B A
S E

Biotechnol. Agron. Soc. Environ. 2005 **9** (1), 19–34

Traitement des valeurs aberrantes : concepts actuels et tendances générales

Viviane Planchon

Section Biométrie, Gestion des Données et Agrométéorologie. Centre wallon de Recherches agronomiques. Rue de Liroux, 9. B-5030 Gembloux (Belgique). E-mail : planchon@cra.wallonie.be

Reçu le 27 avril 2004, accepté le 7 janvier 2005.

En raison de l'évolution rapide des moyens de collecte automatique des données et de leur traitement informatique, le problème des valeurs aberrantes a pris une importance non négligeable durant les dernières décennies. Par exemple, le développement des systèmes d'informations géographiques, de l'agriculture de précision et de la capture automatique de données a entraîné la constitution de grands ensembles de données où les valeurs anormales peuvent plus facilement passer inaperçues. La présence de valeurs anormales peut alors conduire à des estimations biaisées des paramètres des populations et, suite à la réalisation de tests statistiques, à une interprétation des résultats qui peut être erronée. Cet article met en évidence la diversité des méthodes disponibles pour l'utilisateur et met l'accent sur la manière de traiter les valeurs aberrantes de façon structurée. Malgré des fondements théoriques très largement développés et une bibliographie très abondante sur le sujet, on constate que la plupart des logiciels statistiques existant sur le marché sont très limités quant au traitement des valeurs aberrantes.
Mots-clés. Valeur aberrante, test de discordance, accommodation, détection.

Treatment of outliers: present state of concepts and general tendencies. The fast evolution of data collection means and the computerised treatment of information induced a huge problem of outliers during last decades. For example, development of geographic information systems, precision farming and automatisation of data capture led to the constitution of large databases where outliers can more easily go unnoticed. The presence of abnormal values may lead to biased estimations of parameters of populations and, further to the realization of statistical tests to interpretation of results, which can be erroneous. This article highlights the diversity of available methods and emphasizes the way of handling outliers in a structured way. In spite of very widely developed theoretical concepts and a large bibliography on the subject, we notice that most of existing statistical software packages on the market are very limited for the treatment of outliers.
Keywords. Outliers, discordancy test, accommodation, detection.

1. INTRODUCTION

Les observations *non représentatives* ou *aberrantes* ont toujours été considérées comme une source de contamination, déformant l'information obtenue à partir des données brutes. 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 (Barnett, Lewis, 1994).

Depuis plus d'un siècle, un large éventail de méthodes d'analyses statistiques de plus en plus précises ont été construites 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. Le problème des valeurs aberrantes est incontournable pour toutes les personnes qui manipulent des données et doivent juger de la manière de traiter celles-ci. Divers domaines d'études se sont très fortement développés au cours de cette dernière décennie, avec comme conséquence l'acquisition de très grands ensembles de données. Tel est le cas par exemple pour les systèmes d'informations géographiques, l'agriculture de précision et la capture automatique de données et enfin, la constitution de bases de données en général. Cette automatisation de l'acquisition des données crée une situation où la personne n'est plus en contact avec celles-ci et n'est pas capable d'appréhender l'adéquation de certaines d'entre elles.

De plus, il ne faut pas négliger le développement très important des logiciels statistiques qui permettent de traiter de grandes bases de données. Les facilités, que ceux-ci fournissent, deviennent de plus en plus sophistiquées avec des options d'affichage des données (*data-screening*), des procédures de validation des modèles avec des interventions de type semi-intelligent. Par contre, en ce qui concerne la détection, les tests et l'accommodation des valeurs aberrantes, des procédures sont incorporées dans les logiciels mais jusqu'à un certain point. Toute utilisation de méthodes de détection de valeurs aberrantes par ordinateur doit donc tenir compte des limites des méthodes fournies par les logiciels. 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 (Barnett, Lewis, 1994).

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*. Les valeurs aberrantes n'induisent pas forcément en erreur, elles ne sont pas forcément mauvaises ou erronées. Dans certains cas, l'expérimentateur peut même 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*[1] *aux valeurs aberrantes ou les accommodant*.

Durant la dernière décennie, on a réellement pris conscience qu'avant tout traitement d'observations anormales, il faut prendre en compte diverses notions directement liées aux valeurs aberrantes (Barnett, Lewis, 1994). La manière d'appréhender ces valeurs est dès lors plus structurée. En effet, des distinctions bien claires entre les objectifs des analyses statistiques et la manière de considérer les données doivent être réalisées. Barnett et Lewis dressent une classification des types de questions auxquelles il faut réfléchir lors de l'étude de valeurs aberrantes. D'après ces auteurs, il est nécessaire de faire la distinction entre les causes déterministes ou aléatoires d'apparition de valeurs aberrantes, entre les différents objectifs à atteindre lors de l'étude des valeurs aberrantes, entre les différents modèles de probabilité spécifiques, entre les données univariées et multivariées et enfin entre les valeurs aberrantes simples ou multiples.

Dans cet article, nous allons examiner les différentes manières d'aborder le problème des valeurs aberrantes en prenant en considération ces différentes notions.

2. DÉFINITIONS

Avant d'exposer des concepts relatifs aux valeurs aberrantes, il est nécessaire de les définir de manière plus précise. De nombreux auteurs ont cherché à décrire 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) 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. 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

[1] En anglais: *to accommodate the outlier*

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 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*. 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*.

À 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 est défini par Everitt (2002) comme les valeurs les plus grandes et les plus petites parmi un ensemble d'observations. Barnett et Lewis (1994) ont distingué, dans le cas univarié, 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 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. 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 1(a)**, ni la valeur $x_{(1)}$, ni $x_{(n)}$ ne semblent correspondre à une valeur aberrante. Par contre, dans la **figure 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. 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 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, au niveau de la **figure 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. Le terme *valeur suspecte*[2] 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 dans la suite de cet article en liaison avec les *tests de discordance*.

Enfin, il nécessaire de parler des *observations influentes*[3] qui sont définies par Everitt (2002) comme étant des observations qui ont une influence disproportionnée sur un ou plusieurs aspects de l'estimateur d'un paramètre, en particulier, les coefficients de régression. Cette influence peut être due à des différences par rapport aux autres observations de la variable explicative ou à une valeur extrême de la variable à expliquer. L'auteur signale que les valeurs aberrantes sont souvent des observations influentes. Selon Cook et Weisberg (1980), les observations influentes sont celles pour

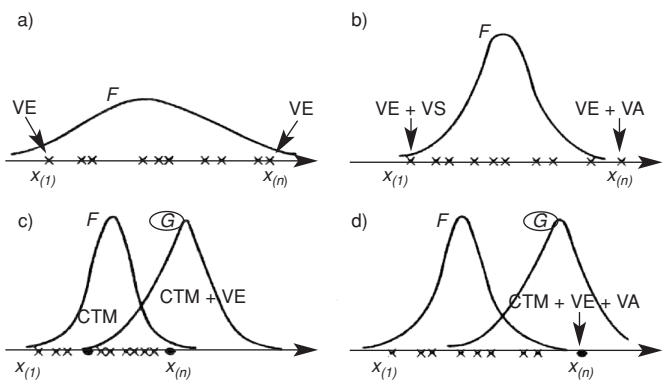

Figure 1. Définition des termes : valeurs extrêmes, valeurs aberrantes et contaminants (figure adaptée à partir de Barnett et Lewis, 1994) — *Definition of terms: extreme values, outliers and contaminants (figure adapted from Barnett and Lewis, 1994)*.

[2] en anglais : *straggler*

[3] en anglais : *influential observation*

lesquelles les caractéristiques de l'analyse sont altérées de manière considérable quand elles sont supprimées. Cette considération est développée au paragraphe présentant les modèles de régression.

Contrairement au cas univarié, le concept d'extrême est difficile à établir pour les données multivariées. La valeur aberrante est moins apparente intuitivement, elle est effectivement cachée dans la masse de données et se trouve en périphérie du nuage de points (Afifi, Azen, 1979). Rohlf (1975) signale qu'il est possible de caractériser les valeurs aberrantes par le fait qu'elles sont quelque peu isolées du principal nuage de points. Elles ne se trouvent pas à la fin de la distribution comme pour les valeurs aberrantes univariées mais elles "dépassent quelque part". Les points qui ne se trouvent pas à l'intérieur du nuage de points sont des valeurs aberrantes potentielles.

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.

D'après Gnanadesikan et Kettering (1972), les conséquences de la présence de valeurs aberrantes dans un échantillon multivarié sont bien plus complexes que dans le cas univarié. En effet, la valeur aberrante multivariée peut déformer non seulement les mesures de position et d'échelle mais également les relations entre les variables.

À 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 restait encore le problème du traitement des échantillons comprenant une ou deux valeurs aberrantes. Des techniques d'analyse en composantes principales ont été utilisées pour résoudre ce problème.

3. ORIGINE DES VALEURS ABERRANTES ET OBJECTIFS POURSUIVIS

Au fil du temps, les explications relatives à l'apparition de valeurs aberrantes dans un ensemble de données ont évolué et ont fait apparaître des causes bien distinctes, liées à la nature de celles-ci. Une classification des différentes manières par lesquelles les valeurs aberrantes peuvent survenir a été discutée dans la littérature par divers auteurs, tels que Beckman et Cook (1983) et Hawkins (1980). Lors de la collecte de données, différentes sources de variabilité peuvent être rencontrées dont la variabilité inhérente, l'erreur de mesure et l'erreur d'exécution (Barnett, Lewis, 1994).

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.

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. Quelques contrôles de ce type de variabilité sont possibles et facilement réalisables.

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 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.

Une illustration d'erreur de mesure d'exécution possible est obtenue à partir des boxplots de la **figure 2**. Cette figure a été réalisée à partir de données issues de la base de données de la Chaîne "SOLS" de

Réquasud[4] 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 (exprimée en grammes par 100 grammes de terre sèche ou g/100 g T.S.) (Laroche, Oger, 1999). On

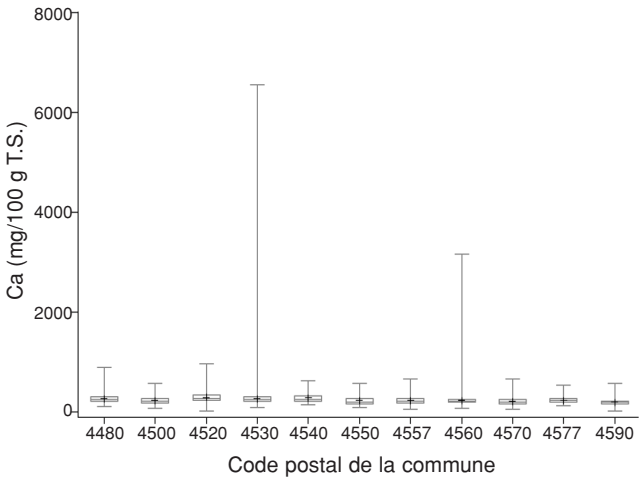

Figure 2. Boxplots des teneurs en calcium (mg/100 g T.S.) pour 11 communes de la Région condruzienne. Identification de deux valeurs particulièrement élevées — *Boxplots for the calcium content (mg/100 g T.S.) of 11 districts of the Condroz region. Identification of two very high values.*

[4] Depuis 1994, une base de données a été constituée à partir de résultats d'analyses de terre réalisées par les laboratoires de la chaîne sols du réseau Réquasud. Ce réseau a été créé afin de mettre à la disposition des praticiens (agriculteurs, vulgarisateurs, etc.) des moyens d'analyses et de conseils efficaces dans le secteur agricole et agroalimentaire. Le réseau a notamment pour but l'amélioration et la promotion de la qualité des produits et des analyses. Il s'occupe en outre du secteur « SOL » et des analyses d'échantillons de l'ensemble de la Région wallonne ont été réalisées. Les résultats sont centralisés au sein de la base de données ; celle-ci contient, à l'heure actuelle, plus de 300.000 enregistrements avec les résultats des analyses de la composition chimique (pHKCl, K, Mg, Ca, etc.) et des variables relatives au lieu d'extraction de l'échantillon, à la texture des sols, à l'occupation des sols, aux types de cultures actuelles et précédentes. L'Unité de Géopédologie (Faculté universitaire des Sciences agronomiques de Gembloux) est le laboratoire de référence pour les analyses de sol et la base de données est centralisée à la Section Biométrie, Gestion des données et Agrométéorologie (Centre wallon de Recherches agronomiques). Une étude approfondie de l'ensemble de ces données permet de connaître les propriétés physiques et chimiques des terres agricoles par types de sols et donc de gérer de manière optimale les sols grâce à la connaissance de leur potentiel de fertilité et de leur capacité à produire l'une ou l'autre culture. Les agriculteurs possèdent ainsi les informations nécessaires à un choix judicieux de l'affectation de leurs parcelles.

observe une valeur aberrante pour la commune ayant le code postal 4530 et une autre 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 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. La valeur correcte aurait peut-être été 655,3 mg/100 g ou 6553,0 mg/1000 g. Le même raisonnement peut être suivi pour l'observation extrême 3225,0 mg/1000 g rencontrée pour la deuxième commune.

Dans une étude relative à des données de type géochimiques géoréférencées, Lalor et Zhang (2001) classent les valeurs aberrantes en trois catégories : valeurs aberrantes d'amplitude, spatiales et relationnelles.

Les valeurs aberrantes d'amplitude sont considérées comme étant 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 sont généralement définies comme des observations qui sont extrêmes par rapport aux valeurs voisines (Cerioli, Riani, 1999). Dans le cas de la base de données de Réquasud, pour des communes avoisinantes, des caractéristiques similaires sont attendues et correspondent à des zones pédologiques semblables.

Les valeurs aberrantes relationnelles sont définies comme des observations non conformes aux relations qui existent entre les éléments.

Les diverses sources de variation, qui provoquent l'apparition de valeurs aberrantes de natures différentes, ont montré la complexité de l'examen des valeurs aberrantes. Cependant, le fait d'être capable de préciser ces notions de nature et d'origine des valeurs aberrantes permet actuellement de déterminer de manière plus structurée les objectifs à atteindre lors de l'examen d'observations anormales. Les objectifs de l'étude des valeurs aberrantes dépendent en effet de l'origine et de la nature de celles-ci, comme le montre la **figure 3**. Cette figure 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, de rejeter les

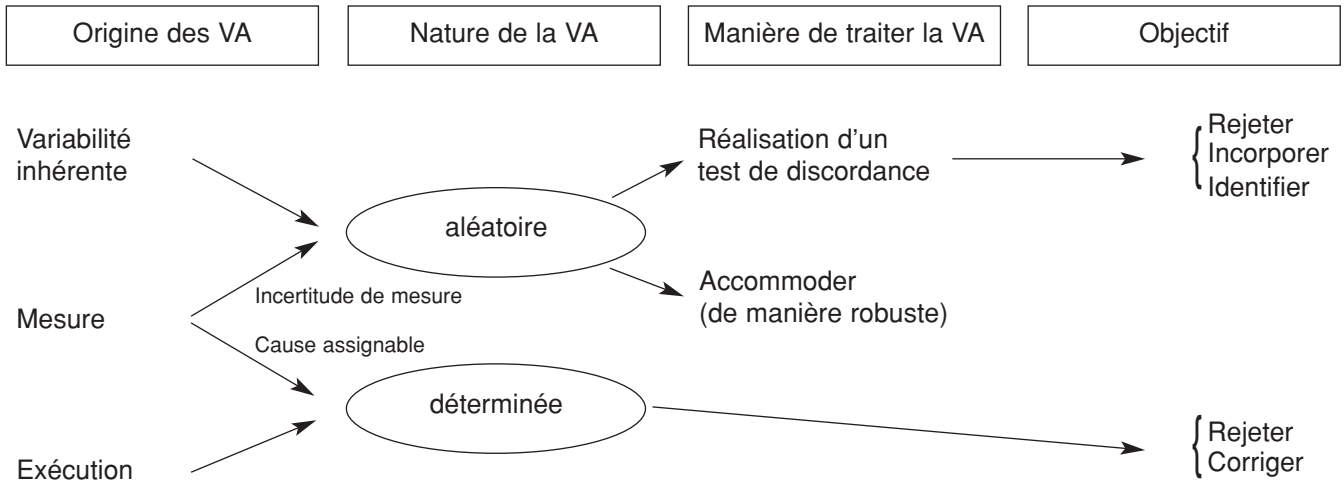

Figure 3. Schéma général de traitement des valeurs aberrantes et objectifs poursuivis lors de l'examen des valeurs aberrantes (Barnett, Lewis, 1994) — *General figure of treatment of outliers and objectives followed during the study of outliers (Barnett, Lewis, 1994).*

valeurs discordantes et de procéder à l'analyse à partir de l'échantillon modifié. D'autres possibilités peuvent cependant également être intéressantes. En effet, on peut choisir d'utiliser un autre modèle que celui choisi initialement. Les données biologiques demandent souvent des modèles de distributions dissymétriques. Ceux-ci 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 effets 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.

L'analyse 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 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 on remplace les extrêmes les plus faibles et les plus grands par leurs plus proches voisins ou la réalisation de censure/rognage vont également avoir des implications sur les distributions. Le processus de rognage consiste à utiliser un échantillon dans lequel une fraction fixée des valeurs extrêmes, basses et élevées, de l'échantillon initial est totalement mise de côté.

Quant aux valeurs aberrantes dont la nature est déterminée, c'est-à-dire les erreurs de mesure ou d'exécution, elles peuvent être rejetées ou faire l'objet de corrections dans la mesure où celles-ci sont encore réalisables.

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.

4. 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.

L'exemple présenté dans la suite de cet article 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 4**). Lorsqu'une relation linéaire est obtenue à partir de ce graphique, les données suivent une distribution normale (Dagnelie, 1998). 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 situées entre 700 et 1000 mg/100 g T.S car celles-ci sont également 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.

Comme pour la distribution normale, des graphiques des quantiles peuvent être réalisés afin de vérifier l'adéquation des observations à ces distributions.

La **figure 5** présente le graphique des quantiles pour la distribution de Pareto. Ces graphiques des quantiles sont construits en plaçant en abscisse, les quantiles exponentiels *Q(p)* et en ordonnée, le logarithme des valeurs observées triées selon un ordre croissant (Beirlant *et al.*, 1996; Beirlant, Goegebeur, 2000). Les quantiles exponentiels sont définis, pour une série de *n* observations, par la relation

$$Q(p) = -\log(1-p)$$

où $p = \frac{i}{n+1}$ et $0<p<1$.

En considérant la partie droite de ce graphique, une relation linéaire est observée et permettrait d'inclure l'ensemble des données excepté la valeur la plus élevée. Celle-ci serait considérée comme aberrante selon la distribution de Pareto.

Par contre pour la partie gauche de cette distribution, la distribution de Pareto n'est pas adaptée. Il faudrait faire appel à d'autres distributions. 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 (ou gauche) des distributions dissymétriques.

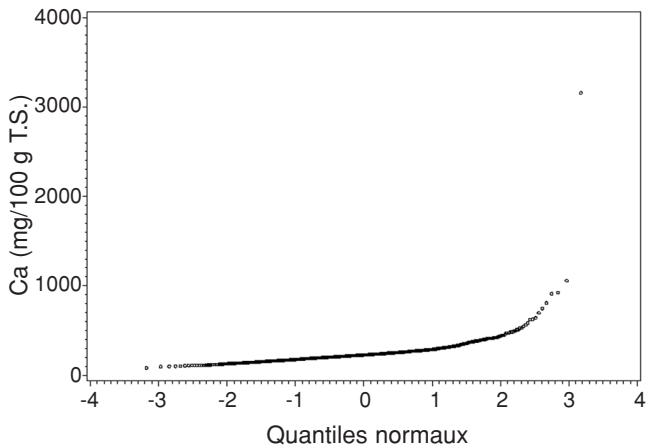

Figure 4. Distribution normale : graphique des quantiles normaux pour les valeurs de calcium (mg/100 g T.S.) d'une commune de la zone agricole du Condroz (n=1342) — *Normal distribution: normal quantiles for the calcium content (mg/100 g T.S.) of a district of the agricultural area of Condroz.*

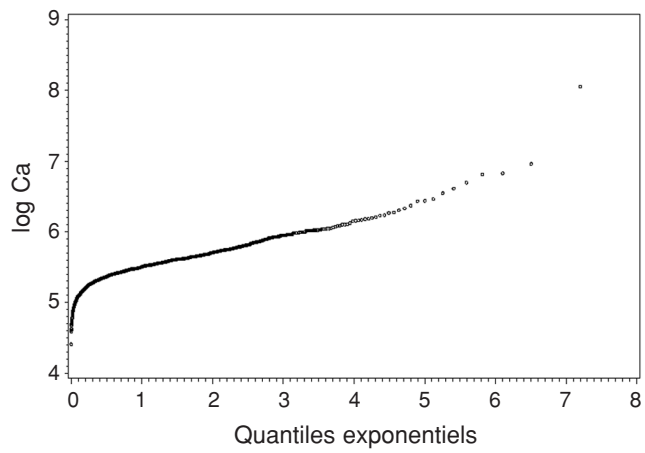

Figure 5. Distribution de Pareto : graphique des quantiles exponentiels pour le logarithme des valeurs de calcium (mg/100 g T.S.) d'une commune de la zone agricole du Condroz (n=1342) — *Pareto distribution: exponential quantiles for the calcium content (mg/100 g T.S.) transformed by the logarithm, of a district of the agricultural area of Condroz.*

5. TEST DE DISCORDANCE

5.1. Introduction

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. Dans cette optique, Barnett et Lewis (1994) proposent d'utiliser un *test de discordance*. L'objectif poursuivi lors de l'utilisation de tests de discordance 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.

5.2. Traitement des valeurs aberrantes dans le cas univarié

Comme précédemment, soit l'échantillon $x_1, x_2, \ldots, x_n$ 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 de discordance peut être réalisé pour examiner si $x_{(n)}$ doit être considéré 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 5725 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 (ISO, 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 les suivants :

- les statistiques liées *au rapport excès/étalement* (Dixon, 1950) ;
- les statistiques liées *au rapport amplitude/étalement* ;
- les statistiques liées *au rapport écart/étalement*, comme le test classique de Grubbs (Grubbs, 1950 ; Tietjen, Moore, 1972) ;
- les statistiques liées *au rapport extrêmes/position* ;
- les statistiques liées *au rapport de sommes de carrés* ;
- les statistiques liées *aux moments d'ordre supérieurs* ;
- la statistique *W de Shapiro-Wilks* (Shapiro *et al.*, 1968 ; Royston, 1982).

Carletti (1988) et Barnett et Lewis (1994) présentent une liste de tests de détection de valeurs aberrantes pour un échantillon normal en fonction du nombre de valeurs aberrantes à détecter.

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, exponentielle tronquée, gamma, uniforme, Poisson et binomiale. Les principaux auteurs de ces tests sont : Kimber (1988), Lewis, Fieller (1979), Chikkagoudar, Kunchur (1987).

Il faut signaler que ces tests sont sujets à ce qu'on appelle l'effet de masque[5], qui consiste en l'incapacité d'une procédure statistique d'identifier une valeur aberrante en présence de plusieurs valeurs suspectes. Afin d'éviter cet effet de masque et les contraintes

[5] En anglais : *masking effect*

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), Carletti (1988) a sélectionné le test classique de Grubbs et s'est intéressé à deux variantes de celui-ci, en utilisant des notions de robustesse.

La première variante correspond au test de Grubbs appliqué à un échantillon tronqué symétriquement afin d'obtenir une estimation plus robuste de la moyenne et de l'écart-type. Un bon compromis semble être une troncature de 8 % symétrique (4 % à gauche, 4 % à droite). La troncature n'étant pas toujours suffisante pour se rapprocher de l'hypothèse de normalité exigée par la méthode de Grubbs, l'auteur propose, en cas de non-normalité, un test de Grubbs sur des échantillons tronqués et transformés par une fonction puissance (Box, Cox, 1964). Carletti (1988) conclut que le test de Grubbs de détection de valeurs anormales réalisé sur un échantillon ordonné et tronqué symétriquement est plus intéressant en terme de puissance que le test classique de Grubbs pour les données faiblement aberrantes et très nettement aberrantes. Les résultats pour l'échantillon transformé sont nettement moins puissants que le test sur l'échantillon tronqué non transformé pour les données faiblement aberrantes et très nettement aberrantes. Suivant les procédures de détection de valeurs aberrantes proposées par Carletti (1988), ces transformations sont 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.

5.3. Traitement de valeurs aberrantes dans le cas de modèles de régression et de l'analyse de la variance

L'objectif de l'examen d'échantillons univariés pour l'ajustement de modèles et pour l'estimation de paramètres, quoique étant une partie importante de la pratique statistique, est relativement limité. Le plus souvent des situations plus structurées sont à considérer, entre autres l'étude des modèles de régression et les techniques d'analyses de la variance.

Dans tous ces cas plus structurés, des données non représentatives sont rencontrées et il est tout aussi important d'être capable de reconnaître, d'interpréter et d'accommoder les valeurs aberrantes en utilisant les techniques statistiques appropriées.

Avec de telles données structurées, deux complications apparaissent : les valeurs aberrantes sont moins apparentes intuitivement car elles sont plus cachées dans la masse de données et les méthodes statistiques pour le rejet ou l'accommodation sont beaucoup moins développées (Barnett, Lewis, 1994).

Lors de l'examen des données, destiné à rechercher les observations étonnamment éloignées du reste du groupe de données, l'estimation de la relation entre les valeurs aberrantes et le groupe principal est plus complexe dans le cas d'une analyse de régression que dans le cas univarié. Pour les modèles de régression, les observations présentant les résidus les plus élevés sont susceptibles d'être aberrantes mais il existe une différence par rapport à la nature des valeurs aberrantes dans les échantillons univariés car les résidus ne sont pas indépendants ; ceci peut rendre la détection des valeurs aberrantes plus délicate. Des tests de discordance sur divers types de résidus (résidus estimés, résidus estimés pondérés) sont disponibles dans la littérature.

Dans le cas du modèle linéaire général, à partir duquel les problèmes de régression et d'analyse de variance sont traités, des résidus estimés peuvent également fournir des mesures appropriées pour l'examen de valeurs anormales.

Afin de définir les résidus utilisés, il est nécessaire de parler de la matrice de projection **H**[6]. Soit le modèle linéaire suivant :

$$\mathbf{Y} = \hat{\beta}\mathbf{X} + \mathbf{e}$$

où **Y** est le vecteur de la variable dépendante,
X, la matrice des variables explicatives,
$\hat{\beta}$, le vecteur des paramètres de régression et
e, le vecteur des variations aléatoires (résidus).

En ajustant ce modèle par les moindres carrés, on obtient le vecteur des valeurs estimées $\hat{\mathbf{Y}}$. La matrice de projection **H** est définie par la relation (Hoaglin, Welsch, 1978) :

$$\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{X}\hat{\beta} = \mathbf{X(X'X)^{-1}X'Y}.$$

Des résidus standardisés, de variance constante et égale à l'unité, peuvent être calculés à partir des résidus estimés, en faisant intervenir l'écart-type de ces résidus et les éléments diagonaux correspondants de la matrice de projection. Un autre type de résidus[7] est obtenu en utilisant la même méthode de calcul mais en éliminant successivement une observation lors de l'ajustement du modèle (Palm, 1988). Des tests de discordance sont utilisés pour examiner les valeurs maximales de ces résidus afin de détecter des valeurs aberrantes (Barnett, Lewis, 1994).

Hoaglin et Welsch (1978) proposent d'examiner les éléments diagonaux de la matrice de projection **H** pour la recherche de valeurs aberrantes. Les éléments diagonaux de cette matrice mesurent l'effet de levier

6 En anglais : *hat matrix*

7 En anglais : *studentized residuals*

de chacune des observations sur la valeur estimée. Les observations qui présentent un grand effet de levier correspondent à des points isolés du groupe principal des données. L'examen de la matrice **H** peut révéler des observations particulièrement sensibles pour lesquelles les valeurs de **Y** sont susceptibles d'avoir un effet important sur l'ajustement.

La technique de détection de valeurs aberrantes pour les modèles linéaires généralisés prend en compte les éléments diagonaux de la matrice de projection et les résidus studentisés, les deux méthodes étant complémentaires. L'application de ces techniques est détaillée dans Palm (1988, 2002).

Palm (1988) et Cook et Weisberg (1982) détaillent les différentes possibilités pour rechercher les observations influentes (par exemple, la distance de Cook). Il faut cependant remarquer que des observations influentes ne sont pas forcément aberrantes et inversement. Barnett et Lewis (1994) distinguent les notions de valeurs aberrantes et d'observations influentes de manière très simple. Pour ces auteurs, les valeurs aberrantes traitées de manière robuste par des techniques d'accommodation n'influencent pas l'estimation des paramètres tandis que pour les observations influentes, dans les mêmes conditions de traitement, de petits changements dans la position de cette observation peut causer des modifications majeures dans les estimateurs.

5.4. Traitement des valeurs aberrantes dans les séries chronologiques et les données spatiales

Dans le cas des séries chronologiques, la présence de valeurs aberrantes peut causer également des problèmes. Deux catégories distinctes de valeurs aberrantes sont à prendre en compte.

Premièrement, on peut rencontrer des **valeurs aberrantes isolées**, dues à une erreur de mesure ou d'exécution. Ces valeurs ne vont pas "correspondre" aux valeurs adjacentes et l'effet d'une valeur aberrante de ce type n'est pas négligeable sur la prévision d'une valeur future.

Deuxièmement, des valeurs aberrantes, appelées **additives**, peuvent s'insérer discrètement au niveau des données mais n'apparaître que bien plus tard lors de l'étude de la structure des corrélations des observations voisines. Tel est cas pour un appareil de mesure qui est déréglé et fournit des observations consécutives erronées. Une dérive est alors observée au cours du temps.

Bien que l'examen du graphique de la série permet dans un premier temps de détecter des valeurs anormales, c'est le graphique des résidus ou des erreurs de prévision qui permet de mettre en évidence des valeurs anormales. En divisant les résidus ou les erreurs de prévision par l'écart-type résiduel ou l'écart-type des erreurs de prévision, des valeurs standardisées sont obtenues dont la distribution devrait être normale. Des tests appliqués sur ces résidus peuvent prouver statistiquement la discordance de ces données. Un test de discordance sur les valeurs aberrantes additives est présenté par Barnett et Lewis (1994) mais très peu d'information à ce sujet est disponible dans la littérature.

Palm (1992) signale que lorsque des techniques de lissage sont utilisées, les données situées en début de la série ne devraient pas faire l'objet d'une élimination trop rapide car il est naturel que les prévisions soient moins bonnes pour le début de la série.

Pour les données spatiales, le problème peut être comparé au cas des séries chronologiques dans le sens où il existe des relations entre des observations spatialement voisines. Les techniques de traitement de valeurs anormales dans le cas des données de ce type n'ont cependant pas donné lieu récemment à de nombreuses publications. Dans le cadre des études géostatistiques, Cressie (1993) présente des résidus calculés à partir des valeurs prédites au niveau du variogramme. Cet auteur expose également la manière de traiter les valeurs aberrantes dans le cas de modèles spatiaux se présentant sous forme de lattices, définis comme étant un ensemble de sites spatiaux, réguliers ou irréguliers, identifiés par une caractéristique commune (exemple : latitude, longitude). Des résidus calculés à partir d'une médiane lissée et pondérée[8] sont exposés par l'auteur.

5.5. Traitement des valeurs aberrantes dans le cas multivarié

Dans le cas multivarié, il est tout aussi important d'être capable de reconnaître les valeurs aberrantes et de les interpréter en utilisant les techniques statistiques appropriées. Cependant, la détection de valeurs aberrantes est loin d'être simple quand on se trouve dans le cas multidimensionnel. Pour cette raison, les méthodes statistiques de détection sont beaucoup moins développées que dans le cas univarié (Barnett, Lewis, 1994).

Si on possède un échantillon d'observations multivariées, chacune consistant en un vecteur de mesures, il peut y avoir une observation particulière pour laquelle aucune des composantes de mesure n'est surprenante pour sa distribution marginale mais c'est son ensemble de mesures pour la même observation multivariée qui semble étonnamment loin du reste du groupe principal de données. Un vecteur de réponse peut être erroné parce qu'une grosse erreur se trouve dans un de ses composants ou parce que de petites erreurs systématiques se sont glissées dans chacune de ses composantes.

---

[8] En anglais : *weighted-median-polish residuals*

La manière la plus simple de traiter les données multivariées est de considérer les échantillons marginaux, c'est-à-dire d'étudier chacune des composantes univariées. Il ne faut pas sous-estimer l'importance des échantillons marginaux pour la présence de valeurs aberrantes. Il est en effet tout à fait possible que la contamination apparaisse uniquement dans une seule des variables marginales. Il faut cependant être prudent et ne pas adopter une approche trop simpliste dans l'examen des données. Cela pourrait avoir comme conséquence de restreindre son attention aux échantillons marginaux en exploitant les méthodes univariées, avec leurs avantages de représenter uniquement les valeurs aberrantes comme des valeurs extrêmes, et en utilisant les tests de discordance développés dans le cas univarié. Cette manière de travailler ignore la structure des corrélations à l'intérieur des observations multivariées, résultant en une perte d'information.

Avant de réaliser toute analyse, il est nécessaire de vérifier la normalité des observations. Pour des données non normales, il peut être envisagé de réaliser des transformations qui permettent de normaliser celles-ci. Si après transformation, la normalité n'est pas rencontrée, diverses méthodes, tels que les graphiques des quantiles, peuvent aider à découvrir quel est le type de modèle de probabilité à prendre en compte.

Dans le cas de l'une ou l'autre distribution connue, telles que les distributions exponentielle ou de Pareto, Barnett et Lewis (1994) proposent des tests de discordance dans les cas bivariés et pour une valeur aberrante supérieure. Ces tests sont très limitatifs et sont peu applicables dans la pratique. Lorsque la distribution n'est pas identifiée, il est nécessaire de se tourner vers des méthodes graphiques ou non paramétriques. Quelques tests de discordance non paramétriques, c'est-à-dire non basés sur l'hypothèse d'une distribution, ont été développés en ce sens mais sont également très restrictifs au niveau de leurs conditions d'application.

Lorsqu'on se trouve dans le cas de la distribution normale, deux cas bien distincts peuvent se présenter en fonction des situations à traiter. D'une part, les variables sont mises sur le même pied d'égalité, c'est-à-dire qu'elles sont interdépendantes, d'autre part, certaines variables sont considérées comme explicatives et d'autres sont dépendantes.

Dans le premier cas, les variables sont mises sur le même pied d'égalité et diverses méthodes de réduction des données sont présentées dans la littérature. Sur base des paramètres calculés par ces méthodes, des tests de discordance sont appliqués pour tester les observations suspectes. La réduction des données multivariées peut être réalisée selon les trois approches présentées ci-dessous.

– Les mesures de distances généralisées R_j, par exemple, les distances généralisées de Mahalanobis $R_j^2 = \mathbf{(x_j - \bar{x})'S^{-1}(x_j - \bar{x})}$ (Crettaz de Roten, Helbling, 1996). Une généralisation du test de Grubbs à partir de ces distances a été étudiée en détail par Carletti (1988) et des variantes de ce test ont été développées (Barnett, Lewis, 1994). Ces derniers auteurs ont considéré des principes généraux pour la détection des valeurs aberrantes multivariées. Le principe le plus simple est le suivant : la plus extrême des observations x_i, est celle dont l'omission de l'échantillon $x_1, x_2, ..., x_n$ produit la plus grande augmentation dans le sens du maximum de vraisemblance sous F pour le reste des données. Si cette augmentation est exceptionnellement grande, x_n est déclaré comme étant une valeur aberrante. Une observation x_n qui produit la valeur maximale R_n est alors un candidat pour la déclaration d'une valeur aberrante. Une valeur aberrante x_n va être déclarée discordante si R_n est excessivement grand en rapport à la distribution de R_n sous un modèle de base connu. Le principe de test de discordance est donc le même que celui présenté dans le cas d'une valeur aberrante univariée.
– Le calcul des composantes principales z_i (ACP) ; les composantes sont utilisées lors de tests spécifiques de discordance (Devlin *et al.*, 1981) ;
– Le coefficient d'étalement multivarié[9] (noté $b_{2,p}$) proposé par Schwager et Margolin (1982).

Notons que les deux premières méthodes de réduction des données multivariées sont les plus couramment utilisées dans le domaine statistique. Les paramètres de position et de dispersion multivariés (moyenne $\mathbf{\bar{x}}$, matrice de variance-covariance **S**) nécessaires au calcul de ces mesures ou coefficients sont estimés par la méthode du maximum de vraisemblance ou par des méthodes robustes appropriées. Les principes de rognage et de winsorization pour l'estimation des paramètres de position univariés ont été généralisés au cas multivarié (Campbell, 1980 ; Jolliffe, 1989).

Des tests de discordance à partir des R_j, des z_i ou des $b_{2,p}$ sont réalisés en prenant en compte un nombre déterminé de valeurs suspectes, soit une valeur supérieure, soit pour tester en bloc 2, 3, 4 valeurs aberrantes dans un échantillon multivarié (Wilks, 1963). Ce dernier auteur a réalisé la première étude détaillée d'application orientée vers la détection de valeurs aberrantes dans des données multivariées

[9] En anglais : *multivariate sample kurtosis coefficient*.

$$b_{2,p} = n\sum_{j=1}^{n}\left[(x_j - \bar{x})'S^{-1}(x_i - \bar{x})\right]^2$$

normales et a également utilisé le rapport entre distances généralisées obtenues avec ou sans l'élimination de données anormales.

6. ACCOMMODATION DES VALEURS ABERRANTES

L'accommodation consiste en la construction 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 (Barnett, Lewis, 1994). Ce concept se trouve à la base des améliorations récentes dans l'élaboration de méthodes statistiques de traitement 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 de grands ensembles de données. 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 en 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. 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. 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 estimations retiennent les propriétés statistiques de tout un ensemble de distributions possibles (Huber, 1981; Rousseeuw, 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 peuvent être très intéressants 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 par la robustesse face à ces valeurs aberrantes. 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.

En ce qui concerne les valeurs aberrantes dans le cas de modèles de régression ou de l'analyse de la

variance, une technique d'accommodation robuste est proposée par Rousseeuw (1984) qui propose de minimiser la médiane des carrés des résidus plutôt que la somme des carrés des résidus, classiquement utilisée en régression linéaire.

7. APERÇU DE MÉTHODES INFORMELLES DE DÉTECTION DES DONNÉES ABERRANTES

Dans le cas univarié, comme présenté précédemment, de nombreuses méthodes graphiques permettent de signaler la présence de valeurs aberrantes : diagramme de dispersion des observations classées en fonction de leur rang, les boxplots, les graphiques des quantiles de valeurs brutes ou de résidus.

Avec la grande complexité du cas multivarié et les conditions d'application très limitantes des tests de discordance, de nombreuses propositions informelles pour la détection et le traitement des valeurs aberrantes sont apparues. Des représentations graphiques, telles que les diagrammes de dispersion, à deux ou trois dimensions, permettent d'identifier dans certains cas des observations aberrantes, totalement hors du nuage de points ou des observations extrêmes pour l'une ou l'autre variable prise en considération.

Les techniques classiques d'analyses multivariées offrent des possibilités d'identification de valeurs anormales. Les méthodes citées sont l'analyse des variables canoniques, l'analyse discriminante (Campbell, 1978 ; 1982), l'analyse factorielle des correspondances, l'analyse en composantes principales (Zhang *et al.*, 1998). Certaines de ces méthodes sont parfois susceptibles aux valeurs aberrantes et afin d'augmenter l'intérêt de celles-ci, des procédures d'estimations robustes des paramètres sont nécessaires (Zhang *et al.*, 1998 ; Lalor, Zhang, 2001).

La méthode des corrélations permet d'analyser les coefficients de corrélations en enlevant une valeur et en évaluant la variation du coefficient entre deux variables marginales. La variation du coefficient de corrélation permet d'identifier des valeurs aberrantes.

La plupart des méthodes informelles reprennent principalement des méthodes graphiques. Des méthodes basées sur les graphiques des quantiles sont utilisées, suivies du test de Shapiro-Wilks de normalité (W-test). Pour divers types de distances généralisées, des auteurs (Gnanadesikan, Kettering, 1972 ; Barnett, Lewis, 1994) proposent de représenter les distances triées $R_{(j)}(\bar{\mathbf{x}}, \mathbf{S})$ par rapport aux valeurs attendues (graphique des quantiles) des statistiques triées pour un échantillon de taille n d'une distribution χ^2 de p degrés de libertés, où p correspond au nombre de variables prises en compte. Ces graphiques ont été utilisés principalement dans le cas des données bivariées où p=2. Les valeurs aberrantes sont détectées comme étant les observations produisant des valeurs de $R_n(\bar{\mathbf{x}}, \mathbf{S})$, $R_{n-1}(\bar{\mathbf{x}}, \mathbf{S})$, etc. se trouvant relativement éloignées de la ligne droite attendue. Des graphiques des quantiles des racines carrées des Rj triées par ordre croissant sont également intéressants.

Rousseeuw et van Zomeren (1990) proposent une méthode graphique pour détecter les valeurs aberrantes à partir de distances généralisées calculées à partir d'estimateurs très robustes qui permettent de remplacer $\bar{\mathbf{x}}$ et $\mathbf{S}$. Pour les composantes principales (z_i), divers graphiques sont proposés par Barnett et Lewis (1994) : des diagrammes de dispersion des premières ou dernières composantes, des graphiques des quantiles sur les z_i, des graphiques des quantiles de la distribution gamma pour les résidus des z_i. Gnanadesikan et Kettering (1972) discutent cet aspect en détail remarquant que la première composante est très sensible aux valeurs aberrantes en augmentant les variances et covariances. Tandis que les quelques dernières composantes sont sensibles aux valeurs aberrantes ajoutant des dimensions fausses aux données ou en cachant quelques singularités (Gnanadesikan, 1977). La construction de diagrammes de dispersion des z_i pour les premières et dernières composantes permet de montrer des valeurs aberrantes de manière graphique. De plus, des tests univariés de détection de valeurs aberrantes peuvent être réalisés sur les z_i individuels où les valeurs ordonnées des z_i peuvent être utilement placées sur un graphique. Si le nombre de variables est assez grand, Barnett et Lewis (1994) estiment que les transformations linéaires impliquées dans l'ACP peuvent mener à une distribution des z_i approximativement normale. Dans ce cas, des graphiques des quantiles normaux peuvent révéler des valeurs aberrantes comme étant des points extrêmes dans le graphique. Une telle procédure informelle a semblé être d'une aide très précieuse pour l'identification des valeurs aberrantes multivariées. Carletti (1988) a utilisé une méthode graphique de visualisation à deux dimensions et de calcul de manière robuste d'une région de confiance à deux dimensions.

Dans le cas des modèles de régression linéaire multiple, Anscombe et Tukey (1963) considèrent la représentation graphique des résidus, incluant les graphiques de probabilité et les graphiques des valeurs observées par rapport aux valeurs estimées. Cette procédure a des limitations telles que l'intercorrélation des résidus pour la détection des valeurs aberrantes. Un avantage de cette approche simple est qu'on considère les résidus des valeurs extrêmes et pas simplement les valeurs extrêmes absolues.

De telles méthodes informelles en général ne conduisent pas à un test formel de discordance. Ces méthodes doivent donc être considérées comme des procédures initiales d'exploration des données.

8. CONCLUSIONS

Les observations contenues dans les bases de données doivent absolument 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. Cette recherche de valeurs aberrantes doit faire l'objet d'une démarche volontaire car les méthodes de traitement des valeurs anormales ne sont pas incluses de manière automatique dans les procédures statistiques.

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.

Cet article a permis de mettre l'accent sur les diverses notions à prendre en compte lors de l'examen de valeurs aberrantes. Les termes principaux de la définition d'une valeur aberrante ont été développés. Celle-ci correspond à une valeur particulièrement surprenante et, en fonction de l'objectif fixé, est statistiquement discordante dans le contexte d'un modèle de probabilité désigné initialement. L'évolution dans la manière de définir une valeur aberrante a montré l'importance du choix d'un modèle de probabilité. De plus, la nature des valeurs aberrantes (caractère aléatoire ou déterministe) détermine clairement la manière de traiter celles-ci ultérieurement. Les objectifs poursuivis lors de l'examen de valeurs aberrantes sont le rejet, l'incorporation, l'identification, l'accommodation ou la correction.

En fonction de l'objectif à atteindre et de la nature de la valeur anormale, le traitement des données est très différent. Il est donc primordial de déterminer au préalable la nature et les objectifs à poursuivre lors de toute étude de valeurs qui semblent suspectes.

Lorsque l'objectif est de rejeter ou d'identifier une valeur anormale, la discordance de cette valeur est évaluée par des tests statistiques, principalement développés dans les années 1950. Une grande partie des tests de discordance ont été élaborés sous l'hypothèse d'une distribution normale. Néanmoins, certains tests concernent les distributions plus dissymétriques. Des méthodes spécifiques liées à l'accommodation ont été développées plus tardivement dans les années 1980. Celles-ci permettent de minimiser l'influence des valeurs aberrantes lorsque toutes les données sont prises en compte lors de l'analyse statistique des données.

Diverses méthodes de détection de valeurs aberrantes dans le cas multivarié sont rencontrées dans la littérature depuis les années 1970. Certaines d'entre elles font appel à des tests de discordance à partir de mesures de réduction basées classiquement sur l'hypothèse de la distribution normale. En cas de dissymétrie des données, ces tests ne sont pas applicables. De nombreuses méthodes graphiques dans le cas multivarié sont également présentées pouvant être fondées sur des méthodes de réduction dérivées mais sans support d'une hypothèse sur la distribution théorique. Ces dernières techniques sont utilisées plutôt comme méthodes exploratoires des données car elles mettent en évidence des observations multivariées qui sont suspectes par rapport à la majeure partie des données. Ces diverses formes de traitements des données incluant les transformations, l'étude des composantes individuelles marginales des observations, la réduction judicieuse des observations multivariées en quantités scalaires, sous la forme de distances généralisées, ou de combinaisons linéaires de composantes, des changements des coordonnées de base des observations et les méthodes appropriées de représentations graphiques peuvent toutes aider à identifier ou mettre en évidence des observations suspectes. Il est conseillé d'utiliser simultanément plusieurs techniques de détection de valeurs aberrantes, de comparer les résultats des analyses univariées et multivariées et de les combiner pour détecter les échantillons aberrants. Le résultat de la détection des valeurs aberrantes est alors plus efficace et permet de déterminer de manière plus fiable les échantillons aberrants.

Enfin, depuis la fin des années 1990 jusqu'à nos jours, de nombreuses publications sur les distributions particulièrement dissymétriques fournissent des perspectives intéressantes pour la détection de valeurs aberrantes dans ces cas bien spécifiques.

Cependant, on constate, 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*. De même, des contraintes logiques au niveau des données chronologiques, temporelles (par exemple, évolution des nitrates dans le sol) ou spatiales (par exemple, communes avoisinantes), sont peu développées dans la littérature.

Bien que le problème des données influentes ait été 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. Comme évoqué dans l'introduction, 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. À l'avenir, on peut espérer un développement de procédures ou de logiciels spécifiques dans cette voie.

Bibliographie

Afifi AA., Azen SP. (1979). *Statistical analysis: A computer oriented approach*. 2nd ed. New York: Academic Press.

Anscombe FJ., Tukey JW. (1963). The examination and analysis of residuals. *Technometrics* **5**, p. 141–160.

Barnett V., Lewis T. (1994). *Outliers in statistical data*. 3rd ed. New York: John Wiley, .

Beckman RJ., Cook RD. (1983). Outlier.....s' (with Discussion). *Technometrics* **25**, p. 119–163.

Beirlant J., Goegebeur Y. (2000). *Local polynomial maximum likelihood estimation for Pareto-type distributions*. Leuven, Belgium: K.U. Leuven, Department of Applied Economics.

Beirlant J., Teugels JL., Vynckier P. (1996). *Practical analysis of extreme values*. Leuven, Belgium: Leuven University Press, 137 p.

Box GEP., Cox DR. (1964). An analysis of transformations. *J. R. Stat. Soc.* **26**, Ser. B, p. 211–252.

Campbell NA. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Appl. Stat.* **27**, p. 251–258.

Campbell NA. (1980). Robust procedures in multivariate analysis. I. Robust covariance estimation. *Appl. Stat.* **29**, p. 231–237.

Campbell NA. (1982). Robust procedures in multivariate analysis. II. Robust canonical variate analysis. *Appl. Stat.* **31**, p. 1–8.

Carletti G. (1988). *Comparaison empirique de méthodes statistiques de détection de valeurs anormales à une et à plusieurs dimensions*. Gembloux, Belgique : Faculté des Sciences agronomiques de l'État, 225 p.

Cerioli A., Riani M. (1999). The ordering of spatial data and the detection of multiple outliers. *J. Comput. Graphical Stat.* **8** (2), p. 239–258.

Chikkagoudar MS., Kunchur SH. (1987). Comparison of many outlier procedures for exponential samples. *Comm. Stat. Theor. Meth.* **16**, p. 627–645.

Cook RD., Weisberg S. (1980). Characterisations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, p. 495–508.

Cook RD., Weisberg S. (1982). *Residuals and influence in regression*. London: Chapman and Hall.

Cressie NAC. (1993). *Statistics for spatial data* (revised ed.). New York: John Wiley and Sons, 900 p.

Crettaz de Roten F., Helbling JM. (1996). Données manquantes et aberrantes : le quotidien du statisticien analyste de données. *Rev. Stat. Appl.* **44** (2), p. 105–115.

Dagnelie P. (1998). Inférence statistique à une et à deux dimensions. Bruxelles : De Boeck & Larcier, vol. 2, 659 p.

Devlin SJ., Gnanadesikan R., Kettenring JR. (1981). Robust estimation of dispersion matrices and principal components. *J. Am. Stat. Assoc.* **76** (374).

Dixon WJ. (1950). Analysis of extreme values. *Ann. Math. Stat.* **21**, p. 488–506.

Everitt BS. (2002). *The Cambridge dictionary of statistics*. 2nd ed.Cambridge, UK: University Press, Second Edition, 410 p.

Gnanadesikan R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley & Sons, 311 p.

Gnanadesikan R., Kettering JR. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* **28**, p. 81–124.

Grubbs FE. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.* **21**, p. 27–58.

Grubbs FE. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**, p. 1–21.

Hampel F., Ronchetti EM., Rousseeuw P., Stahel WA. (1986). *Robust Statistics*. New York: John Wiley.

Hawkins DM. (1980). *Identification of outliers*. London, England: Chapman and Hall, 188 p.

Hoaglin DC., Welsch RE. (1978). The hat matrix in regression and ANOVA. *Am. Stat.* **32** (1), p. 17–22.

Huber PJ. (1972). Robust statistics: a review. *Ann. Math. Stat.* **43**, p. 1041–1067.

Huber P.J. (1981). *Robust statistics*. New York: John Wiley.

ISO (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**, p. 31–78.

Jolliffe IT. (1989). Rotation of ill-defined principal components. *Appl. Stat.* **38**, p. 139–147.

Kimber AC. (1988). Testing upper and lower outlier pairs in gamma samples. *Comm. Stat. Simulation Comput.* **17**, p. 1055–1072.

Lalor GC., Zhang C. (2001). Multivariate outlier detection and remediation in geochemical databases. *Sci. Total Environ.* **281**, p. 99–109.

Laroche J., Oger R. (1999). *Base de données SOLS. Première synthèse*. Gembloux, Belgique : Faculté universitaire des Sciences agronomiques de Gembloux, Unité de Géopédologie ; asbl Réquasud, 36 p.

Lewis T., Fieller NRJ. (1979). A recursive algorithm for null distributions for outliers: I. Gamma samples. *Technometrics* **21**, p. 371–376.

Munoz-Garcia J., Moreno-Rebollo JL., Pascual-Acosta A. (1990). Outliers: a formal approach. *Int. Statist. Rev.* **58**, p. 215–226.

Palm R. (1988). *Les critères de validation des équations de régression linéaire*. Gembloux, Belgique : Faculté des Sciences agronomiques, 27 p.

Palm R. (1992). *Comment interpréter les résultats d'une série chronologique*. Collection STAT-ITCF. Paris : Institut Technique des Céréales et des Fourrages, 80 p.

Palm R. (2002). Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. *Biotechnol. Agron. Soc. Environ.* **6** (3), p. 143–153.

Rohlf FJ. (1975). Generalisation of the gap test for the detection of multivariate outliers. *Biometrics* **31**, p. 93–101.

Rousseeuw PJ. (1984). Least median of squares regression. *J. Am. Stat. Assoc.* **79** (388), p. 871–880.

Rousseeuw PJ., Bassett GW. (1990). The Remedian: a robust averaging method for large data sets. *J. Am. Stat. Assoc.* **85**, p. 97–104.

Rousseeuw PJ., Leroy AM. (1987). *Robust regression and outlier detection*. New York: John Wiley, 329 p.

Rousseeuw PJ., van Zomeren BC. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* **85** (411), p. 633–639.

Royston JP. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat.* **31** (2), p. 115–124.

Schwager SJ., Margolin B. (1982). Detection of multivariate normal outliers. *Ann. Stat.* **10**, p. 943–954.

Shapiro SS., Wilk MB., Chen MJ. (1968). A comparative study of various tests for normality. *J. Am. Stat. Assoc.* **63**, p. 1343–1372.

Thode HC. (2002). *Testing for normality. STATISTICS: textbooks and monographs*. New York: Marcel Dekker, **vol. 164**, 479 p.

Tietjen GL., Moore RH. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics* **14**, p. 583–597.

Wilks SS. (1963). Multivariate statistical outliers. *Sankhya* A (25), p. 407–426.

Zhang CS., Selinus O., Schedin J. (1998). Statistical analyses for heavy metal contents in till and root samples in an area of southeastern Sweden. *Sci. Total Environ.* **212**, p. 217–232.

Zhang CS., Zhang S. (1996). A robust-symmetric mean: a new way of mean calculation for environmental data. *Geojournal* **40** (1-2), p. 209–212.

(53 réf.)