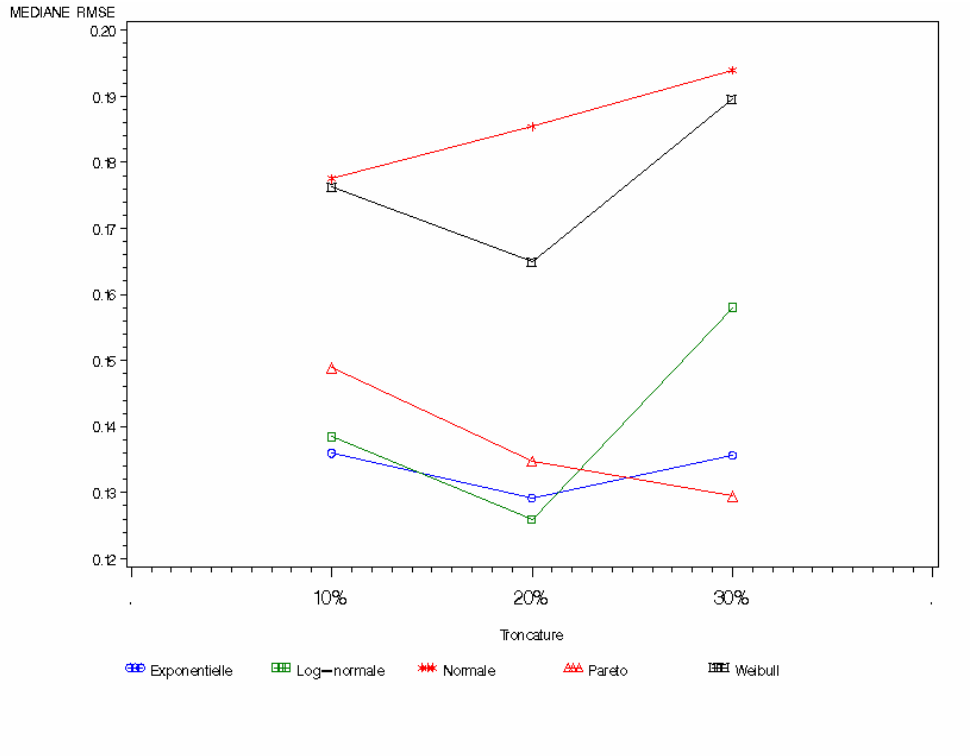


Figure II.7. Elément carbone (partie droite) : étude de l'influence du niveau de troncature sur la qualité des paramètres estimés (exprimée par le RMSE) – trois niveaux de troncature en fonction de la médiane du RMSE calculé pour l'ensemble des entités communales étudiées.

A partir de cette figure II.7, on observe immédiatement que les distributions normale et de Weibull présentent des RMSE très élevés par rapport aux autres distributions ; ce qui constitue une confirmation des résultats présentés au paragraphe 5.1.1.

De plus, même si la distribution de Pareto semble la plus intéressante à partir des valeurs brutes de RMSE, ce graphique indique que la distribution exponentielle est également intéressante.

Idéalement, la distribution à retenir est celle qui présente le RMSE le plus faible et le plus stable quel que soit le niveau de troncature. **Il apparaît donc, à partir de la figure II.7, que la distribution exponentielle présente un niveau très faible de RMSE et les valeurs de RMSE sont les plus stables pour les trois niveaux de troncature.** Ces résultats divergent donc par rapport à ce qui avait été présenté dans l'approche bibliographique dans laquelle la distribution de Pareto était proposée pour traiter la partie droite des distributions (paragraphe 2.4).

La distribution log-normale est la moins stable des distributions présentant un RMSE faible.

5.2. Estimation des valeurs limites et évaluation de la qualité de l'estimation (par entité communale regroupée *a priori*)

5.2.1. Etude de l'influence de la troncature sur l'estimation des quantiles 0,99

Les valeurs des quantiles 0,99 ont été calculées pour l'élément carbone à partir des paramètres ajustés.

Comme lors de l'étude des RMSE, nous présentons la médiane des quantiles estimés (0,99) par niveau de troncature et par distribution (figure II.8). Par cette figure, la classification des distributions en fonction de l'étalement de la queue de distribution à droite est très nette. On observe, en considérant les quantiles les plus faibles aux quantiles les plus élevés, la distribution normale, la distribution de Weibull suivie par la log-normale, l'exponentielle et enfin la Pareto. Ces distributions présentent les queues de distributions des plus faibles aux plus élevées. Ceci indique que le choix de la distribution conditionne la valeur du quantile estimé.

Il faut signaler que plus les quantiles estimés sont faibles, plus la détection de valeurs aberrantes est élevée. En effet, un contaminant d'une teneur en carbone de 3,15 mg/100 g T.S. serait détecté, pour le niveau de troncature de 10%, par les distributions normale et de Weibull mais pas par les autres distributions qui présentent des quantiles supérieurs à 3,15. De même, le contaminant d'une teneur en carbone de 3,25 mg/100 g T.S. ne serait pas détecté quel que soit le niveau de troncature par la distribution de Pareto mais le serait pas les quatre autres distributions.

Idéalement, les quantiles estimés doivent être stables quel que soit le niveau de troncature, c'est-à-dire que le quantile obtenu est identique pour les trois niveaux de troncature. **A partir de la figure II.8, on observe que pour la distribution de Pareto, l'estimation des quantiles est stable.**

En ce qui concerne la distribution exponentielle, les quantiles estimés sont moins stables ; pour les niveaux de troncature de 20 et 30%, les valeurs des quantiles sont plus faibles et plus de contaminants sont détectés comme aberrants.

On peut observer à partir de cette figure que plus la troncature est faible, plus les valeurs des quantiles se rejoignent.

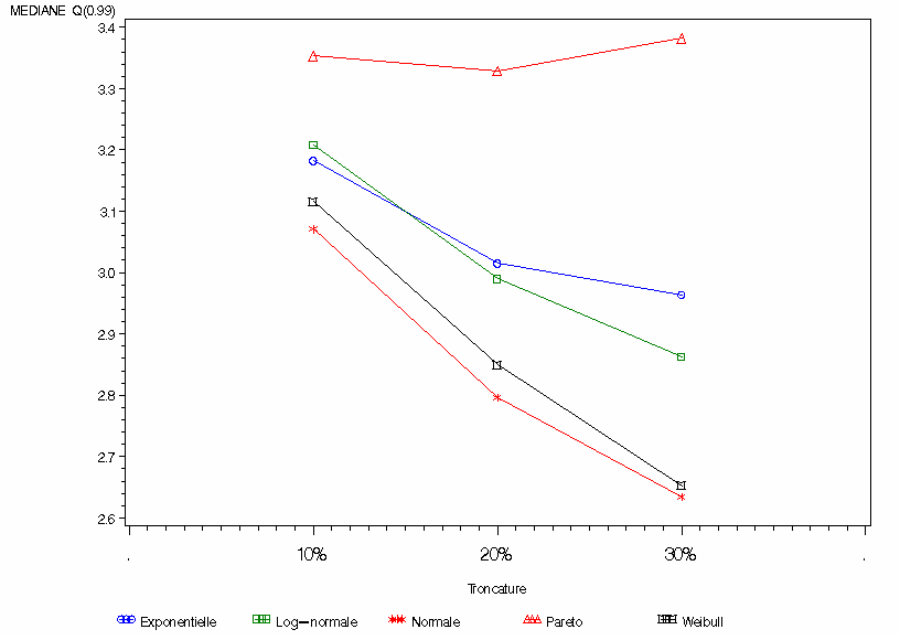


Figure II.8. Elément carbone (partie droite) - Quantiles estimés (0,99) en fonction du niveau de troncature.

5.2.2. Evaluation de la qualité des quantiles estimés 0,99

a. Quantiles estimés par rapport aux quantiles observés

Comme exposé au paragraphe 4.3.5, nous avons réalisé des graphiques des quantiles estimés et des quantiles observés par niveau de troncature pour l'ensemble des distributions.

Les figures II.9 à II.13 présentent les graphiques des quantiles estimés par rapport aux quantiles observés par niveau de troncature, pour chaque distribution. La figure II.9 présente le cas de la distribution normale pour laquelle on rencontre le biais le plus important entre les quantiles estimés et les quantiles observés. Les figures II.10 à II.13 concernent les distributions qui nous intéressent le plus, à savoir les distributions exponentielle et de Pareto pour les niveaux de troncature de 10 et 30%. L'entité communale dont le code NIS est de 52011 n'est pas représentée car les valeurs sont beaucoup trop élevées.

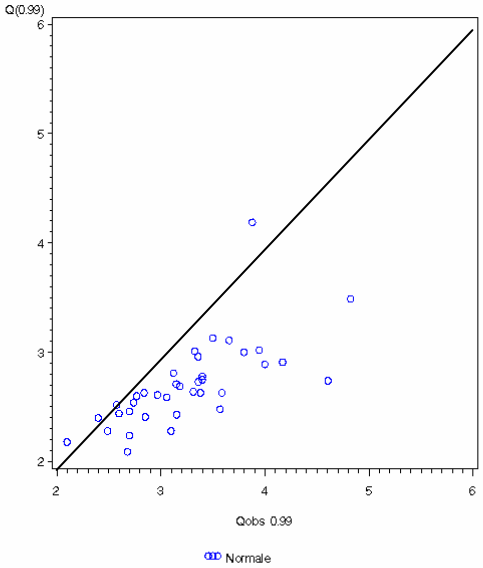


Figure II.9. Elément carbone : quantiles estimés (0,99) à partir de la distribution normale en fonction des quantiles observés pour le niveau de troncature de 30%.

Distribution exponentielle

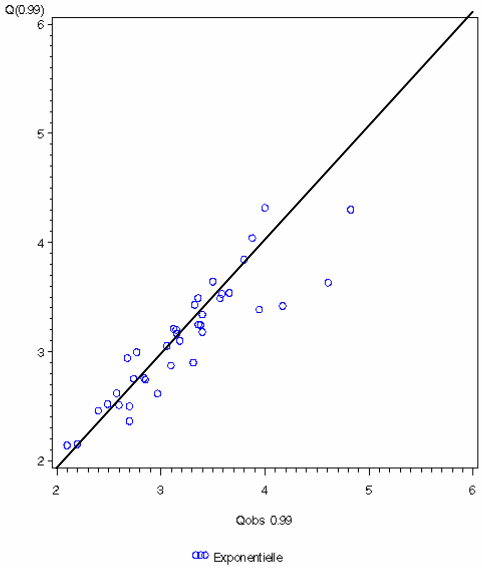


Figure II.10. Elément carbone : quantiles estimés (0,99) à partir de la distribution exponentielle en fonction des quantiles observés - niveau de troncature de 10%.

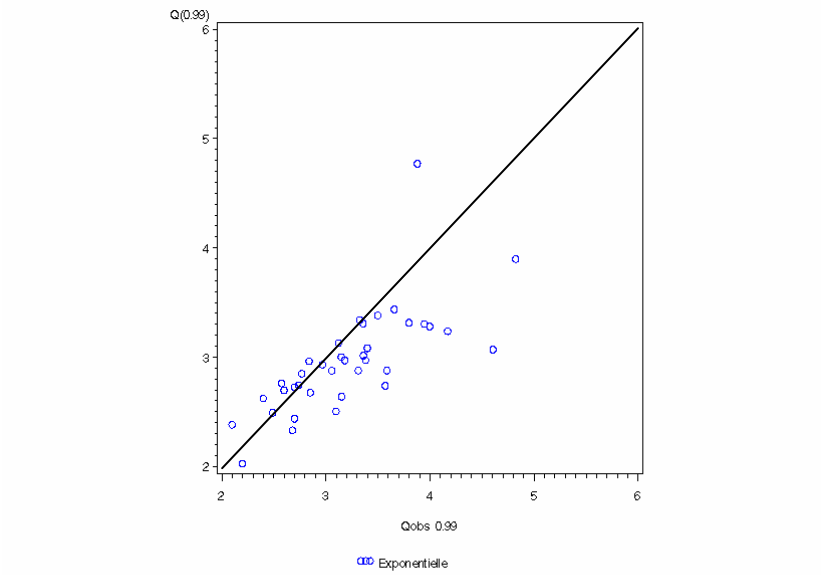


Figure II.11. Elément carbone : quantiles estimés (0,99) à partir de la distribution exponentielle en fonction des quantiles observés - niveau de troncature de 30%.

Distribution de Pareto

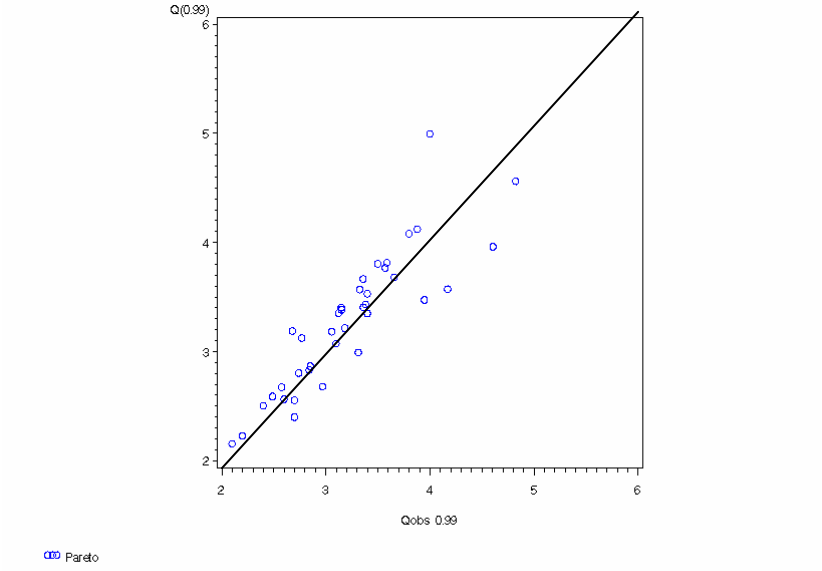


Figure II.12. Elément carbone : quantiles estimés (0,99) à partir de la distribution de Pareto en fonction des quantiles observés - niveau de troncature de 10%.

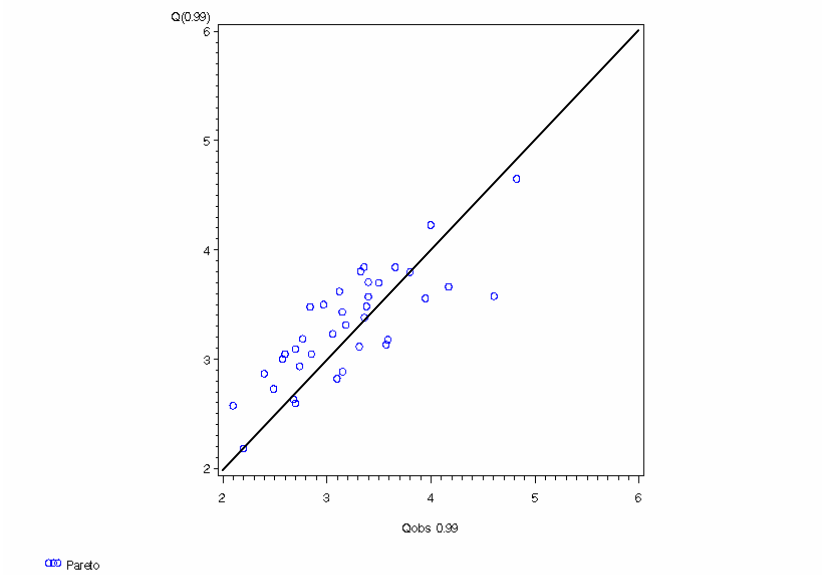


Figure II.13. Elément carbone : quantiles estimés (0,99) à partir de la distribution de Pareto en fonction des quantiles observés - niveau de troncature de 30%.

Afin d'évaluer la dispersion des quantiles estimés par rapport aux quantiles observés, le tableau II.11 présente les valeurs des écarts-types entre les quantiles estimés et les quantiles observés pour le niveau 0,99 par distribution et niveau de troncature[23].

La dispersion la plus faible est rencontrée pour le niveau de troncature de 10% pour la distribution exponentielle, pour 20% pour la distribution log-normale et pour 30% pour la distribution de Pareto.

Globalement, la dispersion la plus faible est obtenue pour la distribution exponentielle, avec le niveau de troncature de 10%.

Un critère qui nous semble particulièrement important au niveau de la qualité de l'estimation des quantiles correspond à la « justesse » de l'estimation des quantiles. Afin de mieux visualiser les différences entre les valeurs observées et les valeurs estimées, c'est-à-dire le biais, la médiane des différences entre les quantiles estimés et les quantiles observés est présentée à la figure II.14.

[23] avec nis 52011.

Tableau II.11. Elément carbone : écarts-types entre les quantiles estimés et observés pour le niveau 0,99 par distribution et niveau de troncature.

Troncature	distribution	Ecart-type
10%	normale	0.329
	weibull	0.309
	log-normale	0.318
	exponentielle	**0.281**
	pareto	0.433
20%	normale	0.584
	weibull	0.513
	log-normale	**0.425**
	exponentielle	0.431
	pareto	0.429
30%	normale	0.842
	weibull	0.781
	log-normale	0.619
	exponentielle	0.611
	pareto	**0.500**

A première vue, le niveau de troncature de 10% permet d'obtenir le biais le plus faible quelle que soit la distribution. Cette figure permet également de montrer que la distribution de Pareto a tendance à surestimer les quantiles tandis que les autres distributions ont tendance à les sous-estimer. Les distributions qui sous-estiment le plus les quantiles étant les distributions normale et de Weibull.

Le biais le plus stable par rapport au niveau de troncature est obtenu pour les distributions exponentielle et de Pareto ; la première sous-estimant les quantiles, la seconde les surestimant.

La distribution log-normale présente également un intérêt en terme de biais ; celui-ci est équivalent aux distributions exponentielles et de Pareto pour les niveaux de troncature de 10 et 20%. Pour la troncature de 30%, le biais est plus élevé. La distribution log-normale présente donc un manque de stabilité de l'estimation des quantiles 0,99 et du biais. De même, la distribution log-normale présentait une valeur de RMSE beaucoup plus élevée pour le niveau de troncature de 30%.

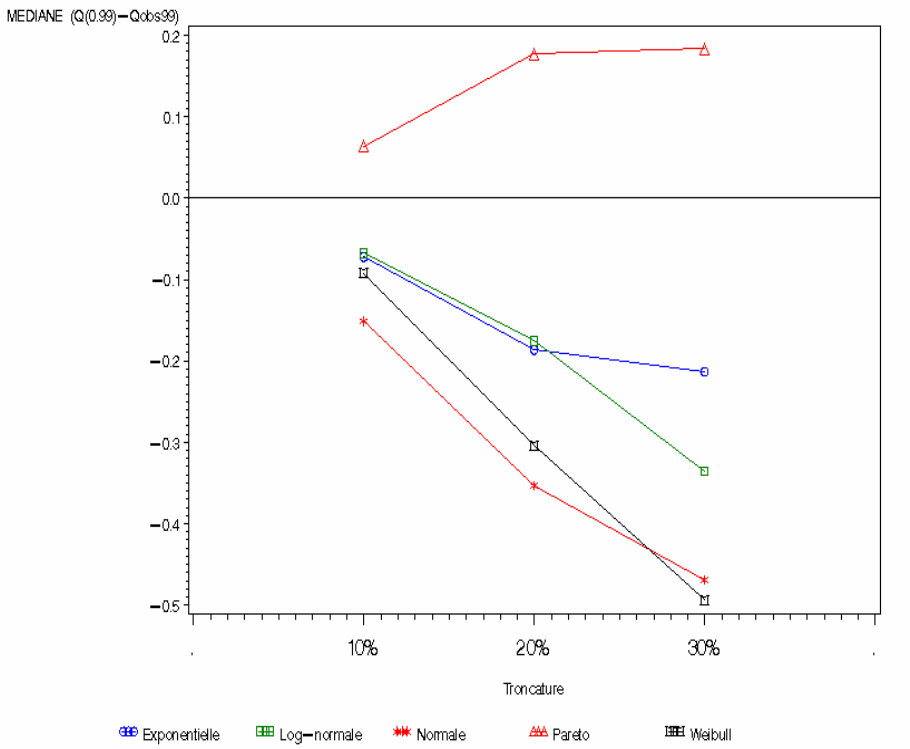


Figure II.14. Elément carbone : médiane des différences entres les quantiles estimés (0,99) et les quantiles observés pour le quantile 0,99, pour les trois niveaux de troncature et pour les cinq distributions étudiées.

b. Etude de la variabilité des quantiles estimés

La figure II.15 présente la valeur de la médiane des différences entre les limites supérieures et inférieures des quantiles estimés en fonction du niveau de troncature et pour les cinq distributions étudiées.

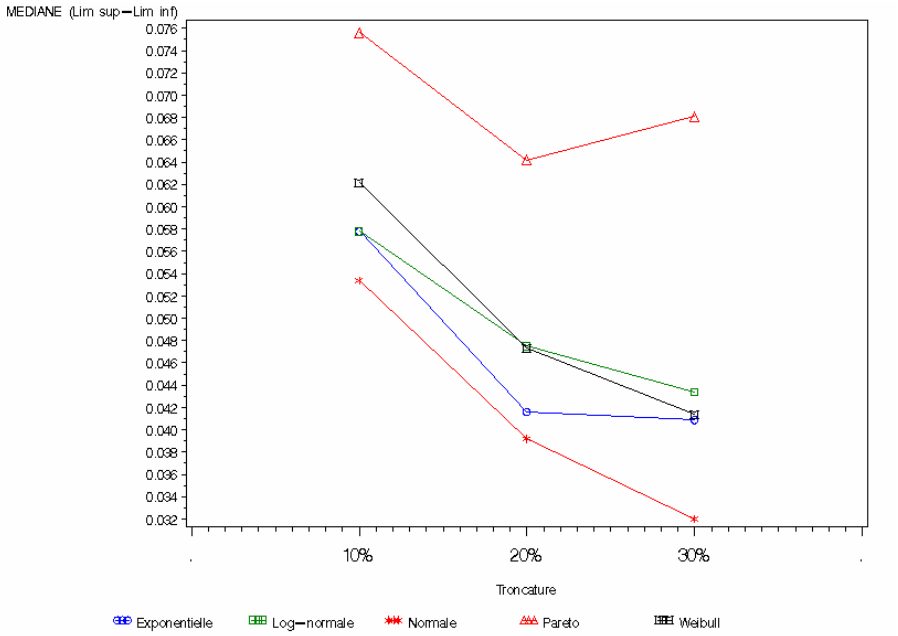


Figure II.15. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,99) pour les trois niveaux de troncature et pour les cinq distributions étudiées.

A partir de cette figure, nous remarquons immédiatement que la **variabilité des quantiles estimés est supérieure pour la distribution de Pareto.** De plus, la **variabilité est plus élevée globalement pour le niveau de troncature de 10%**. Rappelons que c'est pour ce niveau de troncature que le biais est le plus faible.

Ceci signifie donc que, plus le plus le nombre d'observations est faible (troncature 10%) pour l'estimation des paramètres, plus la variabilité est élevée et plus le biais est faible.

Afin de connaître la relation entre le nombre d'observations par entité communale et la variabilité des quantiles estimés, la figure II.16 présente en abscisse le nombre d'observations et en ordonnée l'intervalle entre les limites supérieures et inférieures autour du quantile 0,99, respectivement pour l'ensemble des distributions et pour les trois niveaux de troncature. La figure II.16 correspond à l'ensemble des distributions et des niveaux de troncature.

A partir de ce graphique, il semble qu'un nombre d'observations de 800 à 1000 est nécessaire pour obtenir une estimation fiable des quantiles 0,99, c'est-à-dire des quantiles qui présentent une faible variabilité.

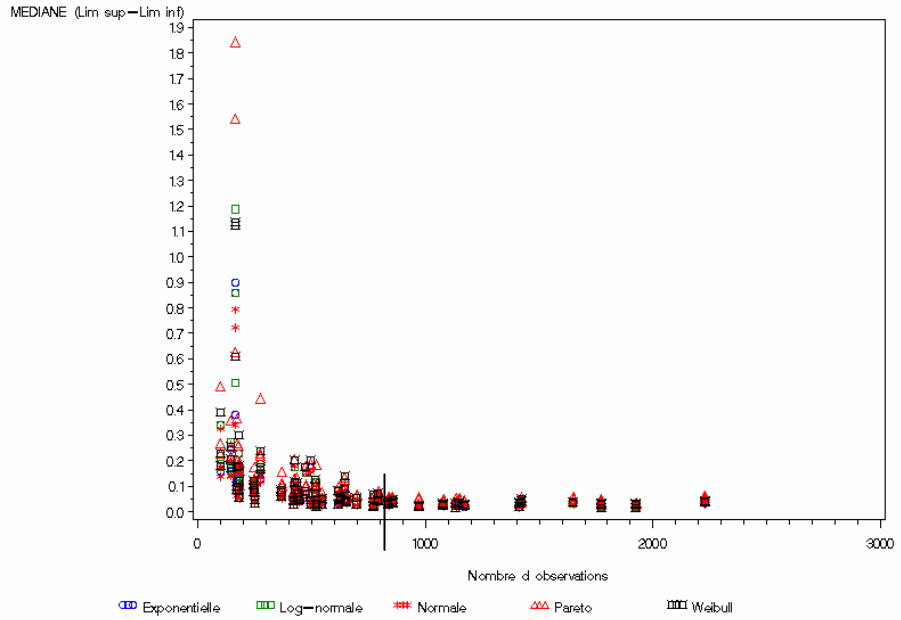


Figure II.16. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,99) en fonction du nombre d'observations pour les cinq distributions étudiées et les trois niveaux de troncature.

Il est également intéressant d'évaluer la relation entre les quantiles estimés et la variabilité de ces quantiles. La figure II.17 présente en abscisse la médiane des quantiles estimés et en ordonnée la médiane de l'intervalle entre les limites supérieures et inférieures.

A partir de cette figure, on observe que plus les quantiles estimés sont élevés, plus l'intervalle autour de ces quantiles est élevé quelle que soit la distribution ou le niveau de troncature.

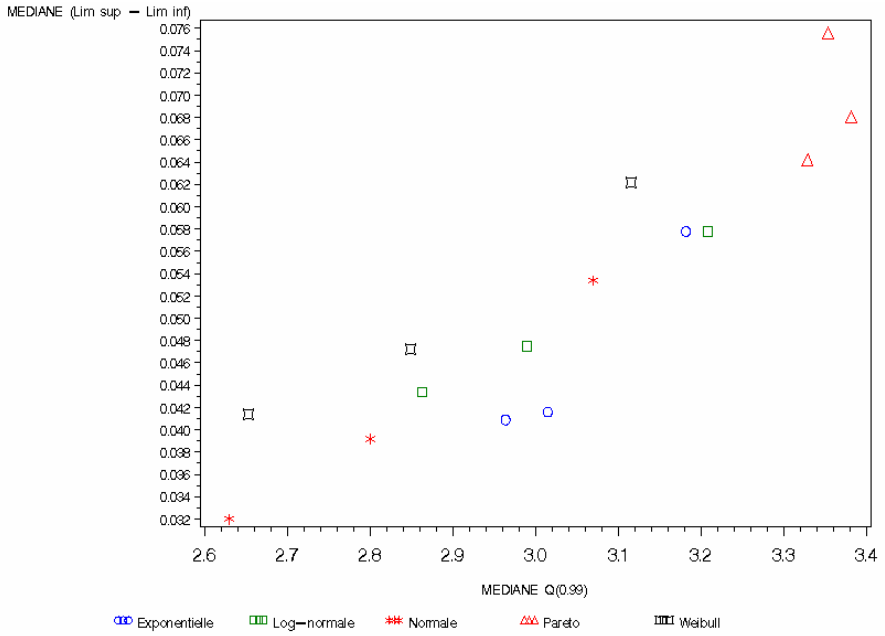


Figure II.17. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,99) en fonction de la médiane des quantiles 0,99 pour les cinq distributions et les trois niveaux de troncature.

5.2.3. Etude de l'influence de la troncature sur l'estimation des quantiles 0,999

Comme lors de l'étude des quantiles 0,99, nous présentons la médiane des quantiles estimés (0,999) par niveau de troncature et par distribution (figure II.18).

A partir de cette figure, on observe immédiatement que les quantiles les plus élevés sont obtenus pour la distribution de Pareto. Des contaminants, par exemple d'une teneur en carbone de 5.0 mg/100 g T.S., détectés par la distribution exponentielle ne le seront pas par la distribution de Pareto.

Les quantiles estimés les plus stables quel que soit le niveau de troncature sont à nouveau obtenus pour la distribution de Pareto ; ceux obtenus pour la distribution exponentielle sont plus stables que pour les quantiles 0,99 (voir figure II.8).

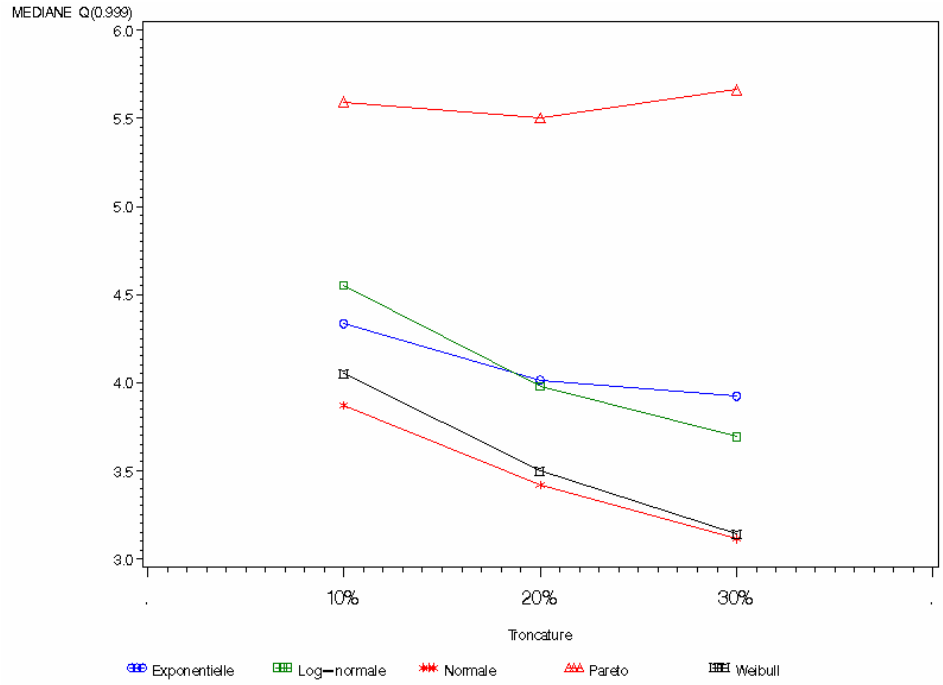


Figure II.18. Elément carbone (partie droite) - Quantiles estimés (0,999) en fonction du niveau de troncature.

5.2.4. Evaluation de la qualité des quantiles estimés 0,999 par l'étude de leur variabilité

Comme pour l'étude des quantiles 0,99, la valeur de la médiane des différences entre les limites supérieures et inférieures des quantiles estimés en fonction du niveau de troncature et pour les cinq distributions étudiées est présentée (figure II.15).

Comme pour les quantiles 0,99, la variabilité des quantiles est plus élevée pour la distribution de Pareto et pour le niveau de troncature de 10%.

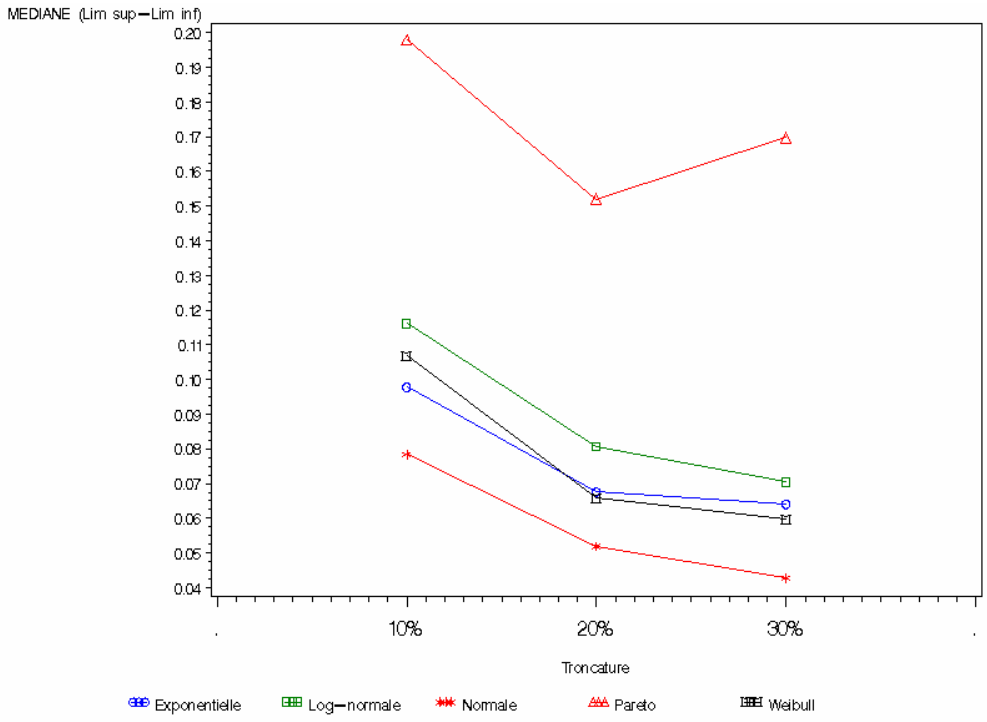


Figure II.19. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,999) pour les trois niveaux de troncature et pour les cinq distributions étudiées.

A nouveau, afin de connaître la relation entre le nombre d'observations par entité communale et la variabilité des quantiles estimés, la figure II.20 présente le graphique avec en abscisse le nombre d'observations et en ordonnée l'intervalle entre les limites supérieures et inférieures autour du quantile 0,999, pour l'ensemble des distributions et des niveaux de troncature. Ces graphiques permettent de confirmer qu'un nombre d'observations de 800 à 1000 sont nécessaires pour obtenir une estimation fiable des quantiles 0,999.

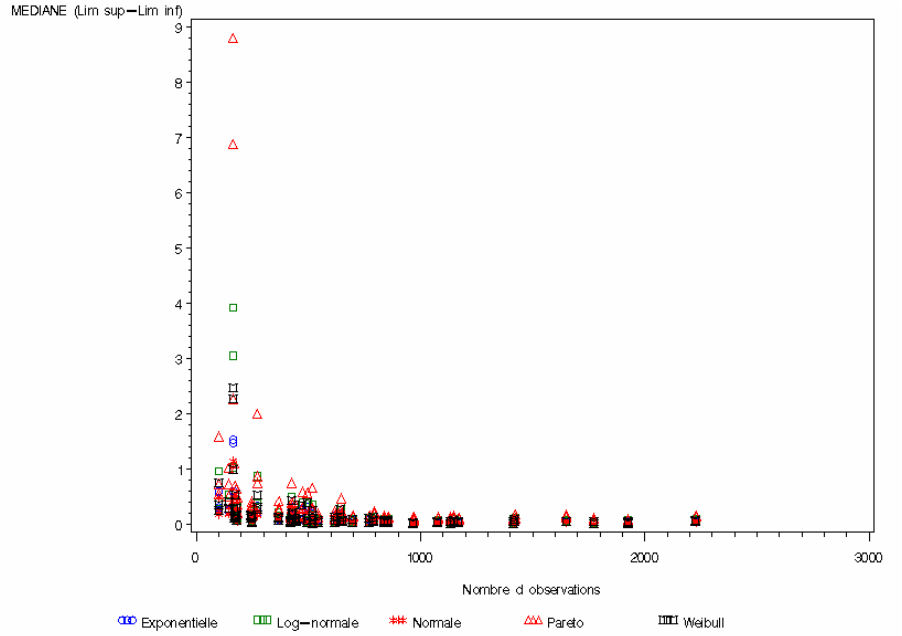


Figure II.20. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,999) en fonction du nombre d'observations pour les cinq distributions étudiées et les trois niveaux de troncature.

La relation entre les quantiles estimés et la variabilité de ces quantiles est présentée à la figure II.21 avec en abscisse la médiane des quantiles estimés et en ordonnée la médiane de l'intervalle entre les limites supérieures et inférieures. A partir de cette figure, on observe à nouveau que plus les quantiles estimés sont élevés, plus l'intervalle autour de ces quantiles est élevé quelle que soit la distribution ou le niveau de troncature.

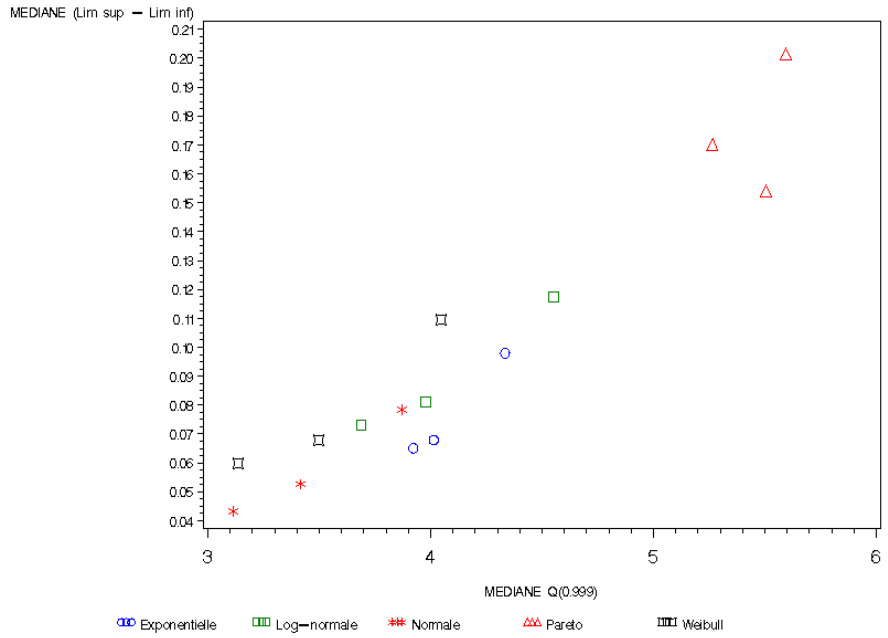


Figure II.21. Elément carbone : médiane des différences entre les limites supérieures et inférieures des quantiles (0,999) en fonction de la médiane des quantiles 0,999 pour les cinq distributions et les trois niveaux de troncature.

5.3. Sélection de la distribution et du niveau de troncature pour la détection des valeurs aberrantes

A partir des études menées sur l'évaluation de la qualité des paramètres estimés (RMSE) et de la qualité de l'estimation des quantiles 0,99 et 0,999[24], il nous semble que **la distribution exponentielle avec un niveau de troncature de 10% est la plus intéressante**. Les arguments qui nous permettent de justifier notre choix sont les suivants :

1. En ce qui concerne la qualité des paramètres estimés, la distribution exponentielle présente de manière globale les valeurs de RMSE (médiane) les plus faibles (figure II.7).
2. Le biais entre les quantiles estimés (0,99) et les valeurs observées est le plus faible pour le niveau de troncature de 10% quelle que soit la distribution et particulièrement pour les distributions exponentielles, log-normale (sous-estimation) et de Pareto (surestimation) (figure II.22).
3. L'écart-type du biais entre les quantiles 0,99 et les quantiles observés correspondants est le plus faible pour la distribution exponentielle et le niveau de troncature de 10% (tableau II.12).
4. En comparaison à la distribution de Pareto, la variabilité des quantiles 0,99 et 0,999 est plus faible pour les quantiles estimés à partir de la distribution exponentielle (figures II.23 et II.24).
5. Les quantiles 0,999 sont stables par rapport aux trois niveaux de troncature (figure II.25).

La distribution de Pareto présente néanmoins les caractéristiques intéressantes suivantes :

1. L'estimation des quantiles 0,99 et 0,999 est plus stable pour la distribution de Pareto quel que soit le niveau de troncature (figures II.26 et II.27).
2. Le biais entre les quantiles estimés (Q 0,99) et les valeurs observées est le plus faible pour le niveau de troncature de 10% et est équivalent à celui de la distribution exponentielle (figure II.28).
3. La variabilité des quantiles 0,99 et 0,999 est la plus élevée pour le niveau de troncature de 10%.

[24] Le coefficient de corrélation entre les quantiles estimés 0.99 et 0.999 est de 0.993.

5.4. Etude des propriétés de la distribution et du niveau de troncature sélectionnés (à partir des entités communales regroupées *a priori*)

5.4.1. Identification des valeurs aberrantes d'origine

Afin de caractériser la distribution et le niveau de troncature sélectionnés, nous recherchons, comme exposé au paragraphe 4.3.9, le pourcentage de valeurs aberrantes d'origine détectées par les quantiles estimés, sachant que pour le quantile 0,99, 1% de valeurs aberrantes devraient être détectées et que pour le quantile 0,999, ce pourcentage est de 0,1%.

Le tableau II.13 présente, pour le carbone, le pourcentage de valeurs aberrantes d'origine détectées à partir des quantiles estimés 0,99 sur base du nombre total d'observations (n=28.209) ; ceci est réalisé par distribution et niveau de troncature. Le dénombrement des valeurs aberrantes d'origine est réalisé commune par commune et est ensuite totalisé.

Tableau II.13. Elément carbone (partie droite) : nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature et pourcentage sur base du nombre total d'observations - quantile 0,99.

Niveau de troncature	Normale	Weibull	Lognormale	Exponentielle	Pareto
10%	356 (1,26%)	332 (1,18%)	293 (1,04%)	**301 (1,07%)**	243 (0,86%)
20%	523 (1,85%)	488 (1,73%)	393 (1,39%)	377 (1,34%)	228 (0,81%)
30%	703 (2,49%)	680 (2,41%)	486 (1,72%)	424 (1,50%)	221 (0,78%)

On observe ainsi qu'à partir des quantiles estimés 0,99, le pourcentage de détection de valeurs aberrantes d'origine varie de 2,49%, pour la distribution normale à 0,78% pour la distribution de Pareto, chacune pour un niveau de troncature de 30%. La distribution de Pareto présente une différence par rapport aux autres distributions avec des pourcentages de détection inférieurs à 1%, ce qui correspond à la surestimation des quantiles par cette distribution ; c'est-à-dire des quantiles plus élevés et un pourcentage de détection plus faible.

Le pourcentage le plus proche de 1% est obtenu pour le niveau de troncature de 10% pour l'ensemble des distributions et en particulier pour les distributions exponentielle et log-normale avec un pourcentage le plus proche de 1% ; ceci correspond au biais le plus faible présenté précédemment.

Le tableau II.14 présente le nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature pour le quantile 0,999.

Tableau II.14. Elément carbone (partie droite) : nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature et pourcentage par rapport au nombre total d'observations - quantile 0,999.

Niveau de troncature	Normale	Weibull	Lognormale	Exponentielle	Pareto
10%	115 (0,41%)	90 (0,32%)	58 (0,21%)	**56 (0,20%)**	17 (0,06%)
20%	296 (1,05%)	183 (0,65%)	91 (0,32%)	78 (0,28%)	21 (0,07%)
30%	453 (1,61%)	309 (1,10%)	135 (0,48%)	98 (0,35%)	12 (0,04%)

On observe ainsi qu'à partir des quantiles estimés 0,999, le pourcentage de détection de valeurs aberrantes d'origine varie de 1,61 %, pour la distribution normale à un pourcentage proche de zéro pour la distribution de Pareto, chacune pour un niveau de troncature de 30%.

Pour le niveau de troncature retenu qui est de 10%, les distributions exponentielle et log-normale présentent un taux de détection de valeurs aberrantes d'origine de 0,20% et 0,21% et la distribution de Pareto de 0,06% ; il faut néanmoins se rappeler que la distribution de Pareto présente une plus grande variabilité des quantiles.

La distribution de Pareto présente une différence par rapport aux autres distributions avec des pourcentages de détection inférieurs à 0,1%, ce qui correspond à nouveau à la surestimation des quantiles par cette distribution.

5.4.2. Evaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles

Comme présenté au paragraphe 4.3.9, nous cherchons à évaluer le taux de détection de valeurs aberrantes issues d'autres régions agricoles, à savoir la Famenne et l'Ardenne. De plus, nous distinguons les contaminants issus de communes voisines de la Famenne ou non. Le tableau II.15 présente le pourcentage de valeurs aberrantes issues de ces régions agricoles pour les quantiles 0,99 et 0,999 pour la distribution exponentielle et le niveau de troncature de 10%.

A partir de ce tableau, on observe que plus les contaminants sont issus de zones éloignées de la région agricole du Condroz, plus le taux de détection en tant que valeurs aberrantes est élevé.

Tableau II.15. Elément carbone (partie droite) : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution exponentielle et le niveau de troncature de 10%, à partir des différentes entités regroupées *a priori* - quantiles 0,99 et 0,999.

Origine des contaminants détectés comme aberrants	**Quantile 0,99**	**Quantile 0,999**
Contaminants des communes de l'Ardenne	54,07 %	18,73 %
Contaminants des communes de la Famenne non voisines au Condroz	16,43 %	5,20 %
Contaminants des communes de la Famenne voisines au Condroz	7,93 %	2,17 %

Afin d'affiner notre étude sur l'évaluation du taux de détection de valeurs aberrantes, la région agricole du Condroz a été scindée en trois régions distinctes formées des communes du Condroz voisines à la Famenne, des communes centrales du Condroz et des communes du Condroz voisines à la Région limoneuse.

Le tableau II.17 présente de manière synthétique le taux de détection moyen obtenu suite au calcul des pourcentages de détection de valeurs aberrantes issues d'autres régions agricoles dans chaque entité communale du Condroz. Ces résultats sont présentés également pour la distribution exponentielle, le niveau de troncature de 10% et pour les quantiles 0,99 et 0,999.

Pour le quantile 0,99, on observe une nette différence entre les contaminants détectés comme aberrants issus de l'Ardenne (49,2% à 60,3%), des communes de la Famenne non voisines au Condroz (15,1% à 18,1%) et, enfin, des communes de la Famenne voisines au Condroz (7,2% à 8,5%). Pour le quantile 0,999, la différence est très importante également entre les communes issues de l'Ardenne (16,8% à 22,7%) et les communes de la Famenne. Entre communes de la Famenne (voisines ou non du Condroz), la différence est moins importante (4,6%-5,9% et 1,9%-2,5%).

Les mêmes observations peuvent être réalisées à partir des taux de détection de valeurs aberrantes obtenus à partir des quantiles 0,999 estimés. En effet, pour les contaminants issus de communes de l'Ardenne, ces taux varient de 22,7% à 16,7%. Pour les contaminants issus de communes de la Famenne non voisines au Condroz, les taux sont de 5,9%à 4,6%. Enfin, pour les contaminants issus de communes la Famenne voisines au Condroz, ces taux varient de 2,5% à 2,1%.

Les résultats indiquent donc des différences importantes entre les taux de détection de contaminants issus des communes situées en bordure du Condroz (Famenne) ou en Ardenne. La détection de valeurs aberrantes est donc de plus en plus importante au fur et à mesure qu'on s'éloigne du Condroz. A l'intérieur du Condroz, on observe quelques différences entre les communes centrales du Condroz et les communes voisines à la Famenne

ou à la Région limoneuse. Le taux de détection le plus élevé est rencontré pour les communes centrales du Condroz, où la contamination devrait être la moins élevée.

Trois zones différentes pourraient donc être considérées : la région centrale et les régions voisines à d'autres régions agricoles. Ceci est un critère important pour le regroupement et pourrait constituer une alternative simple de classification à l'intérieur d'une région agricole.

5.4.3. Evaluation du rapport d'efficacité

Afin de permettre de réaliser des comparaisons entre les différentes méthodes de détection, nous évaluons dans ce paragraphe le rapport d'efficacité basé sur le quantile 0,999 pour la distribution exponentielle et le niveau de troncature de 10% (tableau II.16).

A partir de ce tableau, on observe que plus les contaminants sont issus de zones éloignées de la région agricole du Condroz, plus le rapport d'efficacité est élevé.

Tableau II.16. Elément carbone (partie droite) : rapport d'efficacité calculé sur base du quantile 0,999 pour la distribution exponentielle et le niveau de troncature de 10% à partir des différentes entités regroupées *a priori*.

Origine des contaminants détectés comme aberrants	**Rapport d'efficacité**
Contaminants des communes de l'Ardenne	96,4
Contaminants des communes de la Famenne non voisines au Condroz	26,8
Contaminants des communes de la Famenne voisines au Condroz	11,2

Tableau II.17. Elément carbone : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution exponentielle et le niveau de troncature de 10% à partir des différentes entités regroupées *a priori* - quantiles 0,99 et 0,999 – distinction entre différentes zones de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	**Quantile 0,99**	**Quantile 0,999**
Etude des communes du Condroz voisines à la Famenne		
Contaminants de communes de l'Ardenne comparés dans les communes du Condroz voisines à la Famenne	52,7 %	16,8 %
Contaminants de communes de la Famenne non voisines au Condroz comparés dans les communes du Condroz voisines à la Famenne	15,1 %	4,6 %
Contaminants de communes de la Famenne voisines au Condroz comparés dans les communes du Condroz voisines à la Famenne	7,2 %	1,9 %
Etude des communes centrales du Condroz		
Contaminants des communes de l'Ardenne comparés dans les communes centrales du Condroz	60,3 %	22,7 %
Contaminants des communes de la Famenne non voisines au Condroz comparés dans les communes centrales du Condroz	18,1 %	5,9 %
Contaminants des communes de la Famenne voisines au Condroz comparés dans les communes centrales du Condroz	8,5 %	2,5 %
Etude des communes voisines à la Région limoneuse		
Contaminants des communes de l'Ardenne comparés dans les communes voisines à la Région limoneuse	49,2 %	16,7 %
Contaminants des communes de la Famenne non voisines au Condroz comparés dans les communes voisines à la Région limoneuse	16,1 %	5,1 %
Contaminants des communes de la Famenne voisines au Condroz comparés dans les communes voisines à la Région limoneuse	8,1 %	2,1 %

5.5. Ajustements et évaluation de la qualité des paramètres estimés (pour l'ensemble des données du Condroz)

Un ajustement de la distribution exponentielle a été réalisé à partir de l'ensemble des observations de la région agricole du Condroz sans distinguer les entités communales. Pour le niveau de troncature de 10%, le nombre d'observations est de 2881[25].

La valeur de RMSE obtenue pour la distribution exponentielle est de 0,196 alors que la médiane du RMSE, calculée dans la première partie de ce travail, était de 0,136. La qualité des paramètres estimés est donc moins bonne lorsque les paramètres sont ajustés de manière globale.

5.6. Etude des propriétés de la distribution et du niveau de troncature sélectionnés (à partir de l'ensemble des données du Condroz)

5.6.1. Identification des valeurs aberrantes d'origine

Le nombre de valeurs aberrantes d'origine détectées à partir des valeurs limites estimées à partir de la distribution exponentielle sur l'ensemble des observations pour un niveau de troncature de 10% est présenté au tableau II.18.

Pour le quantile 0,99, le pourcentage de détection était de 1,07%, le pourcentage était donc bien plus précis que dans ce cas-ci (tableau II.13). Pour le quantile 0,999, le pourcentage était de 0,20%.

Tableau II.18. Elément carbone : nombre de valeurs aberrantes d'origine détectées à partir de l'ensemble des observations de la région agricole du Condroz sur base des quantiles 0,99 et 0,999.

[25] Le sous-ensemble de données comprend 28.809 échantillons de sols (paragraphe 4.2).

5.6.2. Evaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles

Le taux de détection de valeurs aberrantes issues d'autres régions agricoles est présenté au tableau II.19. Pour le quantile 0,99, les taux de détection sont assez semblables à ceux présentés au tableau II.17, par contre, les taux observés pour le quantiles 0,999 sont bien plus faibles lorsqu'on travaille de manière globale.

Tableau II.19. Elément carbone : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles, calculé à partir des quantiles 0,99 et 0,999 estimés sur l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	Quantile 0,99	Quantile 0,999
Contaminants de communes de l'Ardenne	54,78 %	6,43 %
Contaminants de communes de la Famenne non voisines au Condroz	13,10 %	2,07 %
Contaminants de communes de la Famenne voisines au Condroz	6,03 %	0,59 %

5.6.3. Evaluation du rapport d'efficacité

Les rapports d'efficacité calculés à partir de l'ensemble des observations de la région agricole du Condroz sont présentés au tableau II.20.

Les rapports d'efficacité calculés à partir de l'ensemble des observations sont bien plus faibles que lorsqu'on travaille entité par entité. Plus particulièrement, dans le cas des contaminants de l'Ardenne, le rapport d'efficacité passe de 96,4 à 28,9.

Tableau II.20. Elément carbone : rapport d'efficacité basé sur le quantile 0,999 pour la distribution exponentielle et le niveau de troncature de 10% et pour l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	Rapport d'efficacité
Contaminants de communes de l'Ardenne	28,9
Contaminants de communes de la Famenne non voisines au Condroz	9,3
Contaminants de communes de la Famenne voisines au Condroz	2,7

5.7. Résultats obtenus à partir des limites actuelles de REQUASUD

Nous présentons, dans ce paragraphe, les résultats obtenus à partir des limites utilisées actuellement par RéQuaSud afin de permettre, par la suite, une comparaison globale des résultats plus aisée avec les résultats issus de la classification spatiale des entités communales.

La valeur limite utilisée au sein du réseau RéQuaSud pour les terres de culture de la région agricole du Condroz et qui correspond au quantile 0,999 est de 3,9.

En utilisant la limite de RéQuaSud, le nombre de valeurs aberrantes d'origine est de 106.

Les taux de détection de valeurs aberrantes issues de la Famenne et de l'Ardenne sont présentés au tableau II.21 tandis que les rapports d'efficacité sont présentés au tableau II.22.

Lorsqu'on compare les rapports d'efficacité, on observe, dans le cas des contaminants de l'Ardenne, que le rapport obtenu à partir des limites de RéQuaSud (37,5) est nettement plus faible que celui obtenu à partir des entités communales regroupées *a priori* (96,4). Le rapport d'efficacité obtenu par RéQuaSud est cependant plus élevé que celui obtenu à partir de l'ensemble des observations de la région agricole du Condroz (28,9).

Pour les contaminants issus des communes de la Famenne, les résultats obtenus à partir de RéQuaSud sont nettement plus faibles que pour les entités regroupées *a priori*.

Une discussion plus globale de ces résultats est présentée au paragraphe 5.9 suite à la réalisation de la classification spatiale.

Tableau II.21. Elément carbone : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles à partir des valeurs limites actuelles du réseau RéQuaSud pour l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	% de détection à partir des limites de RéQuaSud
Contaminants de communes de l'Ardenne	13,82 %
Contaminants de communes de la Famenne non voisines au Condroz	2,55 %
Contaminants de communes de la Famenne voisines au Condroz	1,31 %

Tableau II.22. Elément carbone : rapport d'efficacité calculé à partir des valeurs limites actuelles du réseau RéQuaSud pour l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	**Rapport d'efficacité**
Contaminants de communes de l'Ardenne	37,5
Contaminants de communes de la Famenne non voisines au Condroz	6,8
Contaminants de communes de la Famenne voisines au Condroz	3,5

5.8. Classification spatiale

5.8.1. Evaluation du rapport d'efficacité

La démarche présentée au paragraphe 4.3.10.a est appliquée afin de constituer des groupes d'entités communales similaires d'effectif suffisant. Nous observerons suite à la présentation des résultats relatifs à la variabilité des quantiles extrêmes (paragraphe 5.2.2.b), que cet effectif correspond à un nombre d'observations d'environ 800. Cependant, les quantiles extrêmes estimés pour chaque entité communale (ou groupe d'entités regroupées *a priori*) sont issus d'ajustements réalisés à partir d'un nombre d'observations plus faible. En effet, seules 15 entités sur les 39 étudiées présentent cet effectif minimum. Nous appliquons dès lors la classification avec contrainte de contiguïté spatiale afin de regrouper des entités communales présentant, d'une part, des quantiles extrêmes similaires et, d'autre part, un effectif d'au moins 800 observations. Pour le carbone, le regroupement qui conduit à l'effectif proche de 800 correspond à 9 groupes.

Cependant, que ce soit pour la création de 2, 3, …, 9 groupes, un même ensemble d'entités communales est uniformément créé. Ce groupe correspond aux entités communales de Charleroi (regroupé *a priori* avec Châtelet et Montignies-le-Tilleul) et de Aiseau-Presles (regroupé *a priori* avec Farciennes) et est composé seulement de 338 observations ; ceci est observé tant pour le carbone que pour le calcium. Il est important de signaler ici que la zone du Condroz pédologique ne correspond pas à la région agricole du Condroz, à partir de laquelle nous travaillons. La région de Charleroi ainsi que la région de Lobbes ne font pas partie du Condroz pédologique mais sont comprises dans la région agricole du Condroz. Quant à celle de Beaumont, elle constitue plutôt la bordure Ouest du Condroz pédologique.

Notre critère lié à un effectif minimum de 800 observations n'est donc pas rencontré pour ce groupe car celui-ci présente des caractéristiques trop différentes des autres entités communales (quantiles extrêmes très différents). Le critère de l'effectif minimum est cependant respecté pour les autres groupes formés. Le critère de l'effectif minimal devrait donc être appliqué dans la mesure où il ne modifie pas de manière conséquente la constitution des groupes.

Le nombre de valeurs aberrantes d'origine détectées a été calculé ainsi que le taux de détection des valeurs aberrantes issues d'autres régions agricoles (annexe 6). Les rapports d'efficacité ont ensuite été calculés. Ils sont présentés à l'aide des graphiques des figures II.29, II.30 et II.31, respectivement pour la détection de contaminants issus de communes de l'Ardenne, de communes de la Famenne non voisines au Condroz et de communes de la Famenne voisines au Condroz.

A partir de ces figures, nous observons des résultats tout à fait comparables quelle que soit l'origine des contaminants d'autres régions agricoles (soit de l'Ardenne, soit de la Famenne dont les communes sont voisines ou non du Condroz).

Lorsqu'on compare les résultats obtenus pour les différents regroupements, le rapport d'efficacité le plus élevé est rencontré pour celui formé de 8 groupes d'entités communales. Ce rapport est cependant assez variable d'un regroupement à l'autre et dépend de la présence de l'une ou l'autre entité communale dans les groupes formés. Ceci peut être visualisé à l'aide des figures 1 et 2 de l'annexe 7. Le cercle rouge représente un groupe d'entités communales qui, lorsque ces entités sont incluses, par exemple, dans le groupe 4 de la figure 1 (4 groupes formés), conduit à un rapport d'efficacité de 69.1. Par contre, lorsque ces entités communales sont comprises dans le groupe 4 de la figure 2 (5 groupes formés), le rapport d'efficacité diminue nettement et il est proche de 40.

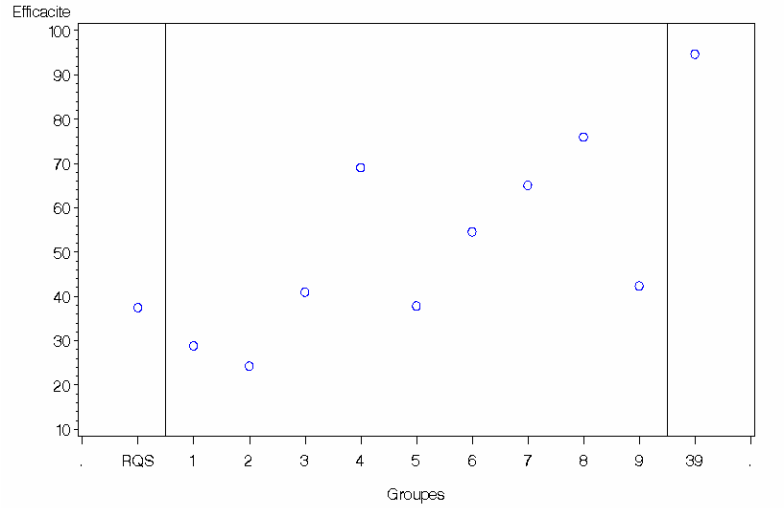


Figure II.29. Elément carbone : évolution du rapport d'efficacité en fonction du nombre de groupes formés - comparaison pour les contaminants issus de l'Ardenne.

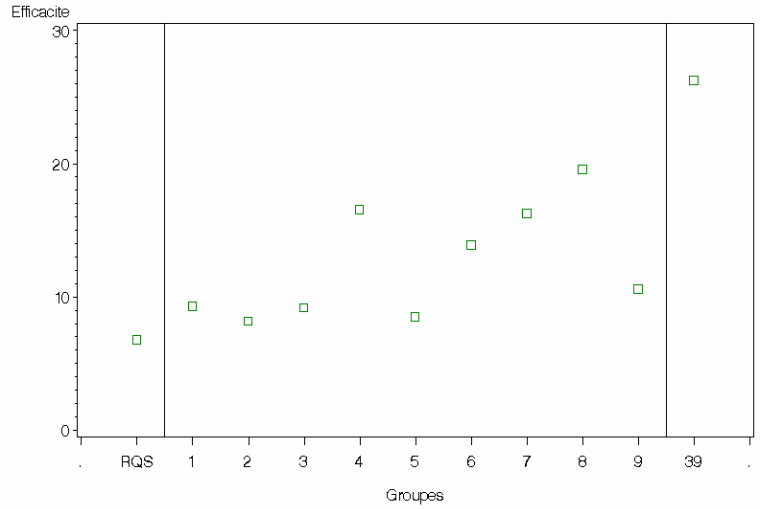


Figure II.30. Elément carbone : évolution du rapport d'efficacité en fonction du nombre de groupes formés - comparaison pour les contaminants issus des communes de la Famenne non voisines au Condroz.

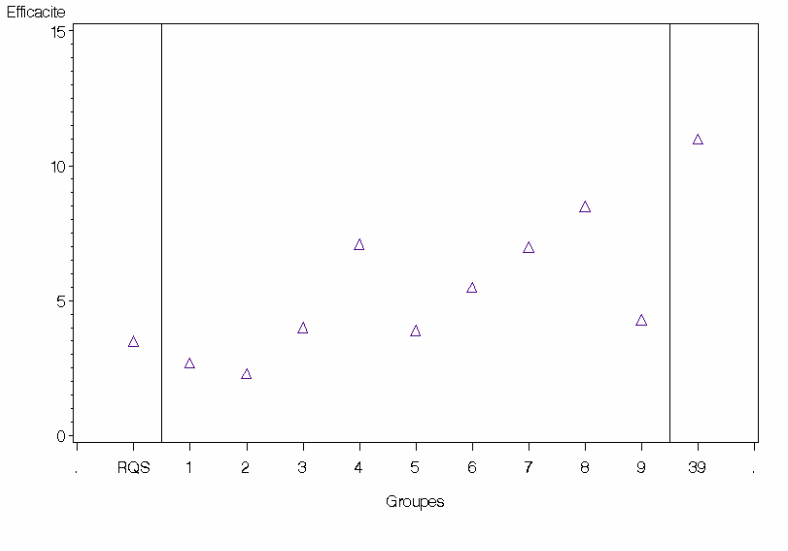


Figure II.31. Elément carbone : évolution du rapport d'efficacité en fonction du nombre de groupes formés - comparaison pour les contaminants issus des communes de la Famenne voisines au Condroz.

Le regroupement constitué de 8 groupes est retenu car c'est pour ce regroupement que le rapport d'efficacité est le plus élevé parmi les regroupements réalisés. Le regroupement de 4 groupes aurait pu être sélectionné également, étant donné l'écart relativement faible entre les deux regroupements. Il est vrai qu'au niveau pratique, il serait plus intéressant, pour un même niveau d'efficacité, de retenir le regroupement le plus simple, c'est-à-dire celui qui présente le nombre de groupes le plus faible. Cependant, la comparaison des représentations graphiques de la figure II.32 (pour 8 groupes) et de la figure 1 de l'annexe 7 (pour 4 groupes), ainsi que les valeurs des quantiles extrêmes par groupe, nous incitent à garder le regroupement de 8 groupes. Aussi, l'interprétation des résultats au niveau pédologique ne pourra en être que plus étoffée.

Afin de faciliter la discussion des résultats, le tableau II.23 présente les rapports d'efficacité pour le regroupement de 8 communes, pour les 39 communes distinctes et pour l'ensemble du Condroz (c'est-à-dire 1 seul groupe). Le rapport d'efficacité obtenu pour les 8 groupes est, comme cela était prévisible, plus important que lorsqu'un seul groupe est pris en compte (ensemble du Condroz). Par rapport aux résultats de RéQuaSud, le rapport d'efficacité est également plus élevé. La constitution de différents groupes permet donc bien d'améliorer le processus de détection des valeurs aberrantes.

Tableau II.23. Elément carbone : évolution du rapport d'efficacité en fonction du nombre de groupes formés.

	Contaminants de communes de l'Ardenne	**Contaminants de communes de la Famenne non voisines au Condroz**	**Contaminants de communes de la Famenne voisines au Condroz**
1 groupe	28,9	9,3	2,7
8 groupes	76,0	19,6	8,5
39 entités	94,7	26,3	11,0
RéQuaSud	37,5	6,8	3,5

De même, le rapport d'efficacité obtenu à partir des 39 entités communales distinctes est 2,5 fois supérieur à celui obtenu en prenant les limites de RéQuaSud, voire 3 fois supérieur à celui de l'ensemble du Condroz pour lequel la distribution exponentielle et le niveau de troncature de 10% ont été utilisés pour le calcul des quantiles extrêmes. L'hypothèse émise en début de travail « *La mise en place d'un système de détection de valeurs aberrantes en intégrant la contrainte spatiale est plus robuste, plus cohérente que les méthodes qui considèrent que les populations sont homogènes dans l'espace (méthode actuellement appliquée au sein de RéQuaSud)* » est ici vérifiée par le gain d'efficacité pour la détection de valeurs aberrantes.

Lorsque les résultats obtenus pour les 8 groupes sont comparés par rapport à ceux obtenus à partir des 39 communes séparées, on observe que le regroupement conduit à une perte d'efficacité. En effet, le rapport d'efficacité obtenu à partir de chacune des entités communales est plus élevé (exemple dans le cas des contaminants issus de l'Ardenne : 94,7) que pour les 8 groupes (76,0). Comme le souligne Foguenne (1995), si l'inclusion de la contrainte de contiguïté se traduit par la création de zones non morcelées et par conséquent par une représentation plus facile des résultats obtenus, cette méthode conduit à une perte d'homogénéité des groupes formés. Il faut dès lors se demander quel est l'intérêt, dans notre cas, du regroupement sachant qu'il existe une perte d'efficacité importante. Cette question sera discutée au niveau de la discussion générale tant pour le carbone que pour le calcium.

Notons enfin que les résultats obtenus à partir de l'ensemble des observations (1 groupe) sont inférieurs à ceux obtenus par RéQuaSud.

5.8.2. Représentation graphique des groupes d'entités communales

Les huit groupes du regroupement qui présente le rapport d'efficacité le plus élevé sont représentés à la figure II.32. Pour les différents groupes, les quantiles estimés 0,999 par groupe sont présentés au tableau II.24.

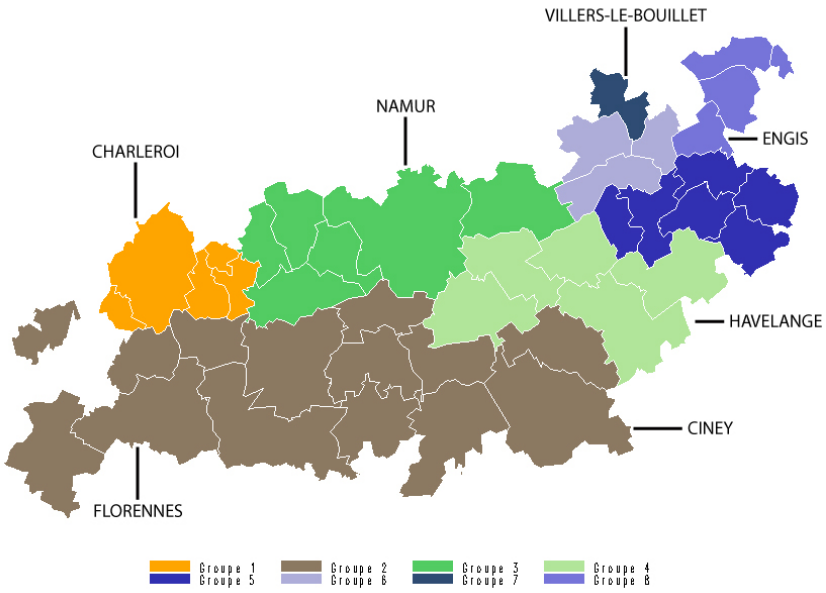


Figure II.32. Elément carbone : création de 8 groupes *a posteriori*.

Tableau II.24. Elément carbone : effectifs et quantiles estimés 0,999 pour les 8 groupes formés.

Groupes	n	Quantiles estimés 0,999
1	338	9,103
2	9929	4,551
3	3282	4,405
4	5666	4,001
5	3391	3,641
6	3338	3,559
7	1773	2,847
8	1092	4,527

Le groupe 1 est composé de l'entité de Charleroi et des communes voisines pour lesquelles le quantile extrême est bien plus élevé que pour les autres entités communales. Comme cité au début de ce paragraphe 6.8, la région de Charleroi ne fait pas partie du Condroz pédologique et la carte des principaux types de sols indique dans cette région peu de zones calcaires typiques du Condroz. C'est une région où les limons hydromorphes sont plus représentés entraînant un drainage pauvre. Ceci pourrait représenter un des facteurs explicatifs de ce taux plus élevé de carbone organique total en relation avec une minéralisation ralentie de la matière organique. D'autres facteurs en relation avec les pratiques culturales par exemple sont également envisageables (apports d'intrants organiques d'origine différente, diversité des spéculations). Les observations d'échantillons de sols liées à la présence de limons auraient donc été prises en compte au niveau de la queue de distributions, pour l'estimation des quantiles.

On aurait pu alors s'attendre à ce que les entités de Beaumont et de Lobbes se distinguent du groupe 2. Cependant, l'effectif trop faible ne permet pas d'y parvenir ; des essais de regroupements de 10, 11 groupes, non présentés ici, ont indiqué une séparation de cette région ouest de la région agricole du Condroz. Notons, entre autres, que la technique du chaulage pourrait influencer les résultats obtenus et que le taux de carbone aurait pu dès lors être diminué de manière artificielle. En effet, une concentration de calcaire élevée dans le sol pourrait entraîner une augmentation du pH, ce qui aurait comme conséquence une meilleure minéralisation de la matière organique et entraînerait finalement un taux de carbone plus faible. D'autres facteurs explicatifs peuvent également intervenir dans l'interprétation de ces résultats tels que le mode de gestion des parcelles ou le problème de la représentativité des échantillons dans les groupes formés.

En étudiant les autres groupes formés (figure II.32), trois zones homogènes sont observées. Elles correspondent :

1. au nord de la région agricole de la Famenne (groupe 2) ;
2. à la région limitrophe à la Région limoneuse (groupe 3) ;
3. au nord-est de la région agricole de la Famenne (groupe 4).

La distinction entre les groupes 2 et 3 ne conduit cependant pas à des quantiles estimés très différents d'un groupe à l'autre.

Pour le groupe 4, le quantile extrême est plus faible par rapport aux groupes 2 et 3 alors que les types de sols présents sont relativement similaires. Ces variations pourraient éventuellement s'expliquer par un passé prairial plus important dans l'une ou l'autre région et donc des teneurs en carbone plus élevées et en calcium plus faibles.

La zone du Condroz proche de la Région herbagère est morcelée en 4 groupes distincts (groupes 5, 6, 7 et 8). Les groupes 5 et 6 sont quasiment semblables en terme de quantiles extrêmes ; ceux-ci sont plus faibles que dans la région centrale du Condroz. Le groupe 7 est particulier car il est constitué d'une entité isolée (Villers-le-Bouillet) qui se distingue par le quantile extrême le plus faible. Près de la moitié de la superficie de cette entité est quasiment située dans la Région limoneuse sur laquelle sont situées les terres de culture et pour laquelle le taux de carbone est plus faible. Les autres entités proches de la Région limoneuse telles que Namur ou Andenne sont par contre incluses dans la région agricole du Condroz et les teneurs en carbone sont plus élevées (groupe 3) ; les entités communales de Wanze et d'Amay (groupe 6) possèdent également des terres limoneuses à teneurs en carbone plus faibles. Il suffit par exemple que la majorité des échantillons de sols aient été prélevés au sein de ces bandes limoneuses pour que le quantile extrême de la teneur en carbone soit plus faible.

Enfin, le groupe 8 présente un quantile plus élevé que les groupes 5, 6 et 7. Ce groupe, composé des entités d'Engis, Flémalle, Grâce-Hollogne et Saint-Nicolas, présente moins de zones limoneuses que pour les entités citées ci-dessus ; ceci pourrait expliquer, entre autres, les teneurs en carbone plus élevées. Notons que nous nous trouvons ici dans une zone où les modifications anthropomorphiques sont importantes. Foguenne (1994) avait classé la commune d'Engis en zone très industrielle avec la présence de carrières de chaux.

La comparaison des résultats relatifs à la classification spatiale pour le carbone est présentée en parallèle à ceux obtenus pour le calcium au niveau de la discussion générale.

5.9. Validation de la méthode de détection par comparaison aux résultats de REQUASUD

En comparant les rapports d'efficacité de la méthode de détection de valeurs aberrantes obtenus par entité communale (paragraphe 5.4), par groupe d'entités communales (paragraphe 5.8 – figures II.29 à II.31) et pour l'ensemble des observations de la région agricole du Condroz (paragraphe 5.6) par rapport à ceux obtenus à partir des valeurs limites de RéQuaSud (paragraphe 5.7), les observations suivantes peuvent être formulées. Ces résultats sont identiques que ce soit pour l'Ardenne ou la Famenne.

Les résultats obtenus à partir de l'ensemble des observations sont les moins intéressants en terme de rapport d'efficacité. Par ailleurs, pour ceux obtenus à partir de la limite de RéQuaSud, les rapports d'efficacité sont quelque peu plus élevés que lorsqu'on considère l'ensemble des observations.

Les rapports d'efficacité calculés à partir des différents groupes d'entités communales formés *a posteriori* ou à partir des entités regroupées *a priori* sont nettement plus élevés que ceux obtenus pour RéQuaSud. En effet, le rapport d'efficacité obtenu à partir des 39 entités communales distinctes est 2,5 fois supérieur à celui obtenu en prenant les limites de RéQuaSud. Les résultats acquis à partir des limites de RéQuaSud sont donc moins bons.

La mise en place d'un système de détection de valeurs aberrantes en intégrant la contrainte spatiale permet ainsi un gain d'efficacité pour la détection de valeurs aberrantes.

6. ETUDE DE LA PARTIE GAUCHE DES DISTRIBUTIONS (ELEMENT CALCIUM)

6.1. Ajustements et évaluation de la qualité des paramètres estimés (par entités communales regroupées *a priori*)

6.1.1. Ajustements par distribution et niveaux de troncature

Pour l'élément calcium, les ajustements ont été réalisés pour les 5 distributions et les trois niveaux de troncature pour chacune des 39 communes. Comme présenté dans la partie bibliographique, certaines distributions ne sont pas adaptées à la partie gauche mais nous avons décidé de les présenter à titre de comparaison.

Les cartes identifiant, par différentes couleurs, la distribution présentant le RMSE le plus faible par commune ont également été réalisées pour les trois niveaux de troncature (figures II.33 à II.35). A partir de ces cartes, on observe des zones géographiques assez homogènes pour les distributions de Weibull et normale, principalement pour les niveaux de troncature de 20 et 30%. On peut également observer que la distribution log-normale présente une valeur de RMSE faible pour des entités communales du Condroz voisines à d'autres régions agricoles. En toute logique, les distributions exponentielles et de Pareto ne sont pas du tout observées car la valeur de RMSE est trop élevée. Comme pour l'élément carbone, les valeurs de RMSE ne sont pas influencées par le nombre d'observations des entités communales.

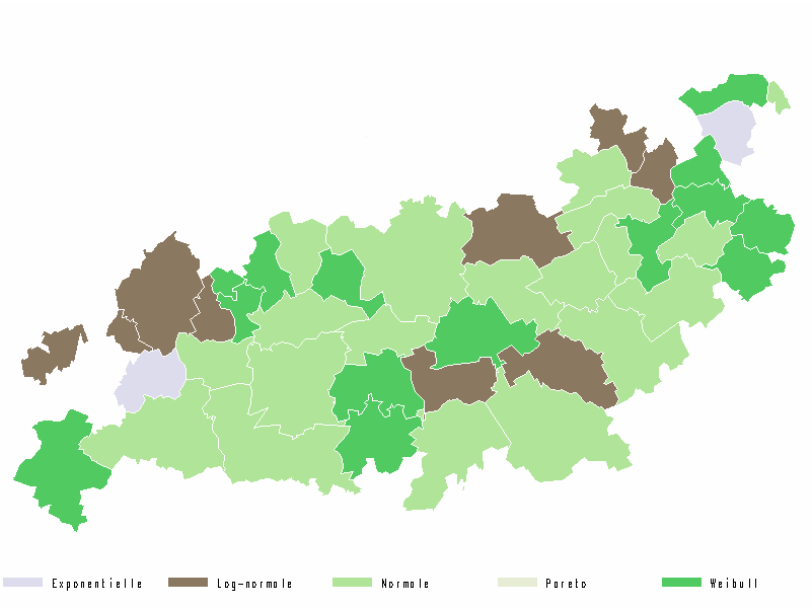


Figure II.33. Elément calcium (partie gauche) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 10% et pour la région agricole du Condroz.

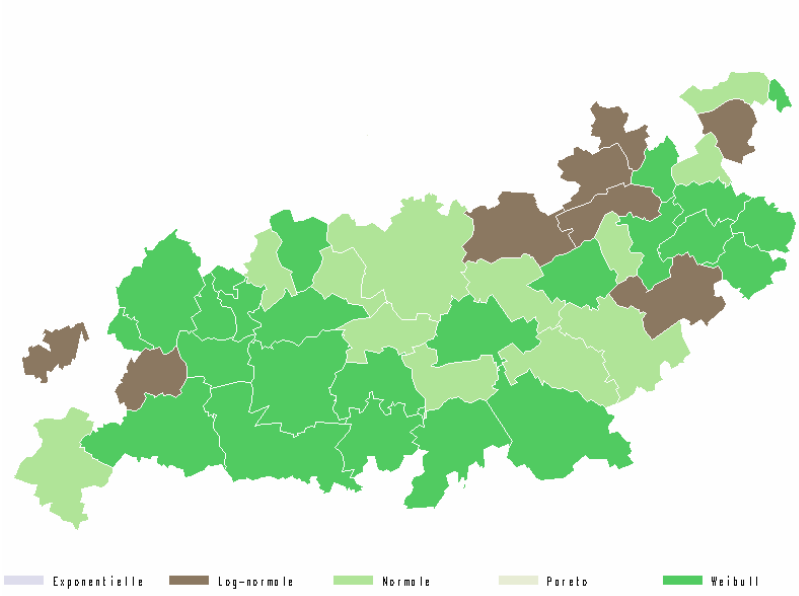


Figure II.34. Elément calcium (partie gauche) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 20% et pour la région agricole du Condroz.

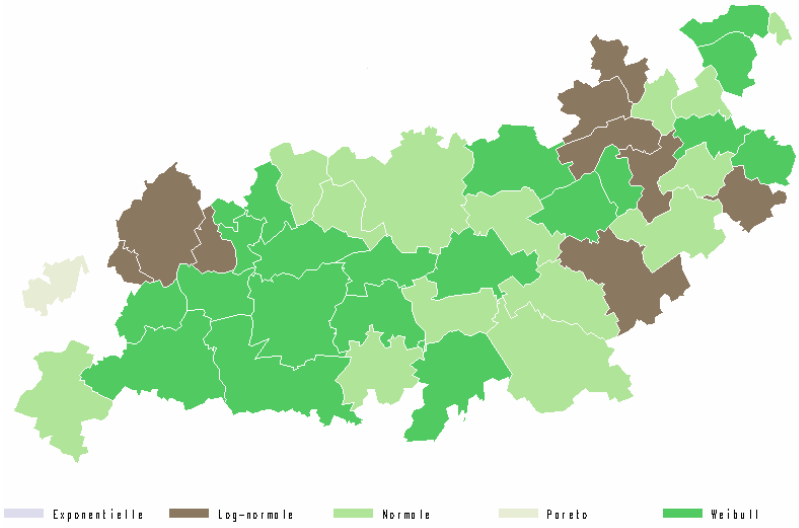


Figure II.35. Elément calcium (partie gauche) : représentation des distributions présentant le RMSE le plus faible par entité communale, pour un niveau de troncature de 30% et pour la région agricole du Condroz.

Le tableau II.25 présente le nombre de fois que chacune des distributions se situe en première position, en deuxième position, etc.

La distribution de Weibull se situe en première position avec 42.7% et en deuxième position avec 35.0%. La distribution normale est en deuxième position avec 32.5% et en deuxième position avec 53.0%. Vient ensuite la distribution log-normale avec 69.2% et finalement les distributions exponentielles et Pareto en $4^{ème}$ et $5^{ème}$ position dans plus de 90% des cas.

Tableau II.25. Elément calcium (partie gauche) : comptage de la position (rang) des différentes distributions par niveau de troncature en fonction de la valeur de RMSE.

Troncature	Rang	normale	Weibull	log-normale	exponentielle	Pareto
10%						
	1er	16	13	8	2	0
	2ème	16	18	3	1	1
	3ème	3	6	28	1	1
	4ème	1	2	0	35	1
	5ème	3	0	0	0	36
20%						
	1er	12	19	8	0	0
	2ème	21	14	4	0	0
	3ème	5	5	27	2	0
	4ème	1	0	0	37	1
	5ème	0	1	0	0	38
30%						
	1er	10	18	10	0	1
	2ème	25	9	3	2	0
	3ème	2	11	26	0	0
	4ème	2	0	0	37	0
	5ème	0	1	0	0	38
Global						
(en pct)	1er	32.5	**42.7**	22.2	1.7	0.9
	2ème	**53.0**	35.0	8.5	2.6	0.9
	3ème	8.5	18.8	**69.2**	2.6	0.9
	4ème	3.4	1.7	0.0	**93.2**	1.7
	5ème	2.6	1.7	0.0	0.0	**95.7**

6.1.2. Etude de l'influence de la troncature sur le RMSE

Nous avons également calculé la médiane des RMSE par niveau de troncature et par distribution pour le calcium (figure II.36).

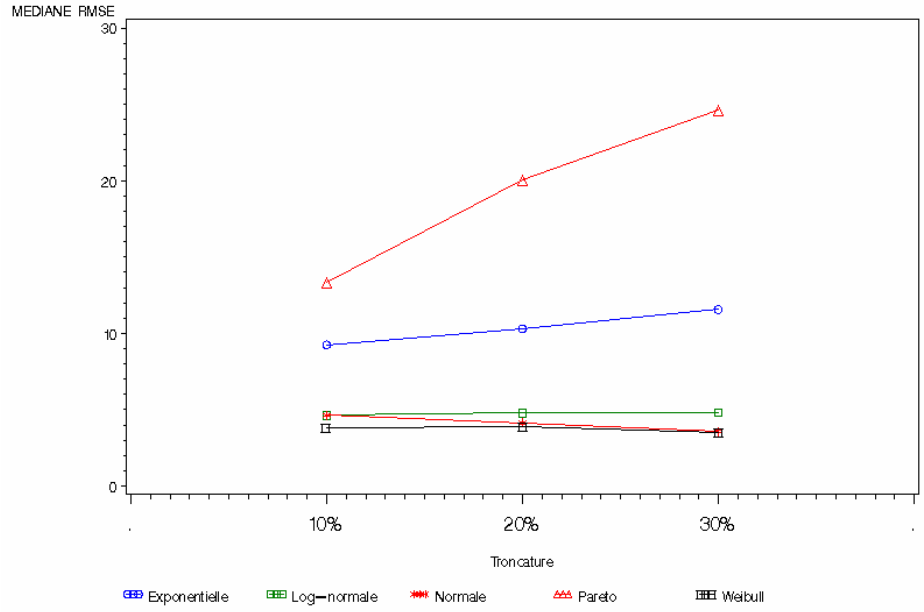


Figure II.36. Elément calcium (partie gauche) : étude de l'influence du niveau de troncature sur la qualité des paramètres estimés (exprimée par le RMSE) – trois niveaux de troncature en fonction de la médiane du RMSE calculé pour l'ensemble des communes étudiées.

Comme cité pour le carbone, la distribution la plus adéquate correspond à celle pour laquelle le RMSE est le plus bas et le plus stable quelle que soit la troncature. **Dans le cas du calcium, la distribution de Weibull est la plus intéressante car il n'y pas d'effet de troncature et le RMSE est le plus faible**. La distribution normale est également intéressante.

Les distributions exponentielles et de Pareto présentent des valeurs de RMSE très élevées.

6.2. Estimation des valeurs limites et évaluation de la qualité de l'estimation (par entité communale regroupée *a priori*)

6.2.1. Etude de l'influence de la troncature sur l'estimation des quantiles

Les valeurs des quantiles 0,01 ont été calculées pour l'élément calcium à partir des paramètres ajustés des différentes distributions, par niveau de troncature.

Comme précédemment, la médiane des quantiles 0,01 est présentée à la figure II.37 par distribution pour les trois niveaux de troncature.

Il faut signaler que dans le cas de la partie gauche des distributions, plus les quantiles estimés sont faibles, plus la détection de valeurs aberrantes est faible ; ce qui est le contraire par rapport à la partie droite.

Dans ce cas-ci, un contaminant d'une teneur en calcium de 105.5 mg/100 g T.S. serait détecté, pour le niveau de troncature de 10%, par les distributions normale, log-normale et de Weibull mais pas par les distributions exponentielle et de Pareto car elles présentent des quantiles inférieurs à 105.5 mg/100 g T.S.

Les quantiles estimés doivent idéalement être stables quel que soit le niveau de troncature. A partir de la figure II.37, on observe que **les distributions normales, log-normale et de Weibull sont relativement stables par rapport au niveau de troncature**. Pour les distributions normale et de Weibull, les quantiles estimés pour le niveau de troncature de 10% sont plus élevés que pour 30% ; le pourcentage de détection de valeurs aberrantes est donc plus élevé. Par contre, pour la distribution log-normale, les quantiles obtenus à partir du niveau de troncature de 30% sont plus élevés que pour la troncature de 10%.

En ce qui concerne les distributions exponentielle et de Pareto, les quantiles estimés sont très variables d'un niveau de troncature à l'autre ; les valeurs des quantiles sont plus basses pour les niveaux de troncature les plus faibles.

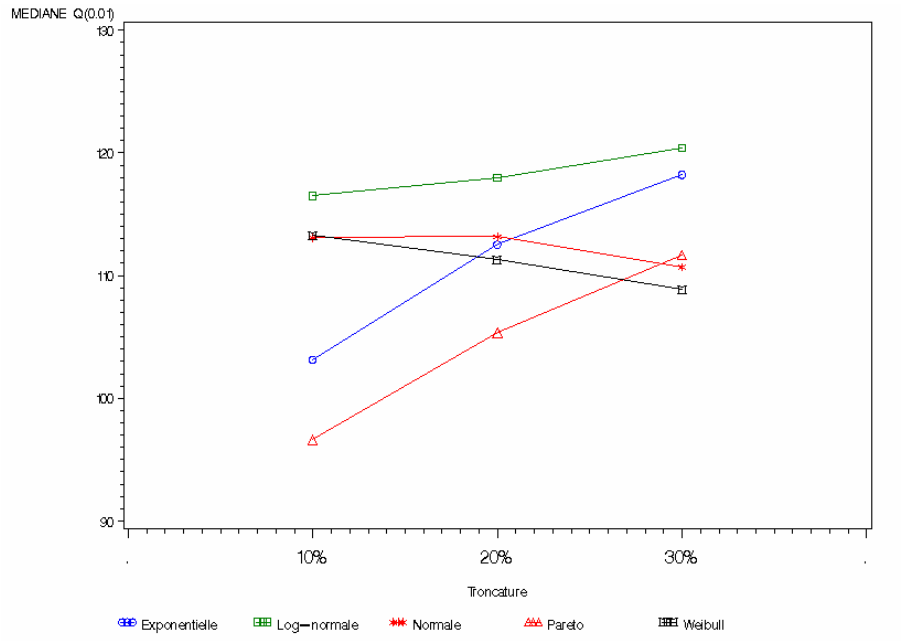


Figure II.37. Elément calcium (partie gauche) - Quantiles estimés en fonction du niveau de troncature.

6.2.2. Evaluation de la qualité des quantiles estimés 0,01

a. Quantiles estimés par rapport aux quantiles observés

Les figures II.38 à II.43 présentent les graphiques des quantiles estimés par rapport aux quantiles observés par niveau de troncature pour le calcium pour les distribuions normale, de Weibull et log-normale, pour les niveaux de troncature de 10 et 30%. Les distributions exponentielle et de Pareto ne sont pas présentées étant donnés les résultats présentés précédemment en terme de RMSE. Les valeurs obtenues pour l'entité communale de code NIS 52011 ne sont à nouveau pas représentées car les quantiles estimés à partir de la distribution normale sont négatifs.

Distribution normale

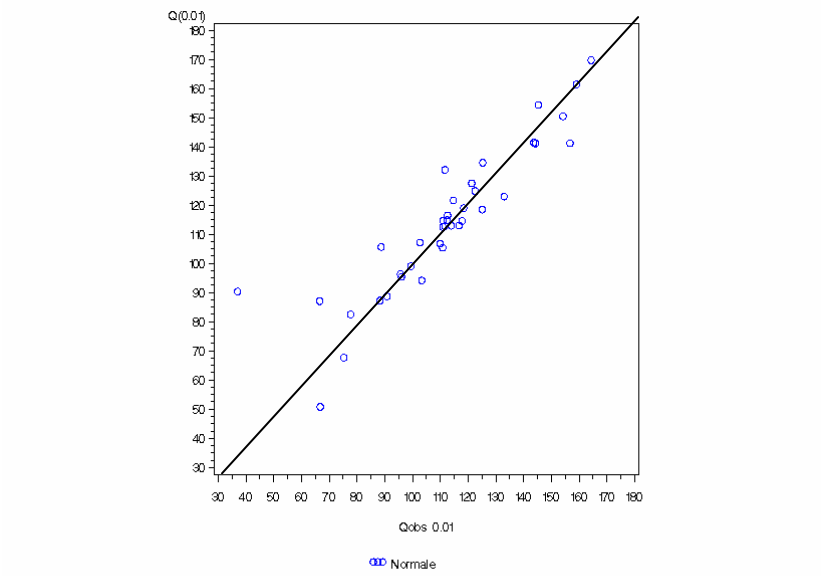


Figure II.38. Elément calcium : quantiles estimés (0,01) à partir de la distribution normale en fonction des quantiles observés pour le niveau de troncature de 10%. *Les valeurs obtenues pour l'entité communale de code NIS 52011 ne sont pas représentées car les quantiles estimés à partir de la distribution normale sont négatifs.*

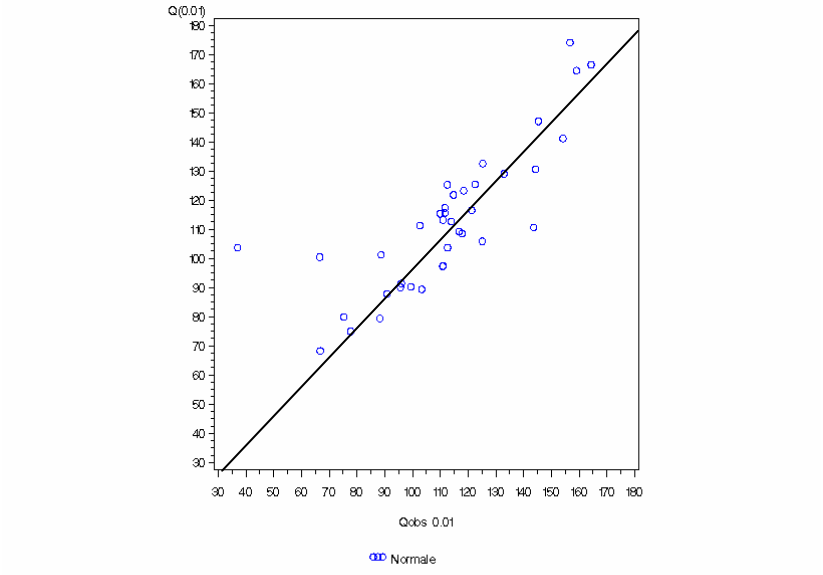


Figure II.39. Elément calcium : quantiles estimés (0,01) à partir de la distribution normale en fonction des quantiles observés pour le niveau de troncature de 30%.

Distribution de Weibull

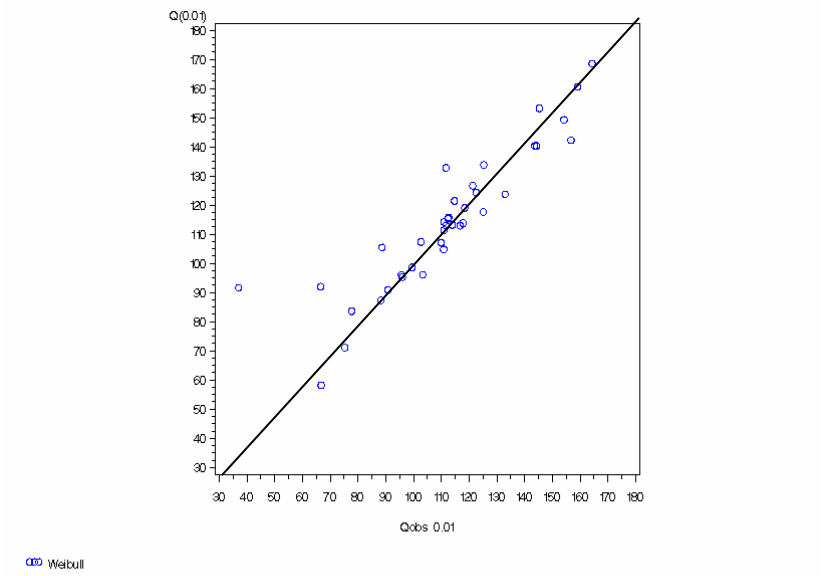


Figure II.40. Elément calcium : quantiles estimés (0,01) à partir de la distribution de Weibull en fonction des quantiles observés pour le niveau de troncature de 10%.

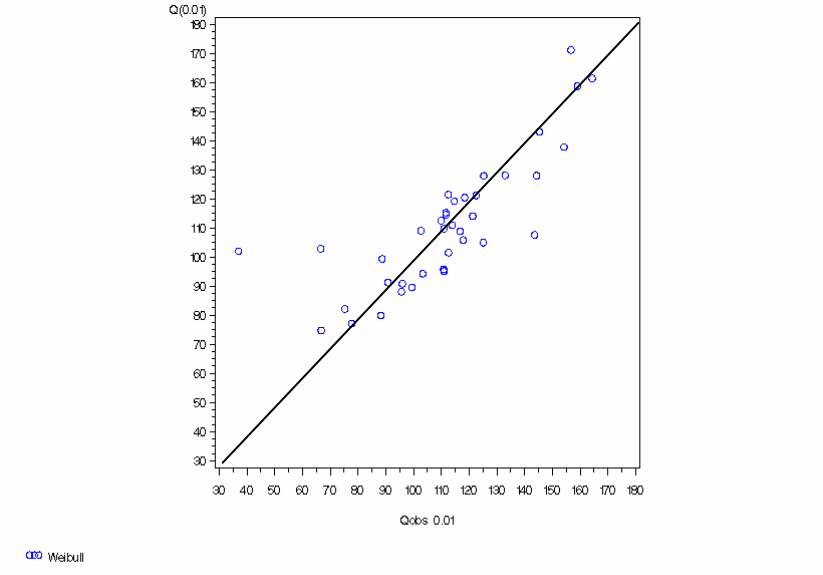


Figure II.41. Elément calcium : quantiles estimés (0,01) à partir de la distribution de Weibull en fonction des quantiles observés pour le niveau de troncature de 30%.

Distribution log-normale

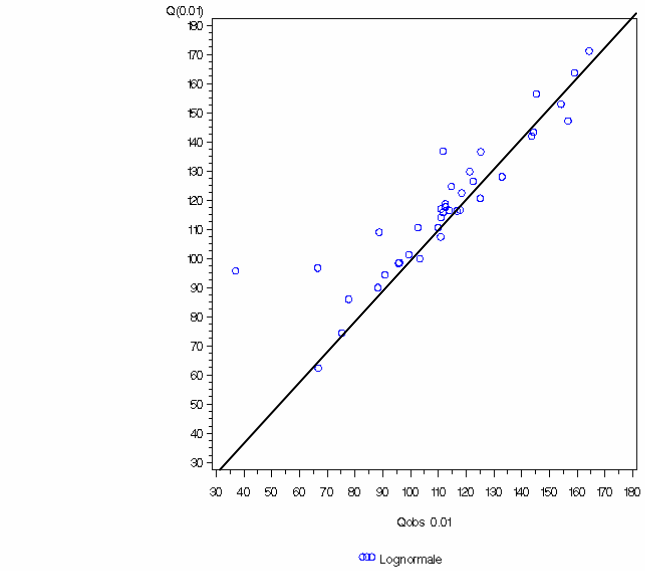


Figure II.42. Elément calcium : quantiles estimés (0,01) à partir de la distribution log-normale en fonction des quantiles observés pour le niveau de troncature de 10%.

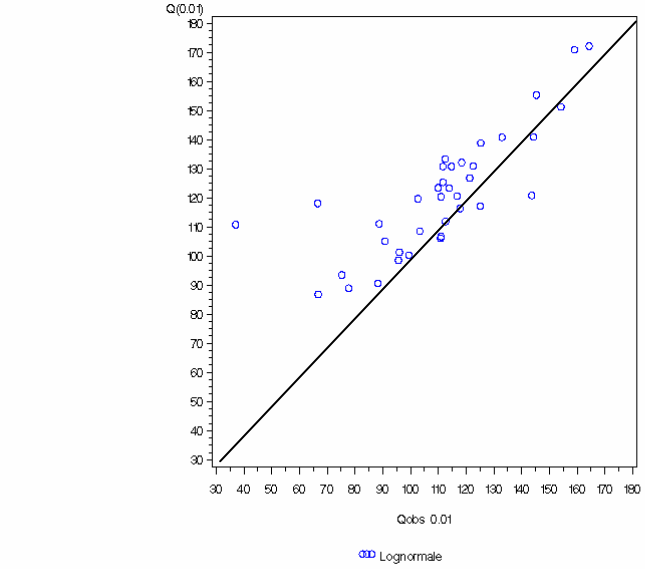


Figure II.43. Elément calcium : quantiles estimés (0,01) à partir de la distribution log-normale en fonction des quantiles observés pour le niveau de troncature de 30%.

Au tableau II.26, les écart-types entre les quantiles estimés 0,01 et les quantiles observés correspondants sont présentés par distribution et niveau de troncature.

La dispersion la plus faible est rencontrée pour le niveau de troncature de 10% pour la distribution exponentielle, pour 20% pour la distribution log-normale et pour 30% pour la distribution de Pareto.

Pour les niveaux de troncature de 10 et 20%, la dispersion des quantiles estimés par rapport aux quantiles observés la plus faible est obtenue pour la distribution exponentielle tandis que pour le niveau de troncature de 30%, la distribution de Weibull présente l'écart-type le plus faible. Néanmoins, il faut rappeler que la distribution exponentielle présentait un RMSE élevé et peu stable par rapport à la distribution de Weibull.

Tableau II.26. Elément calcium : écarts-types entre les quantiles estimés et observés pour le niveau 0,01 par distribution et niveau de troncature.

Troncature	distribution	Ecart-type
10%	normale	12.028
	weibull	12.544
	log-normale	12.814
	exponentielle	**10,991**
	pareto	25.129
20%	normale	14.612
	weibull	15.736
	log-normale	16.254
	exponentielle	**11.759**
	pareto	20.601
30%	normale	20.912
	weibull	**12.991**
	log-normale	18.765
	exponentielle	14.792
	pareto	19.309

La figure II.44 permet de visualiser la médiane des différences entres les quantiles observés et estimés. **Le biais le plus faible est rencontré pour le niveau de troncature de 10%** excepté pour la distribution de Pareto. **Pour le niveau de troncature de 10%, l'estimation des quantiles est la meilleure pour les distributions de Weibull et normale avec une légère tendance à la surestimation. De plus, le biais est relativement stable pour ces deux distributions** alors que pour les autres distributions, le biais est très variable d'un niveau de troncature à l'autre.

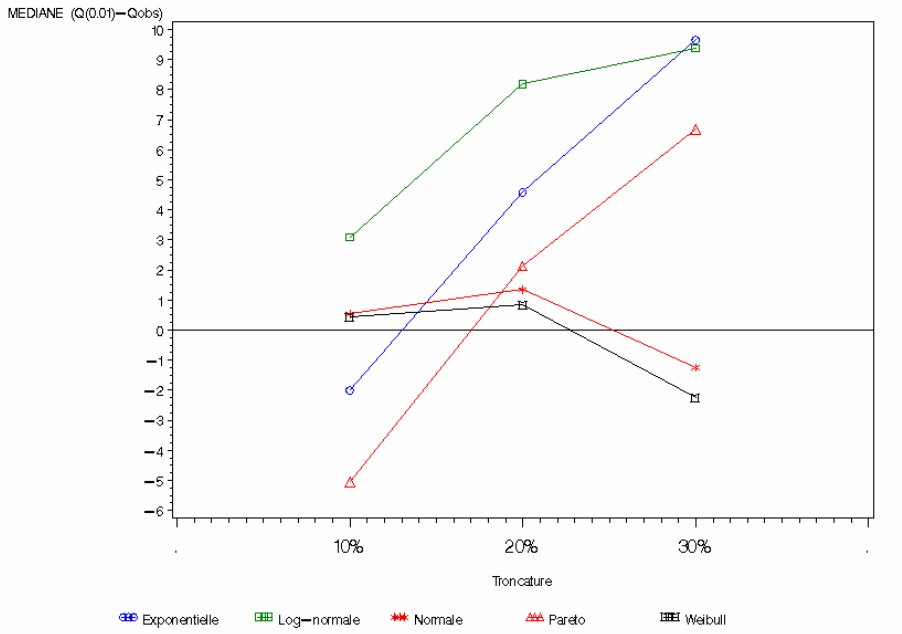


Figure II.44. Elément calcium : médiane des différences entres les quantiles estimés (0,01) et les quantiles observés pour les trois niveaux de troncature et pour les cinq distributions étudiées.

b. Etude de la variabilité des quantiles estimés

La figure II.45 présente la valeur de la médiane des différences entre les limites supérieures et inférieures des quantiles estimés en fonction du niveau de troncature pour les distributions normale, de Weibull et log-normale. La variabilité des quantiles estimés pour les distributions exponentielle et de Pareto étant très importante, ces distributions ne sont pas représentées.

Comme pour le carbone, **la variabilité est plus importante pour le niveau de troncature de 10%. Pour ce niveau de troncature, la variabilité la plus faible est rencontrée pour les distributions normale et de Weibull.**

Afin de connaître la relation entre le nombre d'observations par entité communale et la variabilité des quantiles estimés, la figure II.46 présente le graphique avec en abscisse le nombre d'observations et en ordonnée l'intervalle entre les limites supérieures et inférieures autour du quantile 0,01, respectivement pour les trois distributions normale, log-normale et de Weibull et les trois niveaux de troncature (les distributions exponentielle et de Pareto présentant un comportement trop différent ne sont pas présentées).

A partir de ces graphiques, il semble également, comme pour le carbone, qu'un nombre d'observations de 800 à 1000 sont nécessaires pour obtenir une estimation fiable des quantiles 0,01.

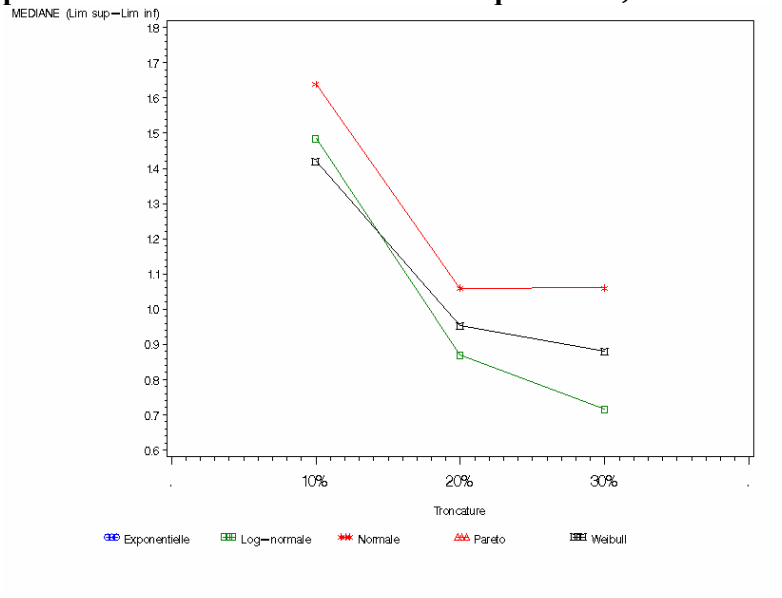


Figure II.45. Elément calcium : médiane des différences entre les limites supérieures et inférieures des quantiles (0,01) pour les trois niveaux de troncature pour les distributions normale, de Weibull et log-normale.

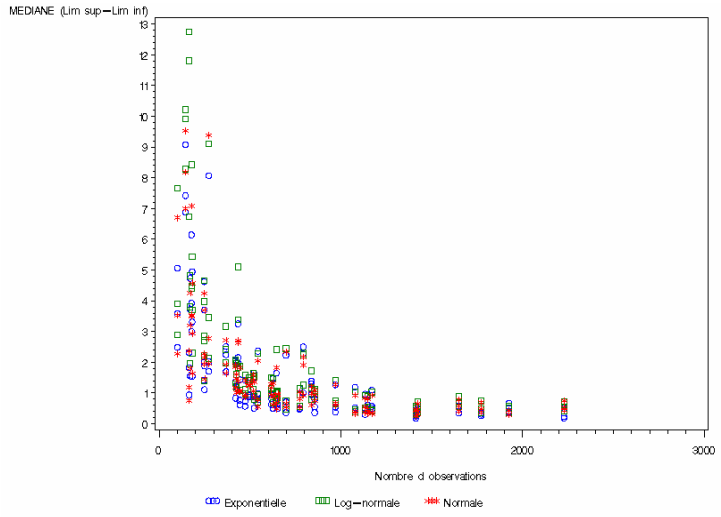


Figure II.46. Elément calcium : médiane des différences entre les limites supérieures et inférieures des quantiles (0,01) en fonction du nombre d'observations pour les distributions normale, log-normale et de Weibull et les trois niveaux de troncature.

Afin d'évaluer la relation entre les quantiles estimés et la variabilité de ces quantiles, la figure II.47 présente le graphique avec en abscisse la médiane des quantiles estimés et en ordonnée la médiane de l'intervalle entre les limites supérieures et inférieures.

A partir de cette figure, il semblerait que, pour les distributions exponentielle et de Pareto, plus les quantiles estimés sont faibles, plus l'intervalle autour de ces quantiles est élevé.

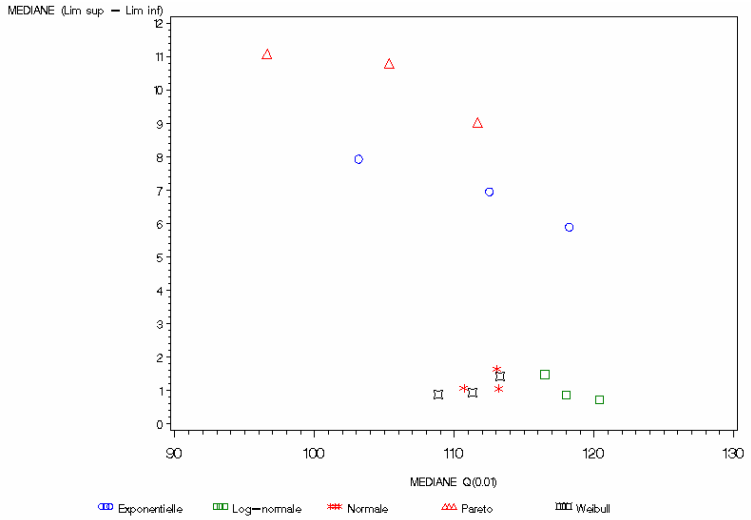


Figure II.47. Elément calcium: médiane des différences entre les limites supérieures et inférieures des quantiles (0,01) en fonction de la médiane des quantiles 0,01 pour les cinq distributions et les trois niveaux de troncature.

6.2.3. Etude de l'influence de la troncature sur l'estimation des quantiles 0,001

Nous présentons dans ce paragraphe la médiane des quantiles estimés (0,001) par niveau de troncature et par distribution (figure II.48).

Par rapport aux quantiles 0,01, les quantiles sont évidemment plus faibles. A nouveau, les distributions exponentielle et de Pareto présentent des quantiles peu stables ; de même, la distribution log-normale fournit des quantiles 0,001 peu stables en fonction du niveau de troncature ; **les distributions les plus stables étant les distributions normale et de Weibull.**

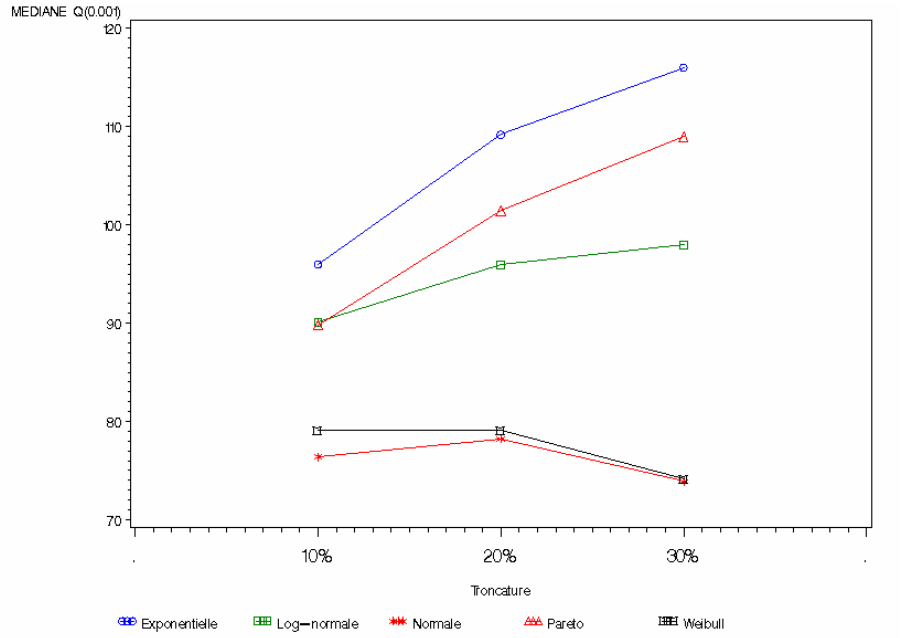


Figure II.48. Elément calcium (partie gauche) – Quantiles estimés (0,001) en fonction du niveau de troncature.

6.2.4. Evaluation de la qualité des quantiles estimés 0,001 par l'étude de leur variabilité

La figure II.49 présente la valeur de la médiane des différences entre les limites supérieures et inférieures des quantiles estimés en fonction du niveau de troncature, d'une part, pour les distributions normale, de Weibull et log-normale.

Les distributions qui présentent la variabilité la plus faible sont la distribution de Weibull et la log-normale. La distribution normale présente une variabilité quelque peu plus élevée. Comme pour les quantiles 0,01, la variabilité des quantiles est la plus élevée pour le niveau de troncature de 10%.

Comme pour le quantile 0,01, les graphiques présentant la relation entre le nombre d'observations par entité communale et la variabilité des quantiles estimés ont montré que 800 à 1000 observations sont nécessaires pour une estimation correcte des quantiles.

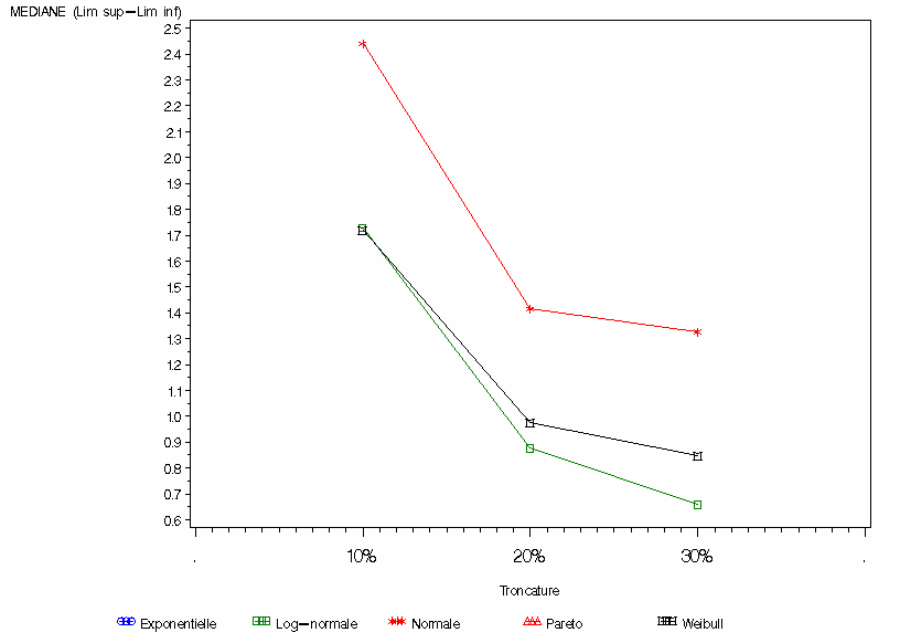


Figure II.49. Elément calcium : médiane des différences entre les limites supérieures et inférieures des quantiles (0,001) pour les trois niveaux de troncature et pour trois distributions étudiées.

Il est également intéressant d'évaluer la relation entre les quantiles estimés et la variabilité de ces quantiles. La figure II.50 présente le graphique avec en abscisse la médiane des quantiles estimés et en ordonnée la médiane de l'intervalle entre les limites supérieures et inférieures.

A partir de cette figure, il semblerait, contrairement à ce qui a été observé pour le quantile 0,01 que plus les quantiles estimés sont élevés, plus l'intervalle autour de ces quantiles est élevé.

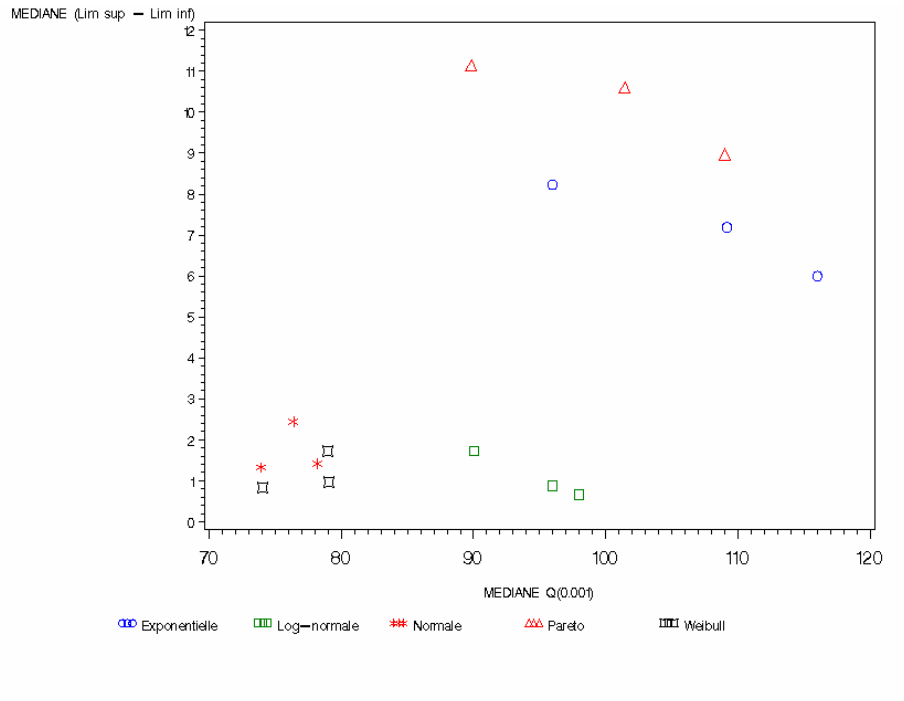


Figure II.50. Elément calcium : médiane des différences entre les limites supérieures et inférieures des quantiles (0,001) pour le niveau de troncature de 10% en fonction du nombre d'observations pour la distribution de Pareto.

6.3. Sélection de la distribution et du niveau de troncature pour la détection des valeurs aberrantes

A partir des études menées sur l'évaluation de la qualité des paramètres estimés (RMSE) et de la qualité de l'estimation des quantiles 0,01 et 0,001[26], il nous semble que **la distribution de Weibull avec un niveau de troncature de 10% est la plus intéressante**. Les arguments qui nous permettent de justifier notre choix sont les suivants :

1. En ce qui concerne la qualité des paramètres estimés, les distributions normale, log-normale et de Weibull présentent les valeurs de RMSE les plus faibles ; la distribution de Weibull présente, en terme de médiane, le RMSE le plus faible et le plus stable (figures II.36).
2. Le biais entre les quantiles estimés (quantiles 0,01) et les valeurs observées est le plus faible pour le niveau de troncature de 10% pour les distributions normale et de Weibull (figure II.44).
3. Pour le niveau de troncature de 10%, pour lequel le biais est le plus faible, la variabilité des quantiles est la plus faible pour la distribution de Weibull mais la variabilité des quantiles est la plus élevée pour le niveau de troncature de 10% par rapport aux autres niveaux de troncature (figure II.45).
4. Pour le quantile 0,001, la stabilité des quantiles est à nouveau rencontrée pour les distribution normale et de Weibull et plus particulièrement pour cette dernière (figure II.48).
5. La variabilité des quantiles 0,001 est plus faible pour la distribution de Weibull que la distribution normale (figure II.49).

La distribution normale présente cependant des caractéristiques presque identiques à la distribution de Weibull pour le même niveau de troncature. Elle n'a pas été sélectionnée car la variabilité est plus élevée pour les quantiles 0,01 et 0,001 (figures II.45 et II.49).

[26] Le coefficient de corrélation entre les quantiles estimés 0.01 et 0.001 est de 0.969.

6.4. Etude des propriétés de la distribution et du niveau de troncature sélectionnés (à partir des entités communales regroupées *a priori*)

6.4.1. Identification des valeurs aberrantes d'origine

Le tableau II.27 présente, pour le calcium, le pourcentage de valeurs aberrantes d'origine détectées à partir des quantiles estimés 0,01 sur base du nombre total d'observations (n=28.209) ; ceci est réalisé par distribution et niveau de troncature.

Tableau II.27. Elément calcium (partie gauche) : nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature et pourcentage par rapport au nombre total d'observations – quantile 0,01.

Niveau de troncature	Normale	Weibull	Lognormale	Exponentielle	Pareto
10%	268 (0,95)	**272 (0,96)**	329 (1,17)	210 (0,74)	148 (0,52)
20%	264 (0,94)	254 (0,90)	386 (1,37)	354 (1,25)	227 (0,80)
30%	249 (0,88)	224 (0,79)	426 (1,51)	456 (1,62)	304 (1,08)

On observe ainsi qu'à partir des quantiles estimés 0,01, le pourcentage de détection de valeurs aberrantes d'origine varie de 1,62, pour la distribution exponentielle (30%), à 0,52 pour la distribution de Pareto (10%).

La distribution de Weibull avec un niveau de troncature de 10% permet de retrouver le pourcentage de valeurs aberrantes d'origine le plus proche de 1% (0,96%) ; elle est suivie par la distribution normale. Ces deux distributions présentaient, pour ce niveau de troncature, le biais le plus faible.

Le tableau II.28 présente le nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature pour le quantile 0,001.

Tableau II.28. Elément calcium (partie gauche) : nombre de valeurs aberrantes d'origine détectées par distribution et niveau de troncature et pourcentage par rapport au nombre total d'observations – quantile 0,001.

Niveau de troncature	Normale	Weibull	Lognormale	Exponentielle	Pareto
10%	43 (0,15%)	**45 (0,16%)**	87 (0,31%)	139 (0,49%)	106 (0,38%)
20%	47 (0,17%)	44 (0,16%)	115 (0,41%)	277 (0,98%)	185 (0,66%)
30%	39 (0,14%)	42 (0,15%)	121 (0,43%)	406 (1,44%)	268 (0,95%)

On observe ainsi qu'à partir des quantiles estimés 0,001, le pourcentage de détection de valeurs aberrantes d'origine varie de 1,44%, pour la distribution exponentielle à un pourcentage de 0,14% pour la distribution normale ; toutes les deux pour un niveau de troncature de 30%.

Pour le niveau de troncature retenu qui est de 10%, les distributions normale et de Weibull présentent un taux de détection de valeurs aberrantes d'origine de 0,15 et 0,16%, pourcentages les plus proches de 0,1%.

6.4.2. Evaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles

Comme pour le carbone, nous cherchons à évaluer le taux de détection de valeurs aberrantes issues de la Famenne et l'Ardenne, en distinguant les contaminants issus de communes voisines de la Famenne ou non. Le tableau II.29 présente les pourcentages de détection de valeurs aberrantes suite à la comparaison de contaminants issus d'autres régions agricoles pour la distribution de Weibull avec le niveau de troncature de 10%. Les pourcentages de détection sont présentés pour chaque type de communes étudiées et pour les quantiles 0,01 et 0,001.

Tableau II.29. Elément calcium (partie gauche) : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution de Weibull et le niveau de troncature de 10% - quantiles 0,01 et 0,001.

Origine des contaminants détectés comme aberrants	Quantile 0,01	Quantile 0,001
Contaminants des communes de l'Ardenne	27,30 %	11,63 %
Contaminants des communes de la Famenne non voisines au Condroz	3,87 %	2,77 %
Contaminants des communes de la Famenne voisines au Condroz	4,33 %	1,40 %

A nouveau, comme dans le cas du carbone, le pourcentage de détection est plus faible dans les zones les plus proches de la région agricole du Condroz.

Le tableau II.30 présente les résultats pour les trois zones de la région agricole du Condroz (communes du Condroz voisines à la Famenne, communes centrales du Condroz, communes du Condroz voisines à la Région limoneuse).

On observe que, pour les communes du Condroz voisines de la Famenne, les communes centrales du Condroz et les communes du Condroz voisines à la Région limoneuse, la tendance générale est relativement identique, tant pour les quantiles 0,01 que pour les quantiles 0,001, c'est-à-dire une détection des contaminants comme valeurs aberrantes plus importante pour les contaminants de l'Ardenne et très faible pour ceux de la Famenne.

En ce qui concerne le quantile 0,01, pour les communes du Condroz voisines à la Famenne et pour les communes centrales du Condroz, les observations issues de l'Ardenne sont détectées avec un pourcentage de 22,6% tandis que pour les observations issues de la Famenne, on obtient un pourcentage de détection de l'ordre de 3%.

Tableau II.30. Elément calcium : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution de Weibull et le niveau de troncature de 10%, à partir des différentes entités regroupées *a priori* - quantiles 0,01 et 0,001 - distinction entre différentes zones de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	**Quantile 0,01**	**Quantile 0,001**
Etude des communes du Condroz voisines à la Famenne		
Contaminants de communes de l'Ardenne comparés dans les communes du Condroz voisines à la Famenne	22,6%	6,7%
Contaminants de communes de la Famenne non voisines au Condroz comparés dans les communes du Condroz voisines à la Famenne	2,5%	0,6%
Contaminants de communes de la Famenne voisines au Condroz comparés dans les communes du Condroz voisines à la Famenne	2,6%	0,4%
Etude des communes centrales du Condroz		
Contaminants des communes de l'Ardenne comparés dans les communes centrales du Condroz	23,7%	7,1%
Contaminants des communes de la Famenne non voisines au Condroz comparés dans les communes centrales du Condroz	2,7%	0,8%
Contaminants des communes de la Famenne voisines au Condroz comparés dans les communes centrales du Condroz	2,9%	0,6%
Etude des communes voisines à la Région limoneuse		
Contaminants des communes de l'Ardenne comparés dans les communes voisines à la Région limoneuse	35,6%	21,1%
Contaminants des communes de la Famenne non voisines au Condroz comparés dans les communes voisines à la Région limoneuse	6,4%	6,9%
Contaminants des communes de la Famenne voisines au Condroz comparés dans les communes voisines à la Région limoneuse	7,5%	3,2%

Pour les communes du Condroz voisines à la Région limoneuse, le taux de détection est de 35,6% pour les communes issues de l'Ardenne et de l'ordre de 7% pour les communes issues de la Famenne. Donc, plus les communes du Condroz sont éloignées des régions d'où sont issus les contaminants, plus le taux de détection est élevé. Cette constatation avait été réalisée également pour le carbone.

Pour le quantile 0,001, le taux de détection de valeurs aberrantes issues de l'Ardenne est le plus élevé pour les communes du Condroz voisines à la Région limoneuse (21,1%). Pour les autre communes ce pourcentage est plus faible et est de l'ordre de 7%.

Pour les contaminants de la Famenne, le taux de détection est le plus élevé à nouveau pour les communes voisines à la région limoneuse (6,9 et 3,2%). Pour les communes centrales du Condroz ou les communes voisines à la Famenne, le taux de détection est particulièrement faible (0,4 à 0,8%).

6.4.3. Evaluation du rapport d'efficacité

Comme pour le carbone, nous évaluons le rapport d'efficacité basé sur le quantile 0,001 pour la distribution de Weibull avec le niveau de troncature de 10%, en distinguant les contaminants issus de communes voisines de la Famenne ou non (tableau II.31).

Tableau II.31. Elément calcium (partie gauche) : rapport d'efficacité calculé sur base du quantile 0,001 pour la distribution de Weibull et le niveau de troncature de 10% à partir de l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	Rapport d'efficacité
Contaminants des communes de l'Ardenne	74,5
Contaminants des communes de la Famenne non voisines au Condroz	17,7
Contaminants des communes de la Famenne voisines au Condroz	9,0

6.5. Ajustements et évaluation de la qualité des paramètres estimés (pour l'ensemble des données du Condroz)

L'ajustement a été réalisé, comme pour l'élément carbone, à partir de l'ensemble des observations de la région agricole du Condroz. Pour le niveau de troncature de 10%, le nombre d'observations est de 2881.

La valeur de RMSE obtenue pour la distribution de Weibull est de 4,99 alors que la médiane du RMSE, calculée dans la première partie de ce travail, était de 3,75. La qualité de l'ajustement est donc moins bonne lorsque les paramètres sont ajustés de manière globale et non par entité communale.

6.6. Etude des propriétés de la distribution et du niveau de troncature sélectionnés (à partir de l'ensemble des données du Condroz)

6.6.1. Identification des valeurs aberrantes d'origine

Le nombre de valeurs aberrantes d'origine détectées à partir des valeurs limites estimées sur l'ensemble des observations est présenté au tableau II.32.

Pour le quantile 0,01, le pourcentage de détection était de 0,96% (tableau II.27), le pourcentage était donc un peu plus précis que dans ce cas-ci (1,07%). Pour le quantile 0,001, le pourcentage était de 0,16% et il est passé à 0,29% en calculant de manière globale, ce qui est nettement moins bon.

Tableau II.32. Elément calcium : nombre de valeurs aberrantes d'origine détectées à partir de l'ensemble des observations de la région agricole du Condroz sur base des quantiles 0,01 et 0,001.

6.6.2. Evaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles

Le taux de détection de valeurs aberrantes issues d'autres régions agricoles est présenté au tableau II.33. Pour le quantile 0,01 et 0,001, les taux de détection sont bien plus faibles lorsqu'on travaille de manière globale en comparaison aux taux de détection obtenus en travaillant par entité communale (figure II.30).

Tableau II.33. Elément calcium : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles, calculé à partir des quantiles 0,01 et 0,001 estimés sur l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	Quantile 0,01	Quantile 0,001
Contaminants de communes de l'Ardenne	19,12%	3,43%
Contaminants de communes de la Famenne non voisines au Condroz	2,11%	0,25%
Contaminants de communes de la Famenne voisines au Condroz	1,84%	0,22%

6.6.3. Evaluation du rapport d'efficacité

Le rapport d'efficacité basé sur le quantile 0,001, pour la distribution de Weibull et le niveau de troncature de 10%, est présenté au tableau II.34.

A partir de ce tableau, on observe qu'en travaillant de manière globale, le rapport d'efficacité diminue de manière significative. En effet, pour les contaminants issus de l'Ardenne, obtenus à partir des entités communales séparées, le rapport d'efficacité était de 74,5 alors qu'il est dans ce cas-ci de 12,1. Pour les contaminants de la Famenne, les rapports d'efficacité obtenus à partir des entités communales séparées, respectivement pour ceux issus des communes non voisines et voisines au Condroz, étaient de 17,7 et de 9,0, ils se réduisent ici à 0,7 ce qui est vraiment très faible.

Tableau II.34. Elément calcium : rapport d'efficacité basé sur le quantile 0,001 pour la distribution de Weibull et le niveau de troncature de 10% et pour l'ensemble des observations de la région agricole du Condroz.

Origine des contaminants détectés comme aberrants	Rapport d'efficacité
Contaminants de communes de l'Ardenne	12,1
Contaminants de communes de la Famenne non voisines au Condroz	0,7
Contaminants de communes de la Famenne voisines au Condroz	0,7

6.7. Résultats obtenus à partir des limites actuelles de REQUASUD

La valeur limite utilisée au sein du réseau RéQuaSud pour les terres de culture de la région agricole du Condroz et qui correspond au quantile 0,001 est de 81,2.

En utilisant la limite de RéQuaSud, le nombre de valeurs aberrantes d'origine est de 112.

Les taux de détection obtenus à partir de cette limite sont présentés au tableau II.35 et les rapports d'efficacité sont exposés au tableau II.36.

A partir de ces tableaux, on observe des taux de détection nettement plus faibles que lorsqu'on travaille à partir des entités communales regroupées *a priori*. En ce qui concerne les rapports d'efficacité, pour les contaminants de l'Ardenne, le rapport est encore plus faible que celui obtenu à partir de l'ensemble des observations de la région agricole du Condroz. Pour les contaminants issus de la Famenne, les rapports sont quelque peu plus élevés mais bien plus faibles que ceux obtenus à partir des entités communales séparées.

La discussion globale des résultats est effectuée au paragraphe 6.9.

Tableau II.35. Elément calcium : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles à partir des valeurs limites actuelles du réseau RéQuaSud pour l'ensemble de observations du Condroz.

Origine des contaminants détectés comme aberrants	% de détection à partir des limites de RéQuaSud
Contaminants de communes de l'Ardenne	2,9%
Contaminants de communes de la Famenne non voisines au Condroz	0,7%
Contaminants de communes de la Famenne voisines au Condroz	0,5%

Tableau II.36. Elément calcium : rapports d'efficacité calculés à partir des valeurs limites actuelles du réseau RéQuaSud pour l'ensemble de observations du Condroz.

Origine des contaminants détectés comme aberrants	Rapport d'efficacité
Contaminants de communes de l'Ardenne	7,5
Contaminants de communes de la Famenne non voisines au Condroz	1,8
Contaminants de communes de la Famenne voisines au Condroz	1,3

6.8. Classification spatiale

6.8.1. Evaluation du rapport d'efficacité

Dans le cas du calcium, le regroupement qui conduit à l'effectif proche de 800 observations correspond à 8 groupes. Afin d'appliquer la même démarche que pour le carbone, 2, 3, … 9 groupes ont été créés à partir des quantiles extrêmes estimés.

Suite à l'identification des valeurs aberrantes d'origine et à l'évaluation du taux de détection de valeurs aberrantes issues d'autres régions agricoles (annexe 8), les rapports d'efficacité ont été calculés. Les résultats relatifs aux contaminants de l'Ardenne sont présentés à la figure II.51. Comme pour le carbone, des tendances relativement semblables sont observées pour les contaminants des différentes régions étudiées.

Par le fait du hasard, le rapport d'efficacité le plus intéressant est également rencontré pour le regroupement menant à 8 groupes d'entités communales.

Le rapport d'efficacité obtenu pour les 8 groupes (tableau II.37) est à nouveau plus important que lorsqu'un seul groupe est pris en compte (ensemble du Condroz). Par rapport aux résultats de RéQuaSud, le rapport d'efficacité des 8 groupes est encore plus élevé.

Comme pour le carbone, une perte d'efficacité est observée lors du regroupement. En effet, les résultats obtenus à partir des 39 communes séparées indiquent que le rapport d'efficacité obtenu est plus élevé (74,5 dans le cas des contaminants issus de l'Ardenne) que pour les 8 groupes (49,8).

Tableau II.37. Elément calcium : évolution du rapport d'efficacité en fonction du nombre de groupes formés.

	Contaminants de communes de l'Ardenne	**Contaminants de communes de la Famenne non voisines au Condroz**	**Contaminants de communes de la Famenne voisines au Condroz**
1 groupe	12,1	0,7	0,7
8 groupes	49,8	5,8	5,2
39 entités	74,5	17,7	9,0
RéQuaSud	7,5	1,8	1,3

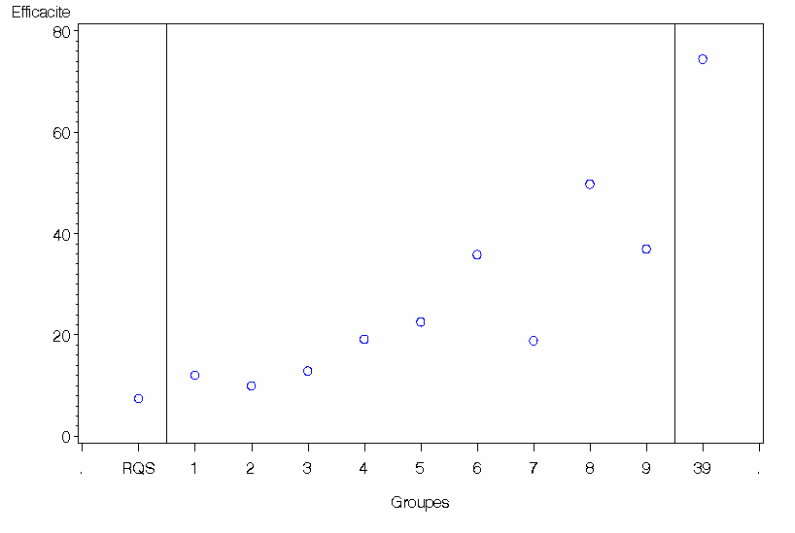


Figure II.51. Elément calcium : évolution du rapport d'efficacité en fonction du nombre de groupes formés - comparaison pour les contaminants issus de l'Ardenne.

6.8.2. Représentation graphique des groupes d'entités communales

Le regroupement constitué de 8 groupes est présenté à la figure II.52 et les quantiles estimés 0,001 par groupe sont présentés au tableau II.38.

Nous étudions ici la partie gauche des distributions et les quantiles extrêmes correspondent donc aux valeurs maximales au-dessous desquelles les valeurs de calcium devraient être rejetées.

Pour le groupe 1, formé de l'entité de Charleroi et des communes voisines, la valeur limite est bien plus faible que les autres groupes ; des valeurs plus faibles de calcium y seront donc acceptées. Ceci peut notamment s'expliquer par l'absence de bandes calcaires typiques du Condroz et par la présence de bandes de limons plus importantes qui pourraient entraîner un drainage du sol plus faible. Les pratiques culturales, liées par exemple à un chaulage moins important, peuvent également constituer un des facteurs explicatifs d'une moindre teneur en calcium.

A partir de la figure II.52, deux zones homogènes sont observées ; elles correspondent :

1. à la région agricole limitrophe à la Famenne (groupe 2) ;
2. à la région limitrophe à la Région limoneuse (groupe 3).

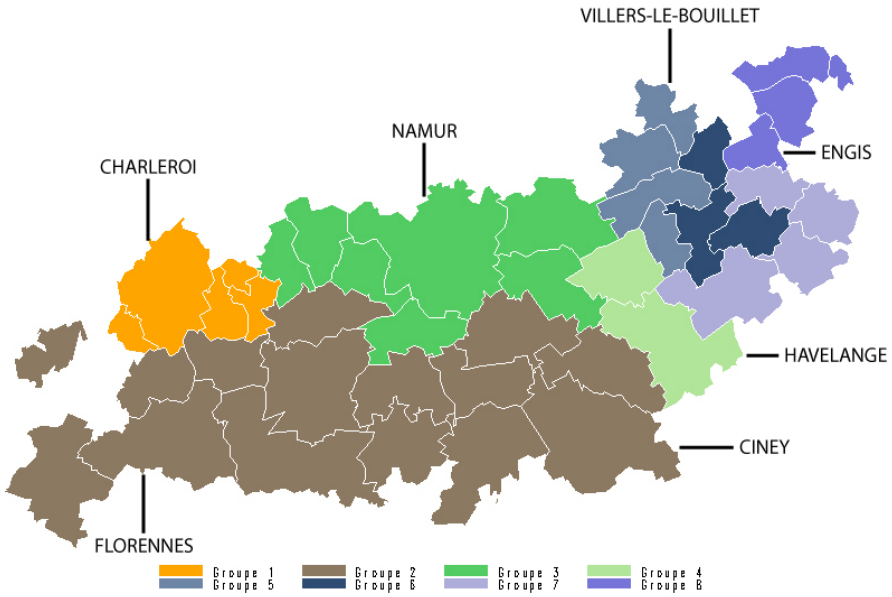


Figure II.52. Elément calcium : création de 8 groupes *a posteriori.*

Tableau II.38. Elément calcium : effectifs et quantiles estimés 0,001 pour les 8 groupes formés.

Groupes	n	Quantiles estimés 0,001
1	338	2,89
2	11360	71,25
3	3482	56,06
4	2385	60,20
5	5229	95,55
6	1533	89,31
7	3390	83,91
8	1092	122,28

La valeur des quantiles estimés dans les groupes 2 et 3 est relativement différente, avec des quantiles plus élevés pour la région proche de la Famenne. Ceci pourrait s'expliquer par la présence, dans le groupe 2, de types de sols plus particuliers tels que des sols à charge schisteuse avec un bon drainage tandis que, dans le groupe 3, une alternance de bandes psammitiques et calcaires avec des placages limoneux est observée. Les pratiques culturales ainsi que l'historique cultural pourraient également expliquer ces différences. Il faut également signaler que la zone 3 est limitrophe à la région limoneuse où la teneur en calcaire pourrait être plus importante.

Le groupe 4 est constitué de deux entités à caractère très rural qui sont Ohey et Havelange, qui présentent un quantile extrême inférieur au groupe 3.

La zone du Condroz proche de la Région herbagère est également morcelée en 4 groupes distincts (groupes 5, 6, 7 et 8). Ce dernier groupe est formé des mêmes entités que pour le carbone. Les quantiles extrêmes sont supérieurs aux quantiles des groupes présentés ci-dessus. Ceci pourrait correspondre à des zones particulières comprenant des placages limoneux ou à un chaulage plus important. En effet, les pratiques culturales telle que des cultures plus intensives peuvent expliquer ces quantiles extrêmes plus élevés. Ceci peut être mis en parallèle avec les valeurs plus faibles de carbone observée précédemment.

En partant du groupe 5 au groupe 7, c'est-à-dire du Nord au Sud, une diminution du quantile extrême est observée. L'éloignement par rapport à la région limoneuse pourrait également être un élément de réponse à cette observation.

Le dernier groupe, constitué d'Engis, Flémalle, Grâce-Hollogne et Saint-Nicolas présente une valeur limite très élevée par rapport aux autres groupes. Pour ce même groupe, les teneurs en carbone étaient pourtant moyennes ; on aurait donc pu s'attendre à des teneurs moyennes en calcium, ce qui n'est pas le cas. Le caractère industriel de la région pourrait éventuellement être mis en cause.

6.9. Validation de la méthode de détection par comparaison aux résultats de REQUASUD

A nouveau, comme pour le carbone, lorsque les rapports d'efficacité de la méthode de détection de valeurs aberrantes obtenus par entité communale (paragraphe 6.4), par groupe d'entités communales (paragraphe 6.8 – figure II.51) et pour l'ensemble des observations de la région agricole du Condroz (paragraphe 6.6) par rapport à ceux obtenus à partir des valeurs limites de RéQuaSud (paragraphe 6.7) sont comparés, les observations suivantes peuvent être formulées tant pour l'Ardenne que pour la Famenne.

Les résultats obtenus à partir de la limite de RéQuaSud sont les moins satisfaisants en terme de rapport d'efficacité. Ceux obtenus à partir de l'ensemble des observations sont légèrement supérieurs à ceux obtenus à partir des limites de RéQuaSud.

Par contre, à partir des différents groupes d'entités communales formés *a posteriori* ou des entités regroupées *a priori*, les rapports d'efficacité calculés sont plus élevés que ceux obtenus pour RéQuaSud ou pour l'ensemble des observations. On observe comme pour le calcium, un gain d'efficacité pour la détection de valeurs aberrantes.

DISCUSSION GENERALE

La méthode de détection de valeurs aberrantes développée à partir d'échantillons de sols a permis de tenir compte de la problématique exposée au début de ce travail. En effet, le mélange de distributions dissymétriques en présence d'une contrainte de contiguïté liée à la présence de différents types de sols au sein de communes voisines ont été pris en compte. De même, la recherche des valeurs aberrantes à droite (à partir du carbone organique) et à gauche (à partir du calcium) des distributions dissymétriques a été examinée.

Des valeurs limites correspondant aux quantiles extrêmes de distributions dissymétriques différentes à droite et à gauche ont été estimées de manière optimale à partir des queues de distributions afin de tenir compte du problème de mélange de distributions et en testant différents niveaux de troncature. C'est à partir de ces valeurs limites finalement déterminées par groupement géographique suite à la classification spatiale que les nouveaux échantillons de sols vont faire l'objet d'un contrôle.

La mise en place de cette méthode de détection de valeurs aberrantes a nécessité l'identification d'une distribution dissymétrique la plus adéquate à droite et à gauche ainsi que le nombre optimal d'observations à prendre en compte au travers de différents niveaux de troncature (10%, 20% et 30%). Cinq distributions ont fait l'objet de cette étude : les distributions normale, de Weibull, log-normale, exponentielle et de Pareto. La sélection, d'une part, de la distribution décrivant le mieux le comportement de la queue de la distribution à droite ou à gauche et, d'autre part, du niveau de troncature est fondée sur différents critères basés sur la qualité d'estimation des paramètres des distributions et la qualité d'estimation des quantiles extrêmes. La distribution la plus intéressante est celle qui présente :

- en terme d'ajustement des distributions et d'estimation des paramètres, le RMSE (paragraphe 2.3.3) le plus faible et le plus stable quel que soit le niveau de troncature ;
- en terme de qualité d'estimation des quantiles, des quantiles estimés stables par rapport au niveau de troncature ;
- le biais entre les quantiles estimés et les quantiles observés le plus faible et le plus stable par rapport au niveau de troncature ;
- la variabilité des quantiles la plus faible et la plus stable par rapport au niveau de troncature.

Le niveau de troncature optimal sélectionné correspond à celui pour lequel le biais entre les quantiles estimés et observés est le plus faible.

Pour la partie droite des queues de distributions, la distribution, qui a été sélectionnée à partir des études menées sur l'évaluation de la qualité de l'estimation des paramètres des distributions dissymétriques et des quantiles 0,99 et 0,999, est la distribution exponentielle avec un niveau de troncature

de 10%. La première raison qui nous a poussé à faire ce choix concerne la qualité de l'estimation des paramètres ; la distribution exponentielle présente les valeurs de RMSE, en terme de médiane, les plus faibles, la distribution de Pareto étant également très intéressante pour ce critère. Ensuite, lorsqu'on compare les résultats obtenus avec la distribution exponentielle par rapport à ceux obtenus avec la distribution de Pareto, on observe que la variabilité des quantiles 0,99 et 0,999 est plus faible pour les quantiles estimés à partir de la distribution exponentielle et que les quantiles 0,999 sont stables par rapport aux trois niveaux de troncature. Le niveau de troncature de 10% a été sélectionné car le biais entre les quantiles estimés (0,99) et les valeurs observées est le plus faible pour ce niveau de troncature quelle que soit la distribution avec une légère sous-estimation pour la distribution exponentielle. Il est finalement logique que ce niveau de troncature ait été choisi car l'estimation est basée un nombre plus faible d'observations et est donc plus précise. Il aurait peut-être été intéressant de considérer le niveau de troncature de 20% afin d'avoir immédiatement un effectif suffisant pour l'application de la classification spatiale et d'obtenir une meilleure précision pour l'estimation des quantiles.

Pour la partie gauche des queues de distributions, la distribution identifiée comme la plus adéquate pour le calcium, est la distribution de Weibull, également avec un niveau de troncature de 10%. Premièrement, concernant la qualité de l'estimation des paramètres, la distribution de Weibull présente le RMSE le plus faible et le plus stable par rapport au niveau de troncature. Ensuite, la stabilité des quantiles 0,001 et la faible variabilité des quantiles 0,01 et 0,001 nous a menés à choisir cette distribution. La distribution normale, présentant des caractéristiques relativement semblables à la distribution de Weibull pour le même niveau de troncature, n'a pas été sélectionnée car la variabilité des quantiles est plus élevée. Comme pour le carbone, le niveau de troncature de 10% a également été sélectionné car le biais entre les quantiles estimés (0,01) et les valeurs observées est le plus faible.

L'étude de la variabilité des quantiles estimés a montré, tant pour le carbone que pour le calcium, que 800 à 1000 observations sont nécessaires pour estimer de manière fiable les quantiles extrêmes.

Suite à la sélection de distributions et du niveau de troncature pour les parties droite et gauche, les caractéristiques des distributions ont été étudiées tant au niveau de la détection de valeurs aberrantes initialement présentes dans le sous-ensemble de données qu'au niveau de la comparaison de contaminants issus d'autres régions agricoles par rapport aux valeurs limites estimées.

L'identification de valeurs aberrantes d'origine a permis de nous conforter dans la sélection des distributions et du niveau de troncature commun car les pourcentages de valeurs aberrantes d'origine à détecter, soit de 1% ou de

0,01%, respectivement pour les niveaux de quantiles 0,99-0,01 et 0,999-0,001, sont les plus intéressants pour les distributions choisies.

En ce qui concerne la comparaison de contaminants issus d'autres régions agricoles, c'est-à-dire dans notre cas, la Famenne et l'Ardenne, aux valeurs limites et leur détection comme étant des valeurs aberrantes, plusieurs remarques peuvent être émises. La détection de valeurs aberrantes est de plus en plus importante au fur et à mesure qu'on s'éloigne de la région agricole du Condroz, avec par exemple, pour les contaminants de l'Ardenne, un taux de détection de 18,73% (quantile 0,999) pour le carbone et de 11,63% (quantile 0,001), pour le calcium. Les résultats indiquent donc des différences importantes entre les taux de détection de contaminants issus des communes situées en bordure du Condroz, c'est-à-dire en Famenne ou en Ardenne. Ceci était attendu étant donné que les différences de niveaux de carbone ou de calcium étaient plus importants en Ardenne qu'en Famenne.

A l'intérieur même du Condroz, des différences au niveau de la détection des valeurs aberrantes issues de la Famenne ou de l'Ardenne, entre les communes centrales du Condroz, les communes situées au Nord (communes voisines à la Région limoneuse) et celles au Sud du Condroz (communes voisines à la Famenne) sont observées. Pour le carbone, le taux de détection le plus élevé est rencontré pour les communes centrales du Condroz. Pour le calcium, c'est dans les communes voisines à la Région limoneuse que ce taux est plus élevé ; ce qui est le plus logique car la distance par rapport à la Famenne ou à l'Ardenne est la plus élevée. En l'absence de classification spatiale, différentes zones pourraient donc être facilement imaginées : la région centrale et les régions comprenant les entités communales voisines à d'autres régions agricoles. Ceci pourrait constituer une alternative simple de création de groupes à l'intérieur d'une région agricole ou de tout autre ensemble géographique.

La classification spatiale a ensuite été appliquée sur base des quantiles extrêmes (0,999 et 0,001) et d'une matrice de contiguïté prenant en compte le voisinage des entités communales et le pourcentage de superficie des types de sols. Dans notre cas, nous avons travaillé avec l'ensemble des types de sols de l'entité ; il aurait cependant été plus correct de travailler avec les types de sols rencontrés exclusivement pour les terres de culture à partir desquelles les échantillons ont été extraits. De plus, la construction de la matrice de contiguïté pourrait être envisagée différemment en considérant d'autres critères que le pourcentage de superficie des types de sols.

Suite à cette classification spatiale, des entités communales présentant des quantiles extrêmes similaires et également un effectif suffisant pour estimer les quantiles de manière fiable ont été regroupées. Le critère de l'effectif minimal a été appliqué dans la mesure où il n'a pas modifié de manière trop importante la constitution des groupes. En effet, le critère lié à l'effectif minimum n'a pas été rencontré pour un groupe d'entités communales car

celui-ci présentait des caractéristiques trop différentes des autres entités ; il a cependant été respecté pour les autres groupes formés.

La comparaison entre la capacité de détection des différentes unités géographiques (ensemble de la région agricole du Condroz, 8 groupes d'entités, 39 entités communales) a été réalisée sur base du rapport d'efficacité, défini à partir du rapport du pourcentage de détection de valeurs anormales dans les échantillons avec et sans introduction de valeurs anormales ; les méthodes étant jugées d'autant meilleures que le rapport d'efficacité est élevé. Lors de la classification spatiale, des rapports d'efficacité assez variables ont été observés d'un regroupement à l'autre ; ils dépendent de la présence de l'une ou l'autre entité communale 'influente' dans les groupes formés. Comme cité dans la partie bibliographique, il est important de vérifier les groupes obtenus par différentes méthodes de classification afin de s'assurer de la cohérence des résultats. Dans le cas de la classification avec contraintes spatiales, l'éventail des logiciels est cependant limité et nous a permis de ne tester que la méthode des *k-means* avec le logiciel R.

Pour le carbone et le calcium, par le fait du hasard, le rapport d'efficacité le plus intéressant, entre les différents regroupements de 3, 4, …, 9 groupes, est rencontré pour le regroupement menant à 8 groupes d'entités communales. Ces groupes ne sont cependant pas forcément constitués par des entités identiques.

Tant pour le carbone que pour le calcium, les rapports d'efficacité, obtenus pour les 8 groupes d'entités ou pour les 39 entités communales, sont nettement supérieurs, comme cela était prévisible, aux rapports d'efficacité obtenus lorsqu'un seul groupe est pris en compte (ensemble du Condroz). En ce qui concerne l'intérêt de ne plus travailler de manière globale sur l'ensemble de la région agricole du Condroz, les résultats obtenus ont prouvé un gain d'efficacité non négligeable. La prise en compte de la contrainte spatiale est donc justifiée.

Lorsque les résultats obtenus pour les 8 groupes sont comparés par rapport à ceux obtenus à partir des 39 communes distinctes, on observe que le regroupement conduit à une perte d'efficacité. Comme le souligne Foguenne (1995), si l'inclusion de la contrainte de contiguïté se traduit par la création de zones non morcelées et par conséquent par une représentation plus facile des résultats obtenus, cette méthode conduit à une perte d'homogénéité des groupes formés. A l'issue de la classification spatiale, comme le rapport d'efficacité obtenu pour les entités communales séparées est plus élevé qu'à partir des groupes obtenus, il est légitime de se demander s'il est finalement nécessaire de réaliser le regroupement des entités communales. En réalité, la décision de regrouper ou non devrait être prise en fonction des situations rencontrées. Par exemple, dans le cas de la base de données SOLS de RéQuaSud, celle-ci va s'enrichir continuellement à partir des nouvelles observations envoyées par les laboratoires. Les

quantiles extrêmes vont être révisés annuellement afin d'en améliorer l'estimation. En travaillant avec les entités individuelles, le risque de modifications nettes des quantiles, d'une année à l'autre, n'est pas négligeable tant que le nombre d'observations par entité n'est pas atteint. De plus, lorsque trop peu de données sont disponibles pour certaines entités géographiques, la variabilité à l'intérieur de celles-ci peut être élevée et peut conduire à des erreurs de classement. En travaillant par groupes d'entités, les modifications des limites d'une révision à l'autre seraient moins importantes. Le rejet de certaines observations entre chaque révision serait alors moins difficile à expliquer aux laboratoires d'analyses.

La détection de valeurs aberrantes peut dès lors s'envisager suivant le schéma suivant. Dans un premier temps, lorsque très peu d'observations sont disponibles, l'estimation de la valeur limite est réalisée à partir de l'ensemble des données. Lorsque la base de données s'étoffe, des groupes sont créés en fonction du critère de l'effectif minimum et de la contrainte spatiale. Enfin, lorsque la base de données comprend assez d'observations pour chaque entité, les paramètres sont estimés par unité géographique. La valeur des quantiles extrêmes va donc être modifiée au cours du temps dans le but d'obtenir le quantile le plus précis et robuste possible. Cette manière de travailler pourrait être appliquée, par exemple, dans le cas de la base de données d'échantillons de sols qui comprendrait les observations sur les éléments traces métalliques (FUSAGx – Laboratoire de Géopédologie) pour laquelle peu d'observations sont disponibles à l'heure actuelle.

Nous avons donc, grâce à la méthodologie mise en place, constitué un référentiel par groupe d'entités communales, constitué des valeurs limites de référence par groupe, qui va évoluer en fonction des révisions des valeurs limites ultérieures.

La classification spatiale a ainsi mené à la constitution de différents groupes qui, dans le cas du carbone, peuvent se présenter sous la forme de trois zones homogènes. La première correspond au nord de la région agricole de la Famenne, la deuxième à la région limitrophe à la Région limoneuse et la troisième, au nord-est de la région agricole de la Famenne. Dans le cas du calcium, deux zones homogènes sont observées. La première concerne la région agricole limitrophe à la Famenne et la seconde correspond à la région limitrophe à la Région limoneuse ; ce qui correspond en quelque sorte au regroupement 'grossier' cité précédemment dans ce chapitre. Pour les groupes situés plus à l'Est, et en passant du Nord au Sud, une diminution du quantile extrême est observée. L'éloignement par rapport à la région limoneuse pourrait être un élément de réponse à cette observation.

La représentation spatiale obtenue pour le calcium est cependant relativement semblable à celle du carbone, excepté pour un groupe de 3 communes de la région Nord-Est du Condroz. Ceci peut s'expliquer par la relation qui pourrait exister entre ces deux variables. En effet, un taux de calcium faible entraînerait un pH faible, la minéralisation serait alors plus

limitée et la teneur en carbone serait plus élevée. Ce raisonnement est justifié lorsqu'on travaille au niveau de la parcelle agricole ou typiquement en milieu forestier ; cependant au niveau de terres agricoles, d'autres composantes interviennent. En effet, les différences entre groupes peuvent s'expliquer soit par des méthodes de culture différentes avec un chaulage plus important d'une zone à l'autre liée aux habitudes régionales, soit par un passé prairial plus important d'un groupe à l'autre. L'interprétation basée uniquement sur la présence d'un ou plusieurs types de sols particuliers pourrait éventuellement s'élargir à l'ensemble des informations agro-pédologiques telles que le type d'exploitation agricole (présence de bétails ou non, culture intensive sans apport de matière organique, etc.).

En ce qui concerne la validation par rapport aux résultats obtenus à partir des limites de RéQuaSud, les résultats de RéQuaSud pour le carbone sont plus intéressants que ceux obtenus à partir de l'ensemble des observations tandis que pour le calcium, c'est l'inverse. Ensuite, tant pour le carbone que pour le calcium, les rapports d'efficacité calculés à partir des différents groupes d'entités communales formés *a priori* ou à partir des entités regroupées *a posteriori* sont plus élevés que ceux obtenus pour RéQuaSud. Par exemple, pour le carbone, le rapport d'efficacité obtenu à partir des 39 entités communales distinctes est 2,5 fois supérieur à celui obtenu en prenant les limites de RéQuaSud. Les résultats étant semblables que ce soit pour les contaminants issus de l'Ardenne ou de la Famenne. Pour le calcium, le rapport d'efficacité à partir des 39 entités est 10 fois supérieur aux limites de routine. Au vu de ce qui vient d'être exposé, à partir du sous-ensemble de données utilisé dans l'application, la méthode de détection mise en place nous a donc permis d'apporter un gain d'efficacité pour la recherche de valeurs aberrantes par rapport à la méthode appliquée en routine.

CONCLUSIONS ET PERSPECTIVES

Tout au long de ce travail, différents éléments de réflexion ont été avancés et discutés et nous permettent de tirer des conclusions sur la détection de valeurs aberrantes liée à la problématique définie initialement.

Comme nous l'avons montré dans l'introduction générale, la détection de valeurs aberrantes pose un certain nombre de difficultés, en particulier en présence de mélanges de distributions et lorsqu'il est nécessaire d'assurer une cohérence spatiale. L'objet de notre étude a été de développer une méthode de détection de valeurs aberrantes lorsque des distributions dissymétriques se présentent sous la forme de mélanges et qu'une contrainte de contiguïté spatiale est à prendre en compte. Le caractère opérationnel de la méthodologie à mettre en place a également été un point important à prendre en considération.

Les différents problèmes exposés ne sont pratiquement pas abordés dans la littérature. En effet, la revue bibliographique en relation avec la recherche de valeurs aberrantes nous a poussé à abandonner l'idée d'appliquer des tests de discordance. En effet, ces tests ne permettent en général de ne considérer qu'une ou deux, voire trois valeurs simultanément alors que dans de grandes bases de données, il peut y en avoir beaucoup plus. Ceci nous a donc mené à proposer des valeurs limites, telles que les quantiles extrêmes, valeurs au-delà ou en deçà desquelles les observations sont considérées comme aberrantes.

Nous avons également constaté, en parcourant la littérature, le manque d'informations relatives aux mélanges de populations et au problème de la contamination des distributions. Ce problème est traité de manière très théorique et ne permet pas vraiment d'application concrète sur des données pour lesquelles la forme des distributions et les proportions des mélanges ne sont pas connues *a priori*.

En ce qui concerne la théorie relative aux distributions fortement dissymétriques, l'étude bibliographique nous a montré qu'il était possible d'extrapoler le comportement de la queue de la distribution des données à partir des observations les plus élevées ou les plus faibles. Les distributions dissymétriques les plus classiques et les plus intéressantes ont été étudiées et ont montré les potentialités qu'elles pouvaient apporter à l'étude des queues de distribution à droite ou à gauche.

Au niveau de la prise en compte de la contrainte spatiale, nous avons mis en évidence que l'utilisation d'une matrice de contiguïté créée de manière adéquate permet, à l'aide de la méthode de la classification non hiérarchique des *k-means*, de créer des groupes d'entités géographiques à l'intérieur desquels la cohérence spatiale est respectée. Le choix de la matrice de

contiguïté est important et doit mener à la définition la plus optimale possible de la contiguïté entre des unités géographiques.

De plus, le fait de prendre en compte les unités géographiques voisines peut également être remis en question ; en effet, des zones géographiques similaires pourraient être réunies sans forcément être voisines. Ceci dépend en réalité des objectifs poursuivis dans le domaine d'étude concerné.

Nous avons, au travers de ce travail, proposé différentes solutions originales aux problèmes posés. Une des principales particularités de la méthodologie développée consiste en la prise en compte des queues droite et gauche des distributions dissymétriques afin d'éluder le problème du mélange et en la fixation des valeurs limites au-dessus ou en dessous desquelles les valeurs sont jugées comme étant aberrantes. Ces valeurs limites sont déterminées à partir des paramètres des queues des distributions dissymétriques. Nous avons montré aussi que l'utilisation de distributions dissymétriques différentes à droite et à gauche des distributions est mieux adaptée à la détection des valeurs aberrantes que lorsqu'une même distribution est considérée. Ceci correspond en une démarche tout à fait innovante.

Nous avons prouvé également que la mise en place d'un système de détection de valeurs aberrantes en intégrant la contrainte spatiale, est plus cohérente que les méthodes qui considèrent que les populations sont homogènes dans l'espace. Ceci a été vérifié par le gain d'efficacité lors la détection de valeurs aberrantes. Les hypothèses présentées au début du travail ont donc été vérifiées.

La classification spatiale a mené, dans le cas des analyses de sols, à la constitution de différents groupes qui peuvent se répartir en des zones plus ou moins homogènes. Les différences observées entre les différentes régions peuvent être attribuées bien évidemment à la présence des différents types de sols mais également aux techniques culturales (chaulage, apports d'intrants organiques d'origine différente, diversité des spéculations, etc.), au mode de gestion des parcelles, aux pratiques culturales antérieures, aux modifications anthropomorphiques mais également à la situation de l'entité géographique en partie sur une région agricole voisine et au problème de la représentativité des échantillons dans les groupes formés. Afin d'améliorer l'interprétation des résultats, seule des analyses de sols avec géoréférencement pourraient fournir les explications très précises quant à la différence entre les quantiles estimés entre les groupes.

Il reste cependant encore des aspects à étudier au niveau de l'étude de la détection des valeurs aberrantes proprement dite. En effet, le regroupement des entités géographiques, pourrait être effectué en prenant en compte plusieurs variables simultanément, c'est-à-dire en considérant le cas multivarié, afin de faciliter ultérieurement les traitements informatiques de détection de valeurs aberrantes. Ceci pourrait malheureusement entraîner

une perte d'homogénéité des groupes formés mais améliorerait la capacité de détection multivariée par la prise en compte des relations éventuelles entre les variables.

Afin de rendre le système tout à faire performant, il faudrait également à l'avenir distinguer les informations liées à la fiche descriptive de l'échantillon (signalétique), c'est-à-dire les types d'occupation de sols, les régions agricoles ; ce qui signifie la prise en compte de la contrainte liée à la structuration des données citée lors de la présentation de la problématique de l'étude.

De plus, étant donné la constitution de la carte des principaux types de sols de la Région wallonne et la stratification de l'espace rural en *Unités de stratification de l'espace rural* (USER)[27], il serait intéressant d'étudier la possibilité de prendre en compte ces derniers au lieu des régions agricoles utilisées classiquement en Région wallonne ; le nombre relativement élevé de USERS pouvant limiter leur utilisation.

Enfin, il subsiste le problème de l'opérationnalité des méthodes à appliquer. En ce qui concerne purement le cas de la détection des valeurs aberrantes, bien que le problème des données influentes soit pris en considération par des logiciels statistiques, tels que SAS et MINITAB, la complexité du traitement des valeurs aberrantes est probablement la raison pour laquelle aucun logiciel statistique ou procédure, identifiée en tant que telle, n'est disponible. Malgré le développement très important des logiciels statistiques qui permettent de traiter de grandes bases de données, en ce qui concerne la détection de valeurs aberrantes, des procédures sont incorporées dans les logiciels mais jusqu'à un certain point, il faut donc tenir compte de l'aptitude de ceux-ci à traiter le problème des valeurs anormales. Un dilemme est cependant bien évident si on met en parallèle le but de l'analyse statistique par ordinateur qui est plutôt de réaliser des routines répétitives d'analyses avec la déclaration d'une valeur aberrante qui est soumise à une déclaration subjective. Un réel manque d'outils statistiques, permettant de tester l'intégrité des données, est observé.

Dans le même ordre d'idées, dans le cas de la classification avec contraintes spatiales, la variabilité du rapport d'efficacité observée entre les groupes d'entités géographiques nous suggère d'approfondir ce domaine d'étude malgré le faible nombre de logiciels permettant d'intégrer la contrainte spatiale.

[27] FUSAGx – ENTENAFOR et principalement le laboratoire de Géopédologie en partenariat avec la DGA.

Ceci nous mène aux perspectives générales de ce travail. Le développement de procédures ou de logiciels statistiques spécifiques aux domaines d'études cités ci-dessus, c'est-à-dire la détection des valeurs aberrantes et la classification spatiale, est très attendu.

Quant aux applications de la méthode proposée, elles touchent les domaines de la gestion et le suivi de la qualité des bases de données ainsi que la spatialisation dans le cadre de Systèmes d'Information Géographique (SIG). En ce qui concerne les bases de données relatives à des échantillons d'analyse de sols, problème qui a été à l'origine de l'étude, le géoréférencement des données est actuellement effectué de manière régulière sur le terrain. Cependant, la validation des bases de données est toujours nécessaire car des valeurs aberrantes peuvent encore apparaître lorsque l'échantillonnage n'est pas représentatif (par exemple, échantillon prélevé au niveau d'un pissat de vache) ou lorsque des erreurs lors de la transcription des coordonnées géographiques se produisent. De plus, lors de campagnes d'échantillonnage très intenses, les échantillonneurs négligent parfois la prise d'informations géographiques sur le terrain ou le matériel peut présenter des défaillances techniques. Des erreurs en laboratoires peuvent également survenir.

D'autre part, d'anciennes bases de données[28] peuvent également être très intéressantes pour des caractéristiques bien précises alors qu'aucune information géographique précise n'est disponible. La validation de ces bases de données est essentielle et est réalisable à partir de la méthode qui a été élaborée.

Quant au domaine de la contamination des sols dans les sites pollués, la méthode proposée pourrait être appliquée pour la détermination du seuil permettant de considérer que le sol est contaminé ; ce seuil correspondrait à un quantile extrême à établir.

Enfin, dans le domaine de l'agriculture de précision, une problématique similaire à celle qui vient d'être étudiée est également rencontrée suite à la présence de contraintes spatiales et de mélanges de distributions. En effet, au sein d'une même parcelle, des zones différentes (par exemple, zone plus sèche et zone plus humide) peuvent être observées alors qu'aucun référentiel tel que la carte de sols n'est disponible. Il est donc nécessaire d'appliquer la classification avec contrainte spatiale afin de définir des zones homogènes au sein de la parcelle. De plus, au sein d'une même parcelle, il est envisageable de considérer, non pas le caractère pédologique, mais plutôt les rendements qui sont observés en continu pour établir des zonations ; des rendements différents, au sein d'une même parcelle, peuvent alors être observés. Ces zones différentes constituent alors une contrainte

[28] Par exemple, la base de données Aardewerk au laboratoire de Géopédologie - FUSAGx.

spatiale dont il est nécessaire de tenir compte lors de la validation des observations. Enfin, lorsque des croisements de différentes couches d'informations sont réalisés à partir par exemple des cartes de rendement, de l'imagerie satellitaire, des cartes de sols, les valeurs des unités élémentaires (pixels) peuvent être étudiées de manière à détecter des valeurs aberrantes rencontrées suite à l'enregistrement en continu (passage sur un caillou, problèmes techniques) et aux problèmes qui peuvent en découler.

Le travail réalisé nous a permis de constater combien le problème de la détection des valeurs aberrantes, bien qu'ayant fait l'objet d'une littérature abondante, reste un sujet qui est loin d'être complètement étudié. Il existe de nombreuses situations qui nécessitent encore des recherches et l'évolution des techniques et des outils de gestion des données constituent autant d'opportunités pour développer des applications spécifiques.

BIBLIOGRAPHIE

Les publications se terminant par **(*)** sont en relation avec la thèse.

Aitchison, J. et Brown, J.A.C. (1969). *The lognormal distribution with special references to its uses in economics.* Cambridge, Cambridge Univ. Press, 176 pp.

Anonyme (1951). Arrêté Royal fixant la délimitation des régions agricoles du Royaume (24 février 1951). Moniteur Belge, 15 mars 1951.

Anonyme (1975). Carte des régions agricoles de Belgique (1/300.000). Ministère de l'Agriculture (ed.). Institut Géographique National (IGN).

Anonyme (1995). Exactitude (justesse et fidélité) des résultats et méthodes de mesure. Partie 2 : Méthode de base pour la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée (ISO 5725-2:1994). In: *Méthodes statistiques pour la maîtrise de la qualité. Méthodes et résultats de mesure. Interprétation des données statistiques. Maîtrise des processus,* Genève, Suisse, Organisation internationale de normalisation, vol. 2, 31-78.

Anonyme (2003). *Requasud, un réseau d'analyse et de conseil.* Les Cahiers de l'Agriculture. Jambes, Ministère de la Région Wallonne, Direction générale de l'Agriculture. 4 pp.

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics*, **2**, 123-147.

Baamal, L. (1994). *Etude des règles d'arrêt en classification numérique.* Gembloux, Faculté des Sciences Agronomiques, Thèse de doctorat, 257 pp.

Bah, B. et Veron, P. (2005). *Mise en oeuvre de la phase"interprétation" du Projet de Cartographie Numérique des Sols de Wallonie (P.C.N.S.W.).* Gembloux, Faculté Universitaire des Sciences Agronomiques, Unité Sol - Ecologie - Territoire, Laboratoire de Géopédologie, Unité de Gestion des Ressources forestières et des Milieux naturels. Rapport d'état d'avancement des activités de la convention du 9 mai 2005, 38 pp + annexes.

Bah B., Engels P., Colinet G. (2005). Ed. Sc. : Bock L. *Légende de la Carte Numérique des Sols de Wallonie (Belgique).* Laboratoire de Géopédologie, Faculté universitaire des Sciences agronomiques de Gembloux, Belgique. 53 p. + annexes

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proc. Roy. Soc. London,* **Ser. A, 353** (1674), 401-419.

Barndorff-Nielsen, O. et Christiansen, C. (1988). Erosion, deposition and size distributions of sand. *Proc. Roy. Soc. London,* **Ser. A, 417** (1853), 335-352.

Barnett, V. (1978). Multivariate outliers : Wilk's test and distance measures. *Bull. Int. Stat. Inst.*, **47** (4), 37-40.

Barnett, V. (1983). Principles and methods for handling outliers in data sets. In: *Statistical methods and the improvement of data quality*, Wright, T. (ed.). Orlando, Florida, Academic Press, 131-166.

Barnett, V. et Lewis, T. (1994). *Outliers in statistical data*. New York, John Wiley, 3rd edition.

Barnett, V. et Turkman, K.F. (1993). *Statistics for the environment*. Chichester, England, John Wiley & Sons, 427 pp.

Barnett, V. et Turkman, K.F. (1994). *Statistics for the environment. Water related issues.* Chichester, England, John Wiley & Sons, vol. 2, 391 pp.

Beckman, R.J. et Cook, R.D. (1983). Outlier.....s' (with Discussion). *Technometrics,* **25**, 119-163.

Beghin, H. (1979). *Méthodes d'analyse géographique quantitative.* Série Droit. Paris, France, Librairie technique, 252 pp.

Beirlant, J. et Goegebeur, Y. (2000). *Local polynomial maximum likelihood estimation for Pareto-type distributions*. K.U.Leuven, Department of Applied Economics. Technical Report n°24

Beirlant, J., Goegebeur, Y., Verlaak, R. et Vynckier, P. (1998). Burr regression and portfolio segmentation. *Insurance : Mathematics and Economics,* **23**, 231-250.

Beirlant, J., Teugels, J.L. et Vynckier, P. (1996). *Practical analysis of* extreme values. Leuven, Leuven University Press, 137 pp.

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004). *Statistics of extremes. Theory and applications.* Wiley Series in Probability and Statistics. Chichester, England, John Wiley & Sons Ltd., 490 pp.

Bezdek, J.C. (1981). *Pattern recognition with Fuzzy objective function algorithms.* New York, Plenum Press.

Birks, H.J.B. et Gordon, A.D. (1985). *Numerical methods in quaternary pollen analysis.* London, Academic Press.

Borgers, N. (2005). *Contribution à l'évaluation de la qualité des sols aux abords des rives sud du lac Naivasha, Kenya*. Gembloux, Belgique, Faculté Universitaire des Sciences Agronomiques. Travail de fin d'études, 80 pp.

Box, G.E.P. et Cox, D.R. (1964). An analysis of transformations. *J. R. Stat. Soc.*, **26, Ser. B**, 211-252.

Bracke, C. et Veron, P. (2002). *Présentation de la méthodologie du Projet de Cartographie Numérique des Sols de Wallonie (PCNSW).*

Journée d'étude "La carte numérique des sols de Wallonie et ses applications", Wépion, Belgique, le 9 décembre 2002, 16 pp.

Brazauskas, V. et Serfling, R. (2001). Small sample performance of robust estimators of tail parameters for Pareto and exponential models. *J. Stat. Comput. Simul.*, **70**, 1-19.

Brostaux, Y. (2002). *Introduction à l'environnement de programmation statistique R.* Notes de Statistique et d'Informatique. Gembloux, Faculté des Sciences Agronomiques. 22 pp.

Burr, I.W. (1942). Cumulative frequency distributions. *Ann. Math. Stat.*, **13**, 215-232.

Caers, J., Vynckier, P., Beirlant, J. et Rombouts, L. (1996). Extreme value analysis of diamond-size distributions. *Math. Geol.*, **28** (1), 25-43.

Campbell, N.A. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Appl. Stat.*, **27**, 251-258.

Carletti, G. (1988). *Comparaison empirique de méthodes statistiques de détection de valeurs anormales à une et à plusieurs dimensions.* Gembloux, Belgique, Fac. Univ. Sci. Agron., Thèse de doctorat, 225 pp.

Casgrain (2004). *Le Progiciel R : Analyse multidimensionnelle, analyse spatiale.* [Online]. Disponible à l'adresse : http://www.fas.umontreal.ca/BIOL/casgrain/en/labo/R/index.html. Consulté le: 19/08/2004.

Cerioli, A. et Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers. *J. Comput. Graphical Stat.*, **8** (2), 239-258.

Chandon, J.L. et Pinson, S. (1981). *Analyse typologique - Théories et applications.* Paris, France, Masson, 254 pp.

Cheng, H.D., Jiang, X.H., Sun, Y. et Wang, J. (2001). Color image segmentation : advances and propects. *Pattern Recognition*, **34** (9), 2259-2281.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17** (8), 790-799.

Chikkagoudar, M.S. et Kunchur, S.H. (1987). Comparison of many outlier procedures for exponential samples. *Comm. Stat. Theor. Meth.*, **16**, 627-645.

Colinet, G. (2003). *Eléments traces métalliques dans les sols. Contribution à la connaissance des déterminants de leur distribution spatiale en région limoneuse belge.* Gembloux, Belgique, Fac. Univ. Sci. Agron., Thèse de doctorat + annexes, 414 pp + 18 pp + 6 pp.

Colinet, G. (2004). Communication personnelle.

Colinet, G., Toussaint, B., Laroche, J., Goffaux, M.J. et Oger, R. (2005). *Base de données sols de Réquasud - 2ème synthèse.* Gembloux, FUSAGx, Unité de Géopédologie, Réquasud. Note interne, 32 pp.

Comaniciu, D. et Meer, P. (2002). Mean shift : a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.,* **24** (5), 603-619.

Cook, R.D. et Weisberg, S. (1980). Characterisations of an empirical influence function for detecting influential cases in regression. *Technometrics,* **22**, 495-508.

Cook, R.D. et Weisberg, S. (1982). *Residuals and influence in regression*. London, Chapman and Hall.

Crettaz de Roten, F. et Helbling, J.-M. (1996). Données manquantes et aberrantes : le quotidien du statisticien analyste de données. *Rev. Stat. Appl.* **XLIV** (2), 105-115.

Csörgö, S., Deheuvels, P. et Mason, D.M. (1985). Kernel estimates of the tail index of a distribution. *Ann. Stat.,* **13** (3), 1050-1077.

Dagnelie, P. (1975). *Analyse statistique à plusieurs variables.* Gembloux, Les Presses Agronomiques de Gembloux A.S.B.L., 362 pp.

Dagnelie, P. (1998a). *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique.* Paris et Bruxelles, De Boeck et Larcier, 508 pp.

Dagnelie, P. (1998b). *Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions.* Paris et Bruxelles, De Boeck et Larcier, 659 pp.

Dagnelie, P. (2003). Communication personnelle.

Davalo, E. et Naïm, P. (1993). *Des réseaux de neurones.* Paris, Eyrolles, 232 pp.

David, H.A., Hartley, H.O. et Pearson, E.S. (1954). The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika,* **41**, 482-493.

Davies, L. et Gather, U. (1993). The identification of multiple outliers (with discussion). *J. Am. Stat. Assoc.,* **88**, 782-801.

Dekkers, A.L.M., Einmahl, J.H.J. et de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Stat.,* **17** (4), 1833-1855.

Derrig, A. et Ostaszewsli, K. (1994). *Fuzzy techniques of pattern recognition in risk and claim classification.* Transactions of the 4th AFIR International Colloquium, Orlando, 141-171.

Dixon, W.J. (1950). Analysis of extreme values. *Ann. Math. Stat.,* **21**, 488-506.

Dodd, E.L. (1923). The greatest and least variate under general laws of error. *Trans. Am. Math. Soc.,* **25**, 525-539.

Essenwanger, O.M. (1986). *Elements of statistical analysis*. World Survey of Climatology, General Climatology. Amsterdam, Elsevier, vol. 1B, 424 pp.

Everitt, B.S. (2002). *The Cambridge dictionnary of statistics*. Cambridge, UK, University Press, Second Edition, 410 pp.

Everitt, B.S., Landau, S. et Leese, M. (2001). *Cluster analysis.* London, Arnold Publishers, Fourth edition, 237 pp.

FAO (1994). Evaluations agro-écologiques aux fins de planification nationale : l'exemple du Kenya. *Bull. Pédol. FAO* 67, 171 pp.

Ferguson, T.S. (1961). Rules for rejection of outliers. *Rev. Inst. Int. Stat.* **29** (3), 29-43.

Fischer, M.J., Gross, D., Bevilacqua Masi, D.M., Shortle, J. et Brill, P.H. (2001). *Using quantile estimates in simulating internet queues with Pareto service times*. Peters, B.A., Smith, J.S., Medeiros, D.J. and Rohrer, M.W. (eds.). The 2001 Winter Simulation Conference, 477-485.

Fisher, R.A. et Tippett, L.H.C. (1928). Limiting form of the frequency distribution of the largest or the smallest member of a sample. *Proc. Cambr. Philos. Soc.*, **24**, 180-190.

Foguenne, M. (1994). *Classification des communes wallonnes selon leur ruralité.* Gembloux, Faculté des Sciences Agronomiques, Unité de Statistique et d'Informatique. Travail de fin d'études, 93 pp + annexes.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Pol. Math. Cracovie*, **6**, 93-116.

Fung, K.Y. et Paul, S.R. (1985). Comparisons of outlier detection procedures in Weibull or extreme-value distribution. *Comm. Stat. Simulation Comput.*, **14**, 895-917.

Garrido, M. (2002). *Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution*. Grenoble, Université Grenoble 1, Mathématiques appliquées, 231 pp.

Glaisher, J.W.L. (1872). On the law of facility of errors of observations and on the method of least squares. *Mem. Roy. Astr. Soc.*, **39**, 75-124.

Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatorie. *Ann. Math.*, **44** (3), 423-453.

Goegebeur, Y. (2003). Communication personnelle.

Goegebeur, Y., Planchon, V., Beirlant, J. et Oger, R. (2002). *Quality assessment of pedochemical data using extreme value methodology.* Leuven, Belgium, K.U. Leuven. Technical Report 2002-08, 14 pp. **(*)**

Goegebeur, Y., Planchon, V., Beirlant, J. et Oger, R. (2005). Quality assessment of pedochemical data using extreme value methodology. *J. Appl. Sci.*, 5 (6), 1092-1102.

(*)

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.*, **21**, 27-58.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1-21.

Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques. *Ann. Inst. Henri Poincaré*, **4**, 115-158.

Gumbel, E.J. (1958). *Statistics of extremes.* New York, U.S.A., Columbia University Press.

Hachama, M. et Bohoua-Nasse, F.O. (2003). *Une segmentation grossière et rapide des images en couleurs.* [Online]. Disponible à l'adresse: http://eleves.dptmaths.ens-cachan.fr/~hachama/index.htm. Consulté le: 24/09/2004.

Hampel, F., Ronchetti, E.M., Rousseeuw, P. et Stahel, W.A. (1986). *Robust Statistics*. New York, John Wiley.

Hawkins, D.M. (1980). *Identification of outliers.* Monographs on Applied Probability and Statistics. London, England, Chapman and Hall, 188 pp.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, **3** (5), 1163-1174.

Holt, T. (2002). [Online]. Disponible à l'adresse: http://www.cru.uea.ac.uk/cru/projects/mice/MICE_methods.pdf. Consulté le: 09/01/2003.

Hongchang, H. (2000). *Multi-spectral satellite image classification using artificial neural networks.* Fribourg, Université de Fribourg, Thèse de doctorat, 144 pp.

Höppner, F., Klawonn, F., Kruse, R. et Runkler, T. (2000). *Fuzzy cluster analysis. Methods for classification, data analysis and image recognition.* Chichester, John Wiley & Sons, LTD, 289 pp.

Houghton, J.C. (1988). Use of the truncated shifted Pareto distribution in assessing size distribution of oil and gas fields. *Math. Geol.*, **20** (8), 907-937.

Huber, P.J. (1972). Robust statistics : a review. *Ann. Math. Stat.*, **43**, 1041-1067.

Huber, P.J. (1981). *Robust statistics.* New York, U.S.A., Wiley.

Immarco, P. (1992). NeuroWindows - A new look for neural nets. *PC AI*, **7** (1), 2-4.

Jeevanand, E.S. et Nair, N.U. (1993). Prediction of future observations from Pareto population in the presence of an outlier. *Statistica*, **LIII** (2), 171-176.

Jeevanand, E.S. et Nair, N.U. (1998). On determining the number of outliers in exponential and Pareto samples. *StHefte*, **39**, 277-290.

Johnson, N.L. et Kotz, S. (1970). *Continuous univariate distributions-1*. Distributions in statistics. Boston, Houghton Mifflin Company, 300 pp.

Kabe, D.G. (1970). Testing outliers from an exponential population. *Metrika*, **15**, 15-18.

Kimber, A.C. (1988). Testing upper and lower outlier pairs in gamma samples. *Comm. Stat. Simulation Comput.*, **17**, 1055-1072.

Kinnison, R.R. (1985). *Applied extreme value statistics*. Columbus, Battelle Press, 149.

Kotz, S. et Johnson, N.L. (1982). *Encyclopedia of statistical sciences*. New York, Wiley-Interscience, vol. 1, 480 pp.

Kruskal, W.H. (1960). Discussion of the papers of Messrs. Anscombe and Daniel. *Technometrics*, **2**, 157-158.

Lalor, G.C. et Zhang, C. (2001). Multivariate outlier detection and remediation in geochemical databases. *Sci. Total Environ.*, **281**, 99-109.

Lambert, P. et Grecu, H. (2003). *A quick and coarse color image segmentation*. ICIP 2003, 14-17 September 2003,

Lange, B. (1982). *Contribution à l'étude de la localisation des activités agricoles en Belgique*. Gembloux, Faculté des Sciences Agronomiques de l'Etat, Thèse de doctorat, 316 pp.

Laroche, J. et Oger, R. (1999). *Base de données sols*. Gembloux, Belgique, Faculté Universitaire des Sciences Agronomiques de Gembloux, Unité de Géopédologie, asbl Requasud, première synthèse, 36 pp.

Lawson, A.B. et Denison, D.G.T. (2002). *Spatial cluster modelling*. Boca Raton, Chapman & Hall, 287 pp.

Legendre, L. et Legendre, P. (1984a). *Ecologie numérique. La structure des données écologiques*. Collection d'Ecologie 12 et 13. Paris, Masson et les Presses de l'Université du Québec, tome 2, viii + 335 pp.

Legendre, L. et Legendre, P. (1984b). *Ecologie numérique. Le traitement multiple des données écologiques*. Collection d'Ecologie 12 et 13. Paris, Masson et les Presses de l'Université du Québec, tome 1, xv + 260 pp.

Legendre, P. (1987). Constrained clustering. In: *Developments in numerical ecology.*, Legendre, P. and Legendre, L. (eds.). Berlin, Springer-Verlag, vol. G 14, 289-307.

Legendre, P. et Legendre, L. (1998). *Numerical ecology*. Developments in Environmental Modelling. Amsterdam, The Netherlands, Elsevier, vol. 20, 853 pp.

Lewis, T. et Fieller, N.R.J. (1979). A recursive algorithm for null distributions for outliers : I. Gamma samples. *Technometrics*, **21**, 371-376.

Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika*, **11**, 46-54.

Lu, J. et Sedransk, N. (2002). *Tail metrics for network performance based on GPD and mixture modeling.* Baltimore. 29 pp.

Madsen, C. (2001). *EVT and VaR.* [Online]. Disponible à l'adresse: http://www.garp.com/library/Meets/EVT_and_VaR.ddb.ppt. Consulté le: 09/01/2003.

Mandelbrot, B. (1963). New methods in statistical economics. *Bull. Inst. Int. Stat.*, **40** (2), 699-721.

Maréchal, R. et Tavernier, R. (1974). *Commentaire des planches 11A et 11B. Extraits de la carte des sols. Carte des associations de sols. Pédologie.* Comité National de Géographie. Atlas de Belgique. 64 pp.

Minasny, B. et McBratney, A.B. (2002). *FuzME version 3.0.* [Online]. Disponible à l'adresse: http://www.usyd.edu.au/su/agric/acpa. Consulté le: 19/10/2001.

Mokadem, A.I. (2002). *Projet de Cartographie Numérique des Sols de Wallonie : Initiation du projet et intégration des résultats dans le SIG de la Région wallonne.* Journée d'étude "La carte numérique des sols de Wallonie et ses applications", Wépion, Belgique, le 9 décembre 2002, 11 pp.

Munoz-Garcia, J., Moreno-Rebollo, J.L. et Pascual-Acosta, A. (1990). Outliers : A formal approach. *Int. Statist. Rev.* **58**, 215-226.

Murphy, R.B. (1951). *On tests for outlying observations.* Princeton University, University Microfilms Inc., Ann Arbor, Mich., Ph. D. Thesis

Murtagh, F.D. (1995). Contiguity-constrained hierarchical clustering in partitioning data sets. In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science,* Cox, I., Hansen, P. and Julesz, B. (eds.). American Mathematical Society Provindence, RI, vol. 19, 143-152.

Otten, A. et Van Montfort, M.A.J. (1980). Maximum likelihood estimation of the general extreme value distribution parameters. *J. Hydrol.*, **47**, 187-192.

Palm, R. (1992). *Comment interpréter les résultats d'une série chronologique.* Collection STAT-ITCF. Boigneville. 80 pp.

Palm, R. (1996). *La classification numérique : principes et application.* Notes de Statistique et d'Informatique. Gembloux, Faculté des Sciences Agronomiques. 28 pp.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat.*, **3** (1), 119-131.

Ping, J.L. et Dobermann, A. (2003). Site-specific management. Creating spatially contiguous yield classes for site-specific management. *Agron. J.*, **95**, 1121-1131.

Planchon, V. (2005). Traitement des valeurs aberrantes : concepts actuels et tendances générales. *Biotechnol. Agron. Soc. Environ.*, **9** (1), 19-34. **(*)**

Prévot, H. (2004). *Comparaison de méthodes statistiques et neuronales pour l'établissement d'équations de calibrage en spectrométrie de réflexion diffuse dans le proche infrarouge.* Gembloux, Faculté Universitaire des Sciences Agronomiques, Thèse de doctorat, 382 pp.

Ripley, B.D. (1993). Statistical aspects of neural networks. In: *Networks and chaos. Statistical and probabilistic aspects*, Barndorff-Nielsen, O.E., Jensen, J.L. and Kendall, W.S. (eds.). London, Chapman and Hall, chapter 2, 40-123.

Ripley, B.D. (1994). Neural networks and related methods for classification. *J. R. Stat. Soc.*, **56, Ser. B** (3), 409-456.

Rocke, D.M. (1992). Estimation of variation after outlier rejection. *Comput. Stat. Data Anal.*, **13** (1), 9-20.

Rothenbuehler, J. (2002). *Extreme tail's tales. Adventures in high quantile estimation.* [Online]. Disponible à l'adresse: http://www.orie.cornell.edu/~jrothenb/seminar.ppt. Consulté le: 09/01/2003.

Rousseeuw, P.J. et Bassett, G.W. (1990). The Remedian : a robust averaging method for large data sets. *J. Am. Stat. Assoc.*, **85**, 97-104.

Rousseeuw, P.J. et Leroy, A.M. (1987). *Robust regression and outlier detection*. New York, John Wiley and Sons, 329 pp.

Royston, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat.*, **31** (2), 115-124.

Sarle, W.S. (2001). *Neural networks FAQ.* [Online]. Disponible à l'adresse: http://ftp.sas.com/pub/neural/FAQ.html. Consulté le: 27/09/2004.

Shapiro, S.S. et Wilk, M.B. (1972). An analysis of variance test for the exponential distribution (complete samples). *Technometrics*, **14**, 355-370.

Shapiro, S.S., Wilk, M.B. et Chen, M.J. (1968). A comparative study of various tests for normality. *J. Am. Stat. Assoc.*, **63**, 1343-1372.

Sichel, H.S. (1973). Statistical evaluation of diamondiferous deposits. *J. South Afr. Inst. Min. Met.*, **73**, 235-243.

Sichel, H.S., Dohm, C.E. et Kleingeld, W.J. (1995). New generalized model of observed ore value distributions. *Trans. Instn. Min. Metall. (sect. A : Min. industry)*, **104** (2), 116-123.

Smara, Y., Ouarab, N., Laama, S. et Cherifi, D. (2003). *Techniques de fusion et de classification floue d'images satellitaires multisources pour la caractérisation et le suivi de l'extension du tissu urbain de la région d'Alger (Algérie).* 2nd FIG Regional Conference, Marrakech, Morocco, 2-5 December 2003, 15 pp.

Smith, M. (1996). *Neural networks for statistical modeling.* London, International Thomson Computer Press, 235 pp.

Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Stat.*, **15** (3), 1174-1207.

Smith, R.L. (1989). Extreme value analysis of environmental time series : an application to trend detection in ground-level ozone. *Stat. Sci.*, **4** (4), 367-393.

Sneyers, R. (1960). On a special distribution of maximum values. *Mon. Weather Rev.* **88**, 66-69.

Thode, H.C. (2002). *Testing for normality.* STATISTICS : textbooks and monographs. New York, U.S.A., Marcel Dekker, Inc., vol. 164, 479 pp.

Tietjen, G.L. et Moore, R.H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics,* **14**, 583-597.

Vandewalle, B. (2004). *Some robust and semi-parametric methods in extreme value theory.* Leuven, KUL, Thèse de doctorat, 146 pp.

Van Meirvenne, M., De Smet, J., Hofman, G., Vanderdeelen, J. et Baert, L. (1993). Inventory, evaluation and fuzzy classification of the ammonium-lactate extractable phosphate in central West-Flanders, Belgium. *Bull. Rech. Agron. Gembloux,* **28** (2-3), 393-407.

Van Montfort, M.A.J. (1970). On testing that the distribution of extremes is of type I when type II is the alternative. *J. Hydrol.*, **11**, 421-427.

Van Montfort, M.A.J. et Otten, A. (1978). On testing a shape parameter in the presence of a location and a scale parameter. *Math. Operationsforsch. Statist.*, **9, Ser. Statistics** (1), 91-104.

Zeng, Y. et Starzyk, J. (2001). *Statistical approach to clustering in pattern recognition.* [Online]. Disponible à l'adresse: http://www.ent.ohiou.edu/-starzyk/network/Research/Papers/Clustering.pdf. Consulté le: 24/06/2004.

Zhang, C.S., Selinus, O. et Schedin, J. (1998). Statistical analyses for heavy metal contents in till and root samples in an area of southeastern Sweden. *Sci. Total Environ.*, **212**, 217-232.

Zhang, C.S. et Zhang, S. (1996). A robust-symmetric mean : a new way of mean calculation for environmental data. *Geojournal,* **40** (1-2), 209-212.

Zivot, E. (2002). *Financial econometrics.* [Online]. Disponible à l'adresse: http://faculty.washington/edu/ezivot/econ512/econ512extremevalue.pdf. Consulté le: 09/01/2003.

ANNEXES

Annexe 1. Régions agricoles de la Région wallonne.

Légende	
5	Région sablo-limoneuse
6	Région Limoneuse
7	Campine Hennuyère
8	Condroz
9	Région herbagère
10	Région herbagère (Fagne)
11	Famenne
12	Ardenne
13	Région Jurassique
14	Haute Ardenne

Figure 1. Représentations des dix régions agricoles de la Région wallonne (Anonyme, 1951, 1975).

Annexe 2. Nombre d'observations par commune avant fusion et par niveau de troncature à gauche ou à droite (T30=30% des données retenues, T20=20%, T10=10%).

	nis	Code postal	Commune avant fusion	Effectif	T30	T20	T10
1	52011	6001	MARCINELLE	7	2	1	1
2	52011	6010	COUILLET	8	2	2	1
3	52011	6030	MARCHIENNE-AU-PONT	15	5	3	2
4	52011	6040	JUMET	4	1	1	0
5	52011	6044	ROUX	19	6	4	2
6	52011	6061	MONTIGNIES-SUR-SAMBRE	74	22	15	7
7	52012	6200	CHATELET	8	2	2	1
8	52018	6240	FARCIENNES	0	0	0	0
9	52025	6280	GERPINNES	72	22	14	7
10	52048	6111	LANDELIES	31	9	6	3
11	52074	6250	AISEAU-PRESLES	172	52	34	17
12	56005	6500	BEAUMONT	410	123	82	41
13	56005	6511	STREE	64	19	13	6
14	56044	6540	LOBBES	57	17	11	6
15	56044	6542	SARS-LA-BUISSIERE	15	5	3	2
16	56044	6543	BIENNE-LEZ-HAPPART	29	9	6	3
17	56086	6120	HAM-SUR-HEURE-NALINNE	179	54	36	18
18	61003	4540	AMAY	251	75	50	25
19	61012	4560	CLAVIER	1650	495	330	165
20	61031	4500	HUY	857	257	171	86
21	61039	4570	MARCHIN	369	111	74	37
22	61041	4577	MODAVE	654	196	131	65
23	61043	4550	NANDRIN	699	210	140	70
24	61048	4590	OUFFET	619	186	124	62
25	61068	4530	VILLERS-LE-BOUILLET	1773	532	355	177
26	61072	4520	WANZE	2230	669	446	223
27	61079	4160	ANTHISNES	253	76	51	25
28	61079	4161	VILLERS-AUX-TOURS	21	6	4	2
29	61079	4162	HODY	40	12	8	4
30	61079	4163	TAVIER	108	32	22	11
31	61080	4480	ENGIS	519	156	104	52
32	61081	4557	TINLOT	628	188	126	63
33	62093	4460	SAINT-NICOLAS	0	0	0	0
34	62118	4460	GRACE-HOLLOGNE	427	128	85	43
35	62120	4400	FLEMALLE	146	44	29	15
36	91005	5537	ANHEE	523	157	105	52
37	91030	5590	CINEY	1423	427	285	142
38	91034	5500	DINANT	862	259	172	86
39	91034	5501	LISOGNE	56	17	11	6
40	91034	5502	THYNES	102	31	20	10

	nis	Code postal	Commune avant fusion	Effectif	T30	T20	T10
41	91034	5503	SORINNES	23	7	5	2
42	91034	5504	FOY-NOTRE-DAME	36	11	7	4
43	91059	5360	HAMOIS	407	122	81	41
44	91059	5361	SCY	104	31	21	10
45	91059	5362	ACHET	63	19	13	6
46	91059	5363	EMPTINNE	142	43	28	14
47	91059	5364	SCHALTIN	58	17	12	6
48	91064	5370	HAVELANGE	1302	391	260	130
49	91064	5372	MEAN	38	11	8	4
50	91064	5374	MAFFE	42	13	8	4
51	91064	5376	MIECRET	32	10	6	3
52	91103	5520	ONHAYE	115	35	23	12
53	91103	5521	SERVILLE	31	9	6	3
54	91103	5522	FALAEN	60	18	12	6
55	91103	5523	SORINNES	147	44	29	15
56	91103	5524	GERIN	92	28	18	9
57	91141	5530	YVOIR	435	131	87	44
58	92003	5300	ANDENNE	646	194	129	65
59	92006	5330	ASSESSE	763	229	153	76
60	92006	5332	CRUPET	83	25	17	8
61	92006	5333	SORINNE-LA-LONGUE	124	37	25	12
62	92006	5334	FLOREE	118	35	24	12
63	92006	5336	COURRIERE	46	14	9	5
64	92045	5150	FLOREFFE	839	252	168	84
65	92048	5070	FOSSES-LA-VILLE	545	164	109	55
66	92054	5340	GESVES	497	149	99	50
67	92087	5640	METTET	1555	467	311	156
68	92087	5641	FURNAUX	12	4	2	1
69	92087	5644	ERMETON-SUR-BIERT	105	32	21	11
70	92087	5646	STAVE	183	55	37	18
71	92094	5000	NAMUR	76	23	15	8
72	92094	5001	BELGRADE	1	0	0	0
73	92094	5003	SAINT-MARC	172	52	34	17
74	92094	5004	BOUGE	39	12	8	4
75	92094	5020	MALONNE	212	64	42	21
76	92094	5021	BONINNE	22	7	4	2
77	92094	5022	COGNELEE	22	7	4	2
78	92094	5024	NANINNE	45	14	9	5
79	92094	5100	WIERDE	106	32	21	11
80	92094	5101	ERPENT	99	30	20	10

	nis	Code postal	Commune avant fusion	Effectif	T30	T20	T10
81	92097	5350	OHEY	718	215	144	72
82	92097	5351	HAILLOT	121	36	24	12
83	92097	5352	PERWEZ	23	7	5	2
84	92097	5353	GOESNES	68	20	14	7
85	92097	5354	JALLET	41	12	8	4
86	92101	5170	PROFONDEVILLE	248	74	50	25
87	92137	5060	SAMBREVILLE	183	55	37	18
88	92140	5190	JEMEPPE-SUR-SAMBRE	275	83	55	28
89	93022	5620	FLORENNES	724	217	145	72
90	93022	5621	THY-LE-BAUDUIN	448	134	90	45
91	93088	5650	WALCOURT	671	201	134	67
92	93088	5651	THY-LE-CHATEAU	478	143	96	48

Annexe 3. Histogrammes de fréquences relatives et graphiques des quantiles normaux pour les variables étudiées.

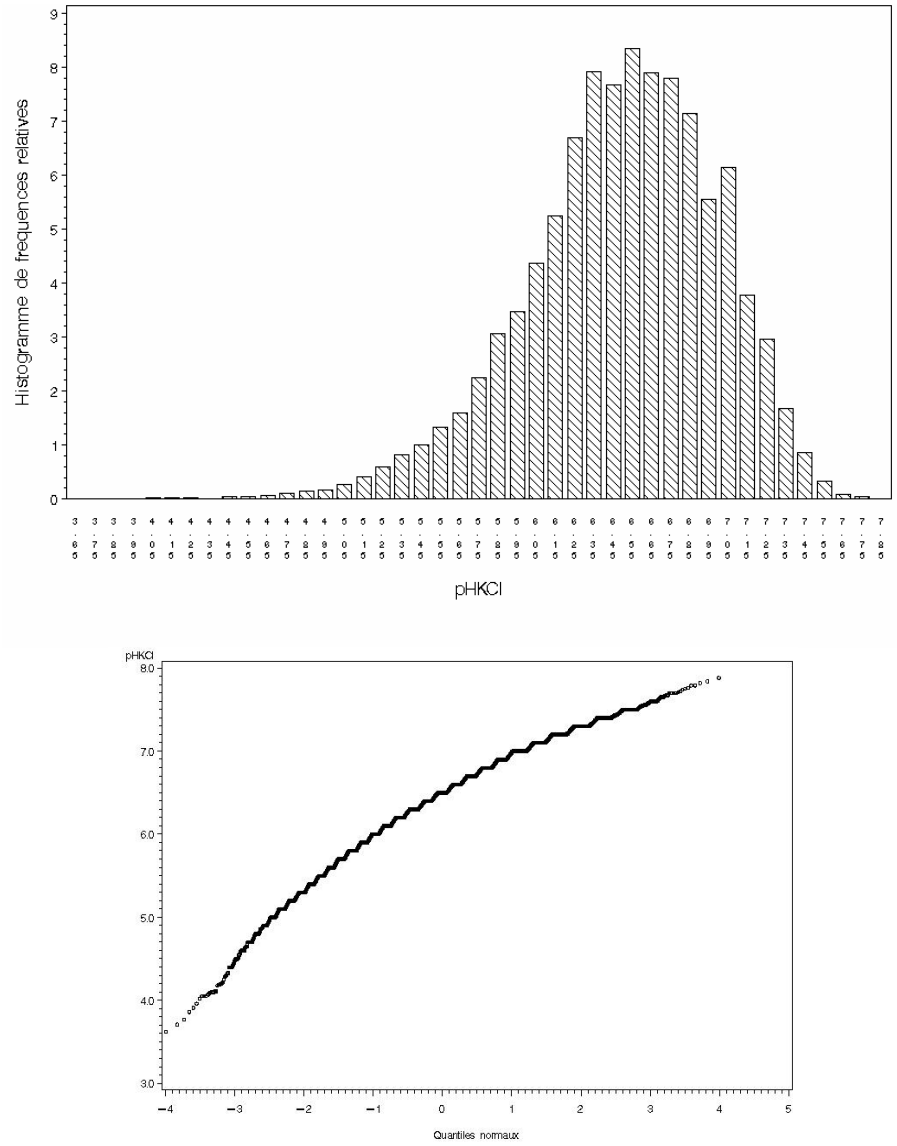


Figure 1. Histogramme de fréquences relatives et graphique des quantiles normaux pour la variable pHKCl et pour les observations de la région agricole du Condroz (terres de culture).

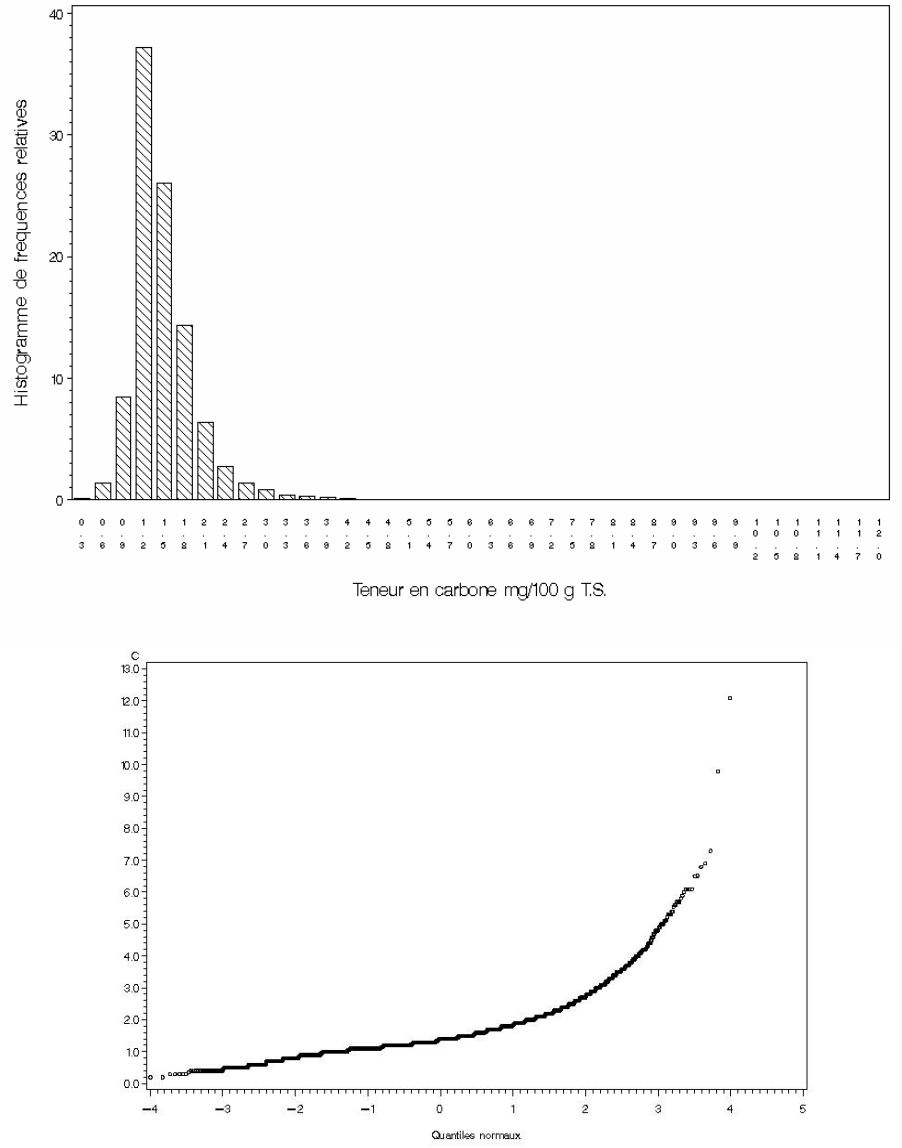


Figure 2. Histogramme de fréquences relatives et graphique des quantiles normaux pour la variable C et pour les observations de la région agricole du Condroz (terres de culture).

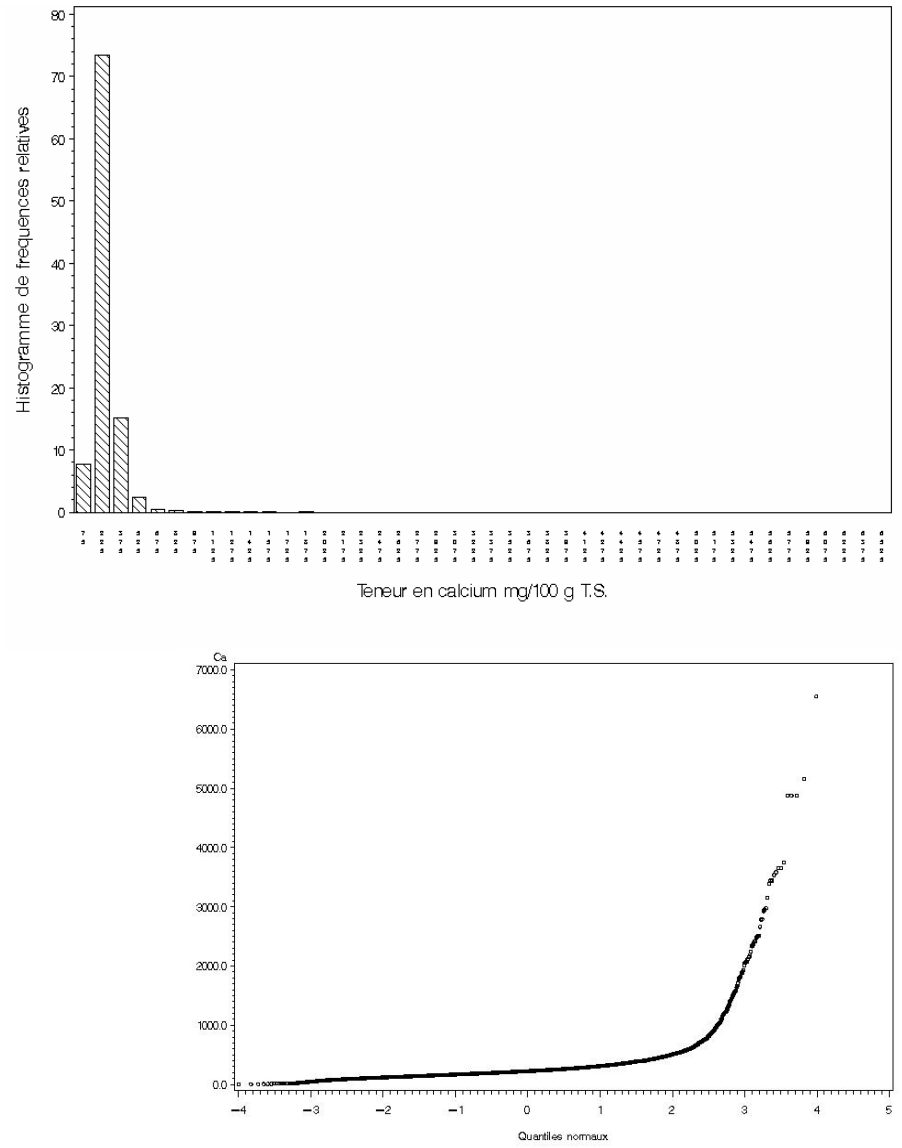


Figure 3 Histogramme de fréquences relatives et graphique des quantiles normaux pour la variable Ca et pour les observations de la région agricole du Condroz (terres de culture).

Annexe 4. Teneurs en carbone organique total et en calcium disponible pour les terres de cultures des différentes régions agricoles wallonnes.

Tableau 1. Teneur en carbone organique total (en mg C/100g TS) pour les terres de cultures des différentes régions agricoles wallonnes ; paramètres descriptifs de la distribution de la population ; période 1998-2002(Colinet *et al.*, 2005).

C					
Régions agricoles	**Effectif**	**P5**	**P95**	**m**	**s**
Région sablo-limoneuse	6.172	0,9	1,9	1,3	0,4
Région limoneuse	32.210	0,9	1,9	1,3	0,4
Campine hennuyère	378	0,8	1,8	1,1	0,4
Condroz	15.034	1,0	**2,3**	1,5	0,4
Région herbagère (Liège)	2.338	1,2	3,9	2,2	0,9
Région herbagère (Fagne)	264	1,2	2,8	1,8	0,6
Famenne	1.729	1,2	**3,3**	2,0	0,7
Ardenne	1.750	1,7	**4,3**	3,1	0,8
Haute Ardenne	102	2,1	5,5	1,8	1,0
Région jurassique	1.496	1,0	3,3	1,5	0,7

Tableau 2. Teneur en calcium disponible (en mg Ca/100g TS) des terres de cultures pour les différentes régions agricoles wallonnes ; paramètres descriptifs de la distribution de la population ; période 1998-2002 (Colinet *et al.*, 2005).

Ca					
Régions agricoles	**Effectif**	**P5**	**P95**	**m**	**s**
Région sablo-limoneuse	6.172	119,0	364,0	220,0	77,5
Région limoneuse	32.210	142,8	407,3	253,3	91,2
Campine hennuyère	378	65,9	252,5	141,8	59,0
Condroz	**15.034**	**144,0**	400,0	244,3	86,3
Région herbagère (Liège)	2.338	125,6	397,4	238,7	88,5
Région herbagère (Fagne)	264	143,0	333,2	218,1	65,3
Famenne	1.729	**124,8**	409,8	226,5	100,1
Ardenne	1.750	**80,9**	278,0	163,5	66,3
Haute Ardenne	102	79,6	233,5	138,5	43,4
Région jurassique	1.496	57,7	432,6	213,8	144,8

Annexe 5. Nombre d'observations par entité communale et par niveau de troncature à gauche ou à droite (T30=30%, T20=20%, T10=10%) après regroupement *a priori*.

Code	NIS	Commune (nis - fusion)	Nombre total d'observations	Troncature		
				10%	20%	30%
1	52011	CHARLEROI/CHATELET/MONTIGNIES-LE-TILLEUL	166	17	33	49
2	52074	AISEAU-PRESLES/FARCIENNES	172	17	34	52
3	56005	BEAUMONT	474	47	95	142
4	56044	LOBBES	101	10	20	30
5	56086	HAM-SUR-HEURE-NALINNE	179	18	36	54
6	61003	AMAY	251	25	50	75
7	61012	CLAVIER	1650	165	330	495
8	61031	HUY	857	86	171	257
9	61039	MARCHIN	369	37	74	111
10	61041	MODAVE	654	65	131	196
11	61043	NANDRIN	699	70	140	210
12	61048	OUFFET	619	62	124	186
13	61068	VILLERS-LE-BOUILLET	1773	177	355	532
14	61072	WANZE	2230	223	446	669
15	61079	ANTHISNES	422	42	84	127
16	61080	ENGIS	519	52	104	156
17	61081	TINLOT	628	63	126	188
18	62118	GRACE-HOLLOGNE/SAINT-NICOLAS	427	43	85	128
19	62120	FLEMALLE	146	15	29	44
20	91005	ANHEE	523	52	105	157
21	91030	CINEY	1423	142	285	427
22	91034	DINANT	1079	108	216	324
23	91059	HAMOIS	774	77	155	232
24	91064	HAVELANGE	1414	141	283	424
25	91103	ONHAYE	445	45	89	134
26	91141	YVOIR	435	44	87	131
27	92003	ANDENNE	646	65	129	194
28	92006	ASSESSE	1134	113	227	340
29	92045	FLOREFFE	839	84	168	252
30	92048	FOSSES-LA-VILLE	545	55	109	164
31	92054	GESVES	497	50	99	149
32	92087	METTET/GERPINNES	1927	193	385	579
33	92094	NAMUR	794	79	159	238
34	92097	OHEY	971	97	194	291
35	92101	PROFONDEVILLE	248	25	50	74
36	92137	SAMBREVILLE	183	18	37	55
37	92140	JEMEPPE-SUR-SAMBRE	275	28	55	83
38	93022	FLORENNES	1172	117	234	352
39	93088	WALCOURT	1149	115	230	345

Annexe 6. Elément carbone : évaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles.

Tableau 1. Elément carbone : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution exponentielle et le niveau de troncature de 10%, pour l'ensemble du Condroz avec la distribution exponentielle et le niveau de troncature de 10%, pour les entités communales séparées et à partir des limites de RéQuaSud.

Nombre de groupes	Contaminants de communes de l'Ardenne (n=2552)	Contaminants de communes de la Famenne non voisines au Condroz (n=435)	Contaminants de communes de la Famenne voisines au Condroz (n=2522)
2	5,49%	0,31%	0,51%
3	9,26%	2,15%	0,91%
4	16,31%	4,02%	1,66%
5	8,17%	1,93%	0,83%
6	24,84%	5,06%	2,42%
7	14,70%	3,71%	1,57%
8	18,52%	4,94%	2,11%
9	10,15%	2,56%	1,04%
39 entités communales	18,73%	5,20%	2,17%
Ensemble du Condroz	6,43%	2,07%	0,59%
Limites RéQuaSud	13,8%	2,5%	1,3%

Annexe 7. Elément carbone : représentation graphique des groupes obtenus et quantiles estimés par groupe.

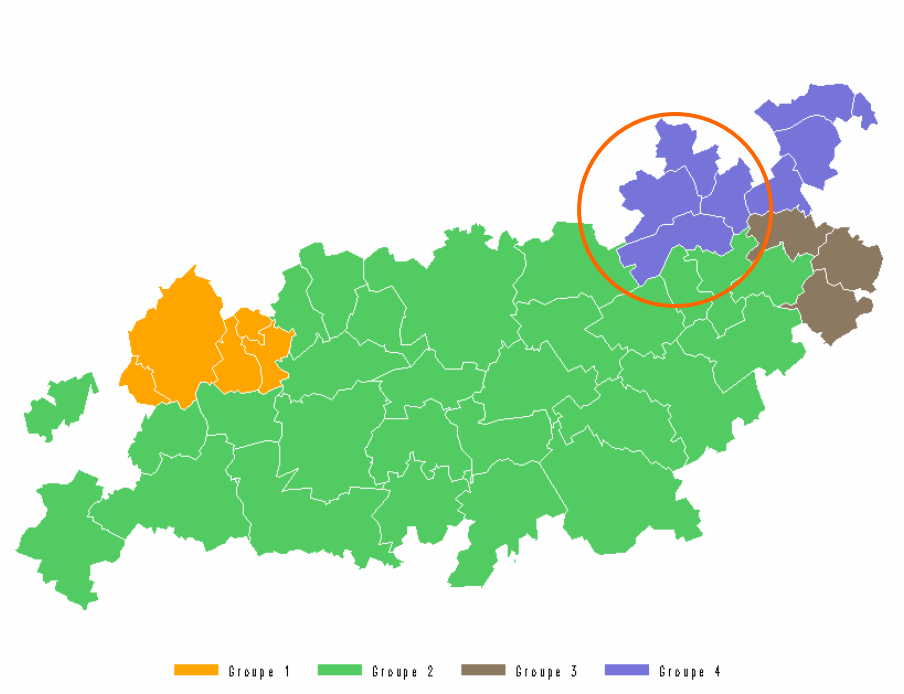


Figure 1. Elément carbone – Formation de 4 groupes *a posteriori*.

Tableau 1. Quantiles estimés 0,999 pour les 4 groupes formés.

Groupes	n	Quantile estimé 0,999
1	338	9,103
2	20528	4,547
3	1740	3,296
4	6203	3,803

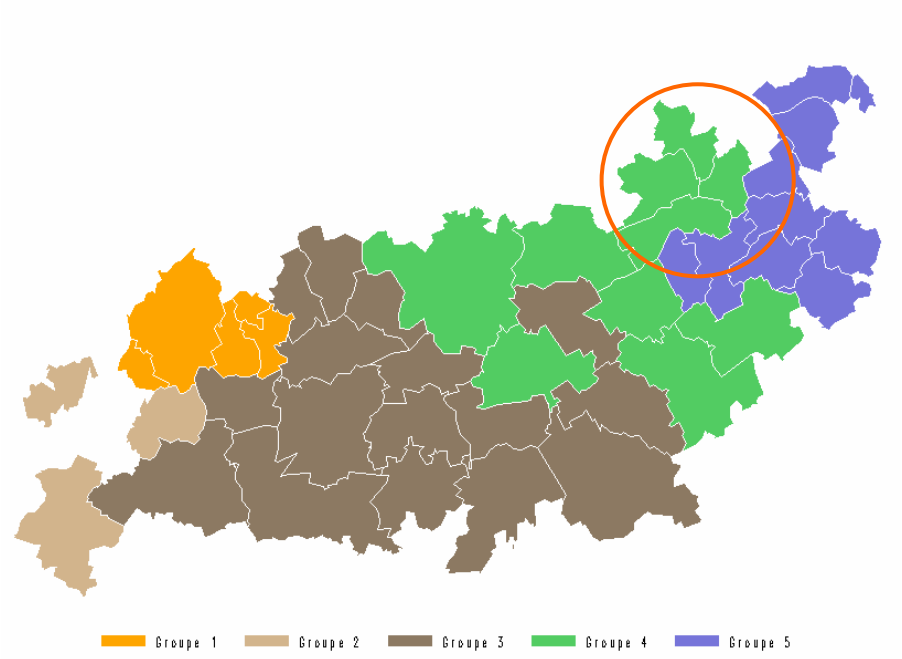


Figure 2. Elément carbone – Formation de 5 groupes *a posteriori*.

Tableau 2. Quantiles estimés 999 par groupes.

Groupes	n	Quantile estimé 0,999
1	338	9,103
2	754	5,456
3	11514	4,398
4	11720	3,951
5	4483	3,675

Annexe 8. Elément calcium : évaluation du taux de détection des valeurs aberrantes issues d'autres régions agricoles

Tableau 1. Elément calcium : pourcentage de détection de valeurs aberrantes issues d'autres régions agricoles pour la distribution de Weibull et le niveau de troncature de 10%, pour l'ensemble du Condroz avec la distribution de Weibull et le niveau de troncature de 10%, pour les entités communales séparées et à partir des limites de RéQuaSud.

Nombre de groupes	Contaminants de communes de l'Ardenne (n=2552)	Contaminants de communes de la Famenne non voisines au Condroz (n=435)	Contaminants de communes de la Famenne voisines au Condroz (n=2522)
2	2,12%	0,04%	0,12%
3	2,64%	0,23%	0,15%
4	3,33%	0,29%	0,24%
5	4,08%	0,37%	0,25%
6	6,97%	0,69%	0,59%
7	3,79%	0,33%	0,23%
8	9,17%	1,06%	0,95%
9	7,83%	0,89%	0,58%
39 entités communales	3,4%	0,2%	0,2%
Ensemble du Condroz	11,63%	2,77%	1,40%
Limites RéQuaSud	2,9%	0,7%	0,5%