NETTOYAGE DE DONNÉES DANS LE CAS DE FICHIERS D'ENTREPRISES[1]
recherche de la cohérence transversale

Elizabeth Kremp

Un fichier de données individuelles, appelé aussi données de panel, peut être caractérisé par trois dimensions : le nombre d'individus, le nombre d'informations, c'est-à-dire de variables permettant de caractériser ces individus, et le nombre d'années pour lesquelles ces informations sont disponibles. En plus de ces trois caractéristiques, une quatrième peut être prise en compte, plus difficilement mesurable, qui est la qualité de ces informations.

Le problème du nettoyage d'un échantillon s'est posé dans le cadre de la comparaison des bases de données comptables de la Banque de France par rapport aux données exhaustives SUSE de l'Insee. En effet l'Observatoire des Entreprises de la Banque de France a à sa disposition deux sources de données comptables qui ne sont pas exhaustives : le fichier FIBEN (Fichier Bancaire des Entreprises) et le fichier FPD (Fichier Périodique des Données) de la Centrale de Bilans[2]*. Avant de pouvoir comparer ces différentes bases, il est important de disposer de statistiques fiables*[3]*. De façon plus générale, ce problème de repérage de données extrêmes ou aberrantes se pose lors de la réalisation d'études économiques appliquées qui utilisent des données de panel.*

Après avoir essayé de préciser ces notions de valeurs aberrantes et de valeurs extrêmes, ce travail rappelle les outils statistiques et présente différentes méthodes permettant d'identifier ces valeurs. Huit techniques construites à partir de ces outils et de ces méthodes sont ensuite testées sur la base FIBEN, sur le critère du ratio clients des délais de paiement. Enfin l'application de trois de ces techniques à sept ratios, permet de les comparer, d'évaluer le rôle du choix des ratios et de mesurer les phénomènes cumulatifs d'élimination d'observations.

Parmi ces trois techniques, deux d'entre elles donnent des résultats très proches : celle qui supprime les observations situées à plus de 3 intervalles interquartiles du premier et du troisième quartiles (technique 2) et celle qui applique une méthode de standardisation avec comme estimateur de localisation une moyenne tronquée à 1 % et comme

(1) Ce document n'engage que son auteur et n'est pas l'expression de la position de la Banque de France. Il reprend une étude interne de la Centrale de Bilan : La question du nettoyage de données, D93/01, mars 1993, Banque de France.
(2) L'annexe 1 donne une description succincte de ces deux sources d'information.
(3) La réflexion sur le nettoyage des données pour pouvoir comparer les sources Banque de France et Insee a été menée avec Marie-Christine Parent du Département des statistiques d'entreprises de l'Insee.

estimateur de dispersion le pseudo écart-type (technique 8). La première est plus simple à mettre en œuvre, ce qui peut être une bonne raison pour la préférer. Cependant, si la distribution de la vraie population pour le ratio étudié est très éloignée d'une distribution normale, alors ces deux techniques peuvent conduire à éliminer trop d'observations, et une technique qui n'impose pas un estimateur de dispersion d'une loi normale semble préférable : c'est le cas de la méthode de standardisation utilisant aussi la moyenne tronquée à 1 % mais prenant l'intervalle interquartile comme estimateur de dispersion (technique 7).

1. Valeurs aberrantes et valeurs extrêmes

La première étape de ce travail a consisté en une lecture de la littérature sur la détection de valeurs extrêmes "outliers" et leur traitement. La difficulté à traduire le terme "outliers" reflète et résume en elle-même le problème. En utilisant la terminologie de valeurs extrêmes, on recherche un sous-ensemble de données dont la suppression modifierait beaucoup l'analyse statistique. Mais comme le signalent Gould et Hadi (1993), la suppression automatique des valeurs extrêmes revient à les considérer comme points aberrants alors que les valeurs extrêmes ne sont pas forcément aberrantes. Elles peuvent effectivement correspondre à des erreurs de collecte de données ou des erreurs de frappe, mais elles peuvent également être dues à un mélange de populations. Si les informations sont correctes, elles peuvent être très utiles, et par exemple signaler que les données ne sont pas issues d'une population suivant une loi normale, hypothèse souvent implicite dans les analyses statistiques, ou signaler que le modèle utilisé ne permet pas de prendre en compte l'ensemble des observations. C'est pourquoi l'autre terme parfois utilisé dans la littérature anglo-saxonne est celui de "influential data".

1.1. L'absence de consensus

La démarche à utiliser pour repérer puis traiter ces valeurs extrêmes ne fait pas et ne peut pas faire l'objet d'un consensus dans la littérature car elle dépend de plusieurs paramètres : type de données, méthode de collecte utilisée, taille de l'échantillon, moyens informatiques disponibles, utilisation ultérieure des données (données en coupe ou fichier temporel, étude de statistique descriptive ou étude économétrique...). Par contre le consensus existe sur le travail minutieux de repérage préalable pour essayer de les identifier tout en évitant la suppression d'un trop grand nombre d'observations. Comme le souligne Dormont (1983, p. 102), dans le cas d'une étude économétrique nécessitant un fichier cylindré, l'utilisation "d'un ensemble de critères semblant *a priori* représenter la cohérence et la continuité minimales exigibles peut conduire à un fichier extrêmement restreint".

1.2. Les options

Plusieurs options peuvent être relevées :

- corriger les données s'il s'avère qu'il y a eu une erreur de saisie ;

- analyser les chiffres avec et sans les observations extrêmes et, si l'on décide que les observations extrêmes doivent être éliminées, c'est-à-dire sont aberrantes ou ont trop d'influence dans le cadre de l'étude, bien en spécifier le nombre ;

- remettre en cause la méthode ou le modèle utilisé : utiliser des statistiques et des tests non paramétriques[1] (la médiane, l'intervalle interquartile, tests sur les rangs, tests sur l'égalité de médianes) à la place de statistiques et de tests paramétriques (moyenne, écart-type, tests sur l'égalité des moyennes), utiliser une statistique plus robuste à la présence de valeurs extrêmes (moyenne tronquée, moyenne bipondérée, winsorization, M-estimateurs), utiliser un modèle plus robuste (régression pondérée, régression en utilisant la médiane et non la moyenne).

1.3. La transparence

Deux remarques peuvent être faites sur ces différentes options. L'ensemble de la littérature étudiée insiste sur la transparence nécessaire à tout traitement de valeurs extrêmes ou aberrantes (transparence pour le responsable de l'étude et transparence vis-à-vis du lecteur). Ainsi, si une observation est considérée comme aberrante, elle est retirée de l'échantillon étudié, quelle que soit la statistique utilisée. Il est en effet très important de toujours spécifier le nombre d'observations rentrant dans le calcul d'une statistique, car la comparaison de différentes statistiques pour une même variable peut être d'un grand apport dans l'analyse. Il faut donc être certain que ces statistiques sont comparables et portent sur la même population. Il est aussi important de donner au lecteur, dans la mesure du possible, les moyens d'évaluer le biais de sélection que la suppression de valeurs aberrantes peut introduire (suppression des entreprises défaillantes dans le cas d'un cylindrage sur plusieurs années, suppression de certaines fusions dans le cas d'un nettoyage sur taux de croissance...).

Les modifications d'observations relevées dans la littérature concernent la correction des erreurs (type erreur de saisie), ou l'estimation de valeurs manquantes (quand la suppression des observations ayant des variables avec valeurs manquantes conduirait

(1) Une statistique est dite **paramétrique** si elle fait référence aux paramètres d'une distribution (dont la moyenne et l'écart-type sont les premiers moments). Par opposition, une statistique **non paramétrique** est libre de toute hypothèse sur la distribution, et plus particulièrement de toute hypothèse de normalité. Une statistique est dite **robuste** si elle est peu sensible à la présence des valeurs extrêmes; elle peut être paramétrique (moyenne tronquée), ou non paramétrique (médiane).

à un échantillon très et trop restreint). La modification dans le fichier de la valeur de certains ratios en les "ramenant à une borne" paraît d'une autre nature. D'une part, leur modification déforme les relations existantes entre les différentes variables relatives à la même observation. D'autre part, elle contribue à la constitution de points d'accumulation aux deux extrêmes de la distribution, ce qui va à l'encontre de l'utilisation de méthodes robustes qui ont pour principal objectif de réduire l'influence des queues de distribution trop épaisses. Enfin, elle n'est pas équivalente et va plus loin que l'utilisation de statistiques modifiant la pondération des observations (moyenne tronquée) ou remplaçant les valeurs extrêmes par leurs valeurs adjacentes ("winsorization"), puisqu'aussi bien la moyenne tronquée que la "winsorization" ne modifie pas les valeurs du fichier.

1.4. Cohérence transversale, cohérence temporelle

Comme il a déjà été souligné, le travail d'identification de valeurs extrêmes dépend de nombreux paramètres parmi lesquels le type de données utilisées. La grille de lecture de la littérature statistique a été guidée par le fait que les échantillons considérés ici sont grands et sont utilisés à la fois pour des études de statistiques descriptives et des études économétriques[1]. Les données sont étudiées à la fois dans leur dimension transversale (on parle souvent de coupe), par exemple l'étude d'un secteur donné pour l'année 1992, et dans leur dimension temporelle. La recherche de la cohérence transversale apparaît comme un préalable à la recherche de la cohérence temporelle, d'une part parce qu'un nettoyage ne peut être totalement remis en question, par la mise à disposition d'une nouvelle année d'information, d'autre part parce que la recherche de la cohérence temporelle repose sur le calcul de taux de croissance, qui impose donc le cylindrage partiel (on parle d'échantillons semi-constants) ou total (on ne garde que les observations des entreprises présentes sur toute la période étudiée) ; ce cylindrage entraîne une réduction importante de l'échantillon et l'introduction de biais de sélection. Le travail présenté ici s'est concentré sur l'étude de la recherche de la cohérence transversale. Les mêmes techniques ne peuvent pas être utilisées pour rechercher ces deux types de cohérence, car les variables (et leur distribution) diffèrent : la cohérence en coupe repose sur l'étude de ratios (rarement de variables en niveau), alors que la cohérence temporelle repose sur l'étude des taux de croissance de variables en niveau (par exemple l'emploi, la valeur ajoutée...).

Des travaux préliminaires sur la recherche de la cohérence temporelle à partir du taux de croissance des effectifs ou du chiffre d'affaires montrent que l'application des

(1) Une méthode de nettoyage adaptée à un échantillon de 30 observations, comme celle proposée par Hadi (1992) est difficilement transposable pour un échantillon de plusieurs milliers d'observations.

techniques présentées ici peut être trop sélective et conduire à éliminer beaucoup de petites entreprises ayant connu une forte croissance[1]. Par exemple, une entreprise qui passe de 2 salariés à 10 salariés connaît une croissance de ses effectifs de 400 %, alors que l'intervalle interquartile est de l'ordre de 10 et que le troisième quartile varie entre 3 et 5 suivant les secteurs. L'application de cette technique conduirait à éliminer toutes les entreprises qui ont un taux de croissance de leurs effectifs supérieur à 35 %. Néanmoins ces travaux suggèrent aussi que si la recherche de la cohérence transversale est faite sur des ratios faisant intervenir les effectifs et le chiffre d'affaires (ou la valeur ajoutée), les entreprises du fichier nettoyé qui ont des taux de croissance élevés, correspondent en fait à des observations extrêmes et non pas aberrantes (par exemple des entreprises qui ont connu une restructuration).

2. Outils et méthodes

La littérature est abondante mais très spécialisée. Différentes démarches peuvent être distinguées : celles proposées par les statisticiens, celles proposées par les théoriciens de l'économétrie, et celles utilisées dans les études économétriques appliquées. Un des objectifs de ce travail est d'essayer de faire le lien entre ces différentes types de littérature[2]. Si certains articles mettent l'accent sur l'identification ou la suppression de valeurs extrêmes (un des objectifs étant de déterminer des "cutoffs", c'est-à-dire des seuils à partir desquels les observations seront écartées), alors que d'autres se centrent sur l'utilisation de statistiques ou de modèles robustes à la présence de valeurs extrêmes, ils font tous référence à un certain nombre de concepts dont les principales propriétés sont brièvement rappelées ci-dessous. Cette présentation des outils statistiques préalablement à celle des méthodes est d'autant plus nécessaire que l'opposition entre méthodes qui suppriment les points aberrants et méthodes robustes dont les résultats sont peu influencés par la présence de ces points aberrants s'est avérée peu adéquate. En effet si les méthodes utilisées pour supprimer les points aberrants reposent sur des statistiques peu robustes, le nombre de points aberrants décelés est inversement proportionnel au nombre de points effectivement aberrants. Ces méthodes échouent dans certains cas à identifier des vraies valeurs aberrantes, simplement parce qu'elles dépendent des observations qu'elles sont supposées identifier. **Un des enseignements des tests de ces différentes méthodes est donc qu'il faut utiliser des statistiques robustes dans les méthodes cherchant à identifier les points aberrants.**

(1) La technique 2 a été utilisée pour ces travaux préliminaires (élimination des observations à l'extérieur de l'intervalle [q1-3 eiq, q3+3 eiq]).
(2) Le manuel STATA (1990, p. 298) commence sa présentation des différents diagnostics de points influents en insistant avec regret et excuses sur le jargon utilisé dans cette littérature et l'absence de consensus sur la terminologie utilisée. L'auteur insiste sur le fait qu'une donnée influente va l'être pour une certaine statistique. Ainsi, au lieu de fournir la liste des points influents, il laisse le lecteur avec différentes listes, calculées respectivement avec différentes statistiques; sa présentation propose 13 méthodes ou statistiques pour réfléchir à la notion de points influents. On voit que le consensus est loin d'exister ...

2.1. *Les outils statistiques*

Les outils qui permettent de caractériser une distribution peuvent être regroupés en trois catégories : les estimateurs de localisation, les estimateurs de dispersion, et les statistiques permettant de juger de la forme de la distribution (symétrie, épaisseur des fins de distribution). Ce survol ne prétend pas être exhaustif. Dans le choix des critères présentés, sont retenus ceux qui sont faciles à mettre en œuvre pour de grands échantillons, soit du fait de la simplicité du concept lui-même, soit du fait qu'il est d'utilisation facile dans le logiciel SAS[1].

2.1.1. Les estimateurs de localisation

La moyenne et la médiane sont les deux statistiques de localisation d'une distribution les plus utilisées et les plus faciles à calculer.

La **moyenne** empirique $\overline{X}$ de l'échantillon est un estimateur de la moyenne μ de la population. Elle utilise toutes les données de l'échantillon, est le meilleur estimateur (c'est-à-dire à variance minimum) à distance finie de localisation si les données proviennent d'une distribution normale (et même d'une distribution uni-modale si la taille n de l'échantillon tend vers l'infini et que l'on applique le théorème central limite), mais est très sensible aux valeurs extrêmes.

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{où } X_i \text{ sont les } n \text{ observations de l'échantillon.}$$

La médiane M ne repose que sur la valeur centrale de la distribution et est mieux adaptée aux distributions aux queues longues ou épaisses, ou avec des valeurs aberrantes (elle est par exemple plus efficace dans le cas d'une loi de Laplace, Wonnacott et Wonnacott, 1991, p. 266). Elle correspond au 50 ième percentile ; les 25 ième et 75 ième percentiles sont aussi appelés respectivement premier et troisième quartiles (*q1* et *q3*).

D'autres estimateurs de localisation sont plus robustes et plus adaptés à un large éventail de distributions, pour lesquelles on accepte de faire cependant l'hypothèse de symétrie.

(1) Cette étude a mis en évidence le grand retard de SAS en terme de techniques robustes dans le module SAS/STAT par rapport à d'autres logiciels accessibles sur PC (SPSS, STATA). Du fait de la taille des échantillons concernés, les logiciels sur ordinateur central continuent à avoir notre préférence. Après consultation de l'institut SAS, il apparaît que ces techniques seraient développées dans un module spécialisé (INSIGHT).

La statistique **trimean**[1] a pour objectif d'intégrer des informations plus éloignées du centre que la simple médiane. $TRI = \frac{1}{4}(F_L + 2M + F_U)$, où F_L et F_U sont approximés par les premier et troisième quartiles, et M est la médiane.

La **moyenne tronquée** ou moyenne élaguée est calculée en supprimant de façon arbitraire un certain pourcentage de la distribution : on parle de moyenne tronquée à 5 % (ou 5 % *trimmed mean* dans les logiciels statistiques) pour une moyenne calculée sur un échantillon où les 5 % d'observations les plus faibles et les 5 % d'observations les plus fortes sont supprimées. La médiane est un cas extrême : c'est une moyenne tronquée à 50 %.

La **moyenne bipondérée** de Tukey appartient à la famille des M-estimateurs. La caractéristique de ces M-estimateurs (M pour Maximum de vraisemblance) est de donner un poids plus faible aux points les plus éloignés du centre, sans toutefois utiliser la méthode brutale de la moyenne tronquée qui les élimine. En général, les poids sont une fonction décroissante de la distance par rapport au centre de la distribution. Les autres M-estimateurs les plus connus sont l'estimateur d'Huber (recommandé quand la distribution est proche d'une distribution normale, et sensible lui-même aux valeurs aberrantes), et les estimateurs d'Hampel, et d'Andrew.

La "**winsorization**" de la moyenne, (Sachs, 1984, p. 280) n'est pas présentée dans la liste des statistiques de localisation, bien qu'elle s'y apparente. Elle a été proposée par C.Winsor, "car beaucoup de distributions empiriques sont presque normalement distribuées seulement dans leurs régions centrales", (Sachs, p. 65). Elle consiste, après avoir ordonné les observations, à remplacer les k plus petites observations par la (k + 1) ième plus petite observation et les k plus grandes observations par la (k + 1) ième plus grande observation.

$$x_{wk} = \frac{1}{n}\left\{(k+1)\,x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)\,x_{(n-k)}\right\}$$

Dans la littérature, cet estimateur est rarement utilisé comme estimateur de la moyenne[2].

Rosenberg et Gasko (1983) comparent différents estimateurs de localisation (moyenne, médiane, moyennes tronquées, trimean...) pour différentes distributions symétriques (loi normale, loi de Cauchy caractérisée par des queues épaisses, loi slash). Cette comparaison, faite sur des échantillons de petite taille, conduit à ne préconiser l'usage de la médiane que pour des échantillons de taille inférieure ou égale à 6 ; pour des

(1) appelée aussi "two-sided quartile weighted median" (Sachs, 1984, p. 100).

(2)Par contre, (Mudholkar, 1991) utilise cette méthode de winsorization en l'appliquant au calcul de l'écart-type pour définir un "pooled trimmed-t statistics".

échantillons de taille supérieure, elle recommande l'utilisation de la **moyenne tronquée**. Celle-ci, comme la trimean, n'est malheureusement pas calculée de façon automatique dans le module STAT de SAS.

2.1.2. Les estimateurs de dispersion

L'écart-type (*standard déviation*) s de l'échantillon est un estimateur de l'écart-type σ de la population :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$ où X_i sont les n observations de l'échantillon et $\overline{X}$ sa moyenne.

Il utilise toutes les données de l'échantillon et est le meilleur estimateur de dispersion si la distribution suit une loi normale ; cependant il est encore plus sensible aux valeurs extrêmes que la moyenne puisqu'il est fonction des écarts élevés au carré. Si la plupart des X_i sont relativement proches les uns des autres et un seul est très différent, l'écart-type est contrôlé principalement par celui-là.

L'écart-type de la moyenne empirique (*standard error of the mean*), caractéristique de l'écart entre $\overline{X}$, moyenne empirique de l'échantillon et sa cible μ, moyenne de la population, est calculé à partir de l'écart-type s de l'échantillon et intervient directement dans la calcul de l'intervalle de confiance :

$$SE = \sigma_{\overline{x}} = \frac{s}{\sqrt{n}}$$

En effet, l'intervalle de confiance à 95 % de μ s'écrit :

$$\left[\overline{X} - 1.96\frac{s}{\sqrt{n}}, \overline{X} + 1.96\frac{s}{\sqrt{n}}\right]$$

L'écart absolu moyen, *(Mean Absolute Deviation (AD))*

Il existe deux définitions suivant les auteurs ; l'un est calculé par rapport à la moyenne $\overline{X}$, l'autre par rapport à la médiane M ;

$$\text{AD} = \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}| \text{ ou AD} = \frac{1}{n}\sum_{i=1}^{n} |X_i - M|$$

L'écart absolu médian, *(Median Absolute Deviation MAD)*

$$\text{MAD} = \underset{i}{\text{médiane}} \{ | X_i - M | \}$$

L'écart interquartile (EIQ), *(Interquartile range (IQR), ou Qrange)* est la différence entre le troisième quartile, *q3* et le premier quartile, *q1*[1]. Cet estimateur est le plus souvent utilisé dans la littérature comme estimateur de la dispersion, si l'on craint la présence de points aberrants. Certains auteurs l'utilisent tel quel, d'autres l'utilisent sous la forme de **pseudo écart-type** (***F-pseudosigma*** [cf. ci-dessous]).

Iglewicz (1983) compare ces différents estimateurs de dispersion. Il rappelle que l'écart-type et l'écart absolu moyen sont peu efficaces si la distribution n'est pas normale. L'EIQ et le MAD donnent de meilleurs résultats si la distribution a des queues épaisses, entre autres si elle contient des points aberrants.

2.1.3. La forme de la distribution

La distribution de référence la plus couramment utilisée est la loi normale. Elle est caractérisée par sa fonction de densité en forme de cloche ; elle est uni-modale, symétrique et les queues de distribution sont peu épaisses. Deux statistiques permettent de mesurer la déviation d'une distribution quelconque par rapport à celle d'une loi normale en termes de symétrie et de queues de distribution :

La **"skewness"** mesure les déviations par rapport à la symétrie de la loi normale. La formule précise fait intervenir les moments d'ordre trois. Une approximation peut être obtenue avec la formule suivante : skewness $= \dfrac{3(\bar{X} - M)}{s}$ où $\bar{X}$ est la moyenne, M, la médiane et s, l'écart-type.

Une distribution parfaitement symétrique a une moyenne égale à sa médiane, c'est-à-dire une skewness égale à 0. Une distribution a une skewness positive si la partie droite de la densité est plus longue, c'est-à-dire si la moyenne est supérieure à la médiane ; si elle est inférieure, on parle de skewness négative.

La **"kurtosis"** fait référence à l'épaisseur des queues d'une distribution. Une distribution normale a une kurtosis égale à 3. Une kurtosis supérieure à 3 indique que la distribution a des queues épaisses relativement à celles d'une loi normale. Le calcul de la kurtosis fait intervenir les moments d'ordre 4 et est donc très sensible à la présence de valeurs extrêmes.

(1) La littérature anglo-saxonne utilise parfois la notion de Fourth-spread $d_F = F_U - F_L$, qui est équivalente.

2.1.4. Les tests rapides de non normalité

Hamilton (1990, p. 44) propose un test plus robuste que le calcul de la kurtosis pour mesurer la normalité des queues d'une distribution dans le cas d'une distribution symétrique, en comparant l'écart-type avec l'intervalle interquartile :

si l'écart-type > EIQ/1.35, on a des queues épaisses,

si l'écart-type ≈ EIQ/1.35, on a une distribution approximativement normale,

si l'écart-type < EIQ/1.35, on a des queues longues.

Ce coefficient de 1.35 vient de ce que les premier et troisième quartiles d'une distribution normale sont égaux respectivement à $\mu - 0.6745\,\sigma$, et $\mu + 0.6745\,\sigma$, et donc l'intervalle interquartile est $1.349\,\sigma$. L'écart-type d'une loi normale est donc égal à EIQ/1.349. Ce ratio est connu sous le nom de ***F-pseudosigma*** (Emerson et Hoaglin, 1983, p. 41) ou de ***pseudo écart-type*** (*Pseudo-standard deviation* (*PSD*)), Hamilton (1991, STB3, p. 16).

Ce test est peu fiable si la distribution est très asymétrique car l'asymétrie est déjà en elle-même une preuve de non normalité et ce type de distribution a typiquement une queue allongée et une queue épaisse.

Sachs (1984, p. 325) fournit un test basé sur l'étendue (***"range"***, différence entre la plus grande observation et la plus petite observation) et l'écart-type. Si le rapport entre les deux $\frac{\text{étendue}}{\text{écart type}} = \frac{R}{s}$ est à l'extérieur d'un intervalle de confiance, alors l'hypothèse de normalité doit être rejetée. Les deux bornes définissant cet intervalle sont fonction de la taille n de l'échantillon et du seuil de significativité α. À titre d'exemple, pour $\alpha = 1$ % et $\alpha = 5$ % et pour n = 100 et n = 1000, la région critique est :

α	n	borne inférieure	borne supérieure
1 %	100	4.10	6.36
5 %	100	4.31	5.90
1 %	1000	5.57	7.80
5 %	1000	5.79	7.33

Source : Sachs (1992, p. 328)

2.1.5. Un autre concept souvent utilisé

Une statistique, souvent utilisée dans les travaux de l'Observatoire des Entreprises n'a pas été présentée ci-dessus : le **ratio moyen** *RM*, concept plus macro-économique

rapporte la somme des numérateurs à la somme des dénominateurs, et peut s'interpréter comme une moyenne pondérée de ratios individuels $\frac{X_i}{Y_i}$.

$$RM = \frac{\sum_i X_i}{\sum_i Y_i} = \sum_i \left(\frac{Y_i}{\sum_i Y_i} \right) \times \left(\frac{X_i}{Y_i} \right)$$

2.2. *Les méthodes de nettoyage*

Il est difficile de toujours bien distinguer les méthodes permettant d'**identifier des valeurs extrêmes** de certaines des méthodes parmi celles dites robustes et utilisées pour **réduire l'influence** des valeurs extrêmes. En effet, le rapprochement des résultats obtenus par certaines méthodes robustes de ceux obtenus par les méthodes traditionnelles permettent de détecter l'existence de valeurs aberrantes.

Prenons deux exemples :

- parmi les statistiques dites robustes, la médiane est une statistique non paramétrique, robuste à la présence de valeurs aberrantes, puisqu'elle ne repose que sur une seule valeur, l'observation centrale. Son utilisation ne permet cependant pas d'identifier directement les valeurs aberrantes ;

- par contre, la moyenne tronquée est une statistique paramétrique robuste, qui repose sur les observations centrales (leur nombre dépend du degré de tronquage utilisé). La comparaison des résultats de moyennes tronquées et de la moyenne permet alors de repérer l'existence des observations très influentes. La méthode la plus simple consiste à calculer des moyennes tronquées à différents seuils, puis à comparer les résultats de moyennes tronquées et de la moyenne pour identifier les valeurs influentes. Si ce test, surtout quand il est calculé automatiquement par un logiciel, ce qui n'est malheureusement pas le cas dans le module STAT de SAS, permet de faire une vérification rapide de la qualité de l'échantillon, il ne fait intervenir aucun paramètre de dispersion et peut paraître sommaire.

2.2.1. Le Box plot de Tukey

Cette méthode, attribuée à Tukey, et utilisée dans les graphiques "Box plots" des logiciels statistiques, est basée sur l'écart interquartile EIQ (différence entre le troisième

quartile, *q3* et le premier quartile, *q1*) et distingue deux catégories de valeurs extrêmes déterminées par deux types de bornes (bornes intérieures et bornes extérieures).

Sont considérées comme légèrement extrêmes ("***mild outliers***") toutes les valeurs extérieures à l'intervalle [*q*1 - 1.5 *EIQ*, *q*3 + 1.5 *EIQ*]. Hamilton (1991, STB3, p. 16) considère que ces observations représentent 0,7 % d'une population normale et ne devraient pas être jugées comme alarmantes. Toujours dans le cas d'une loi normale, Emerson et Hoaglin (1983) estiment que la probabilité qu'une observation appartenant à un échantillon de taille n soit en dehors de l'intervalle [*q*1 - 1.5 *EIQ*, *q*3 + 1.5 *EIQ*] est de 0.007 + 0.4/n.

Sont considérées comme très extrêmes ("***severe outliers***"), les valeurs extérieures à l'intervalle [*q*1 - 3 *EIQ*, *q*3 + 3 *EIQ*]. Elles représentent 0,0002 % d'une population normale et ont des effets non négligeables sur le calcul de la moyenne, de l'écart-type s et des autres statistiques classiques.

2.2.2. Repérage des observations influentes sur le calcul de l'écart-type

Belsley, Kuh et Welsh (1980) ont proposé différentes statistiques pour mesurer l'influence de chaque observation sur les estimations des paramètres d'une régression, en supprimant de façon itérative chaque observation : *DFFITS* (mesure de l'influence sur le prédicteur), *DFBETAS* (mesure de l'influence sur le paramètre estimé) et *COVRATIO* (mesure de l'influence sur l'écart-type). Elles sont intégrées dans la plupart des logiciels statistiques (voir par exemple SAS, 1990, pp. 1418-1420, Hamilton,1990, pp. 119-122 pour STATA, et SPSS,1992, pp. 183-187). Pour une régression où l'on régresse y sur x, un point (x_i, y_i) peut être important parce que $\hat{y}_i$ est éloigné de y_i, ou parce que x_i est trop éloigné de l'ensemble des autres x. En appliquant les seuils proposés par Belsley et alii, fonctions à la fois du nombre de paramètres étudiés et de la taille de l'échantillon, ces différentes statistiques donnent dans notre cas des résultats quasi équivalents[1] ; dans les critères de sélection présentés ici, seule la statistique *COVRATIO* est retenue.

Soit $s(i)^2$ la variance estimée après suppression de la *i*ème observation ;

Soit X(i) la matrice *X* sans la *i*ème observation.

La statistique *COVRATIO* mesure le changement dans le déterminant de la matrice de variance-covariance en supprimant la *i*ème observation :

(1) Ceci n'est bien sûr pas toujours le cas et s'explique ici pour deux raisons: la première est que cette méthode est appliquée ici sur le calcul de la moyenne, et la régression effectuée consiste à régresser sur une constante; la deuxième raison est due à la taille des échantillons utilisés.

$$\text{COVRATIO}_i = \frac{\det(s^2(i)(X(i)'X(i))^{-1})}{\det(s^2(X'X)^{-1})}$$

Si p est le nombre de paramètres du modèle et n le nombre d'observations de l'échantillon, Belsley et alii suggèrent que les observations telles que

$|\ \text{COVRATIO}_i - 1\ | \geq 3p / n$ soient considérées avec attention.

2.2.3. Élimination après standardisation de la distribution[1]

La méthode de nettoyage la plus classique consiste à prendre en compte la moyenne et l'écart-type de la distribution pour déterminer des bornes au-delà desquelles les observations sont éliminées ("*cutoffs*"), et repose donc sur la notion d'intervalle de confiance. Pour une distribution arbitraire de moyenne μ et d'écart-type σ, l'inégalité de Bienaymé-Tchebyshev indique que la probabilité que l'écart absolu entre une variable et sa moyenne soit supérieur à $k\ \sigma$ est inférieure ou égale à $1/k^2$:

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geq k\right) \leq \frac{1}{k^2}$$

Dans le cas d'une distribution arbitraire, on atteint un seuil de 5 %, avec $k = 4.47$ (voir Sachs, 1984, pp. 63-64). Dans le cas d'une loi symétrique et uni-modale, l'inégalité plus stricte de Gauss s'applique :

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geq k\right) \leq \frac{4}{9k^2}$$

La même valeur $k = 4.47$ donne alors un seuil de 0,5 %. Enfin dans le cas d'une distribution normale et avec toujours la même valeur de k, le seuil est de 54×10^{-7}.

Ainsi, si la variable étudiée suit effectivement une loi normale, le seuil à partir duquel les observations sont supprimées est très faible. Mais cette méthode peut être utilisée sans qu'il soit nécessaire de faire d'hypothèse sur la loi de la distribution. Au pire, le seuil à partir duquel les observations sont supprimées est de 5 %.

Le problème est que cette méthode utilise deux estimateurs peu robustes aux valeurs aberrantes, et donc identifie comme extrêmes d'autant moins d'observations que la dispersion est grande au départ, c'est-à-dire que le nombre de valeurs aberrantes dans l'échantillon brut est élevé. Cette méthode échoue dans certains cas à identifier des vraies

(1) On appelle standardisation la transformation d'une variable normale quelconque en une variable normale centrée réduite.

valeurs aberrantes, simplement parce qu'elle dépend des observations qu'elle est supposée identifier. De nombreux essais ont été faits sur la base FIBEN pour rendre plus robuste cette méthode. Ces essais consistent à utiliser d'autres paramètres de localisation et de dispersion de la distribution de l'échantillon, plus robustes à la présence des points aberrants. Cela revient à faire des combinaisons des méthodes précédentes.

3. Huit techniques appliquées à la base FIBEN

3.1. La nécessité du travail de repérage des valeurs extrêmes

On peut se demander pourquoi ne pas utiliser des statistiques robustes et quel est l'intérêt d'éliminer certaines valeurs aberrantes. Autrement dit, ne peut-on utiliser des statistiques, des tests et des méthodes économétriques robustes sur un fichier brut, au lieu d'utiliser des statistiques, des tests et des méthodes économétriques classiques sur un fichier nettoyé grâce à des méthodes de nettoyage reposant sur des statistiques robustes ? Plusieurs éléments poussent en faveur de la deuxième solution. Ils tiennent aux propriétés de nos bases, aux utilisations qui en sont faites, et à la formation de l'utilisateur et du lecteur.

Les bases de l'Observatoire des Entreprises permettent la réalisation d'études dont certaines s'adressent à un large public non spécialisé en statistiques. D'autre part si certaines statistiques (ratios moyens, médianes) paraissent moins sensibles que d'autres (moyennes de ratios) à la présence de valeurs extrêmes, elles n'en sont pas néanmoins totalement indifférentes. Si les nombreux contrôles effectués sur les données de la Centrale de Bilans permettent de penser que cette base est plus à l'abri que la base FIBEN des méfaits des valeurs extrêmes, le petit nombre d'observations parfois présentes pour un croisement secteur-taille donné fait que la statistique utilisée peut être très perturbée par la présence d'un seul point.

De plus, l'utilisation de méthodes de régression robustes est plus lourde à utiliser et donc souvent beaucoup plus coûteuse en temps informatique (elles reposent en général sur des méthodes itératives), et nécessite des connaissances statistiques de la part de l'utilisateur et du lecteur plus poussées.

Enfin, comme le montrent les dates de publication des articles de la bibliographie, ces techniques sont encore en pleine évolution. La revue publiée mensuellement par le logiciel STATA, explique que si il n'y a qu'une façon de faire une régression utilisant la méthode des moindres carrés ordinaires (OLS), la méthode de régression robuste "n'est pas unique, et unifiée ; une grande variété d'estimateurs robustes existent, sans

large consensus sur celui qui marche le mieux", Hamilton (1991, STB, p. 21). L'article se termine ainsi : "en résumé, les méthodes robustes ne peuvent dispenser l'analyste de la nécessité d'un travail soigneux de diagnostics, en regardant et réfléchissant sur les résultats de toute analyse".

Avant de présenter une comparaison des différentes techniques de nettoyage appliquées à la base FIBEN, les *tableaux 1 et 2* illustrent l'effet du nettoyage aussi bien sur la base FIBEN que pour le sous-ensemble constituant le fichier FPD de la Centrale de Bilans. Ces calculs sont présentés pour certains secteurs pour l'année 1988[1]. Les exemples retenus montrent que les ratios moyens, tout comme les moyennes de ratios, peuvent être influencés par les valeurs extrêmes. Ainsi, les pratiques qui consisteraient à publier des tableaux où seraient retirés un certain nombre d'observations pour le calcul de la moyenne de ratios individuels (par crainte de valeurs aberrantes), mais à garder ces mêmes observations pour le calcul des ratios moyens, ne sont pas fiables.

Tableau 1 - Comparaison des fichiers FIBEN et FPD, bruts et nettoyés statistiques sur le ratio délais clients

(standardisation avec moyenne tronquée à 1 % et pseudo écart type)

Secteur	Base	Nombre	Moyenne	Ratio moyen	Écart type	Écart type de la m. empirique	P1	Médiane q2	P99	Maximum	Écart inter quartile
U02	FIBEN	4876	53.7	51.0	222.6	3.19	0.0	43.7	179.0	15236	34.1
I.A.A.	FPD	1908	61.2	52.4	349.3	8.00	3.7	46.5	181.8	15236	32.3
	FIBEN net	4807	47.7	50.0	28.2	0.41	0.0	43.3	136.1	159.0	33.2
	FPD net	1883	51.1	51.7	27.2	0.63	3.7	46.2	142.8	158.1	31.6
U05A	FIBEN	7318	89.7	105.7	69.2	0.81	10.4	84.0	245.3	3571.6	45.3
Ind. b.	FPD	2459	91.0	110.5	39.6	0.80	18.0	88.5	239.7	470.1	43.3
équip.	FIBEN net	7232	85.7	100.8	36.5	0.43	10.2	83.5	194.5	235.7	44.7
Profes.	FPD net	2433	88.9	104.4	33.9	0.69	18.0	87.9	187.4	229.9	42.4
U06	FIBEN	11737	87.3	72.1	983.4	9.08	1.0	65.5	228.2	73680	45.0
Ind. b.	FPD	4291	70.3	73.1	45.6	0.70	6.1	67.4	176.3	1342	40.2
Conso.	FIBEN net	11607	67.0	71.1	34.2	0.32	1.0	65.1	169.6	216.8	44.4
Courant	FPD net	4268	68.5	72.3	31.2	0.48	6.1	67.1	161.7	213.9	39.7

Note : Les résultats présentés dans ce tableau sur la ligne FIBEN nettoyé sont les mêmes que ceux présentés dans le tableau 3, huitième technique.

(1) Ces travaux ont été faits dans le cadre d'une comparaison avec la base SUSE de l'Insee et l'année 1988 était la dernière disponible lors de leur démarrage. L'ensemble des tests et des comparaisons ont été réalisés sur les entreprises industrielles (hors énergie). Dans tous les cas les méthodes ont été appliquées par secteur (au niveau de la NAP15, soit six secteurs), pour prendre en compte les hétérogénéités sectorielles. Par contre elles ne sont pas appliquées par critère taille. Lorsque des résultats sont donnés pour la base FPD de la Centrale de bilans, le nettoyage a été effectué sur l'ensemble FIBEN, puis le sous-ensemble FPD a été reconstitué en triant sur la variable type, qui caractérise le type de bilan. Aucun test de nettoyage n'a été effectué directement au niveau de la base FPD; bien évidemment, les résultats seraient quelque peu différents. Enfin, les résultats ne sont présentés ici que pour certains secteurs, tous les secteurs étant donnés dans la première version de ce document daté de mars 1993.

Tableau 2 - Comparaison des fichiers FIBEN et FPD, bruts et nettoyés par tranche d'effectifs - statistiques sur le ratio délais clients

(standardisation avec moyenne tronquée et pseudo écart type)

1988, secteur des biens de consommation courante (U06)

Tranche effectifs	Base	Nombre	Moyenne	Ratio moyen	Écart type	Écart type de la m. empirique	P1	Médiane q2	P99	Maximum	Écart inter quartile
0-19	FIBEN	3876	98.6	67.7	1225	19.68	0.2	62.7	272.5	66326	50.4
	FPD	733	67.8	64.9	54.9	2.03	2.7	62.8	179.3	1050.5	46.7
	FIBEN net	3809	64.8	64.9	37.1	0.60	0.2	61.9	176.9	215.0	48.9
	FPD net	728	65.1	63.3	36.1	1.34	2.7	62.4	169.6	206.6	46.5
20-499	FIBEN	7626	81.4	71.4	850.4	9.74	1.5	66.5	204.1	73680	42.9
	FPD	3356	70.7	72.7	44.3	0.76	7.3	67.5	176.3	1342.0	39.0
	FIBEN net	7564	67.9	70.1	32.8	0.38	1.5	66.0	163.7	216.8	42.5
	FPD net	3338	69.0	71.4	30.3	0.52	7.3	67.6	155.5	213.9	38.6
>=500	FIBEN	235	91.7	74.4	271.2	17.69	11.3	75.5	145.3	4210.3	35.7
	FPD	202	73.4	74.1	26.3	1.85	12.5	74.9	138.5	145.3	33.8
	FIBEN net	234	74.1	74.4	26.9	1.76	11.3	75.3	139.7	149.1	35.6
	FPD net	202	73.4	74.1	26.3	1.85	12.5	74.9	138.5	145.3	33.8

Note : On rappelle que le nettoyage a été effectué toutes tailles confondues, seul le calcul des statistiques est ensuite réalisé par tranche d'effectifs.

Le *tableau 1* compare différentes statistiques calculées pour le ratio délais clients[1]. Les statistiques de la première ligne sont calculées sur la base FIBEN non nettoyée, celles de la deuxième ligne le sont sur le sous-ensemble constitué par les entreprises adhérentes à la Centrale de Bilans. Les troisième et quatrième lignes présentent les mêmes statistiques sur les deux mêmes bases après élimination des observations considérées comme aberrantes en appliquant la méthode de standardisation au ratio délais clients (avec comme indicateur de localisation la moyenne tronquée à 1 % et comme estimateur de dispersion le pseudo écart-type). La première colonne de chiffres donne le nombre d'entreprises rentrant dans le calcul des différentes statistiques.

Dans le secteur **des industries agro-alimentaires (U02)**, la suppression de 1,4 % des observations de FIBEN (69 observations), ou de 1,3 % des observations du FPD (25 observations) fait passer la moyenne de ratios individuels de 54 à 48 jours pour FIBEN, et de 61 à 51 jours pour le FPD. Les ratios moyens sont dans ce cas peu sensibles au nettoyage. Par contre, l'écart-type passe de 349 à 27 pour les entreprises du FPD, alors que l'intervalle interquartile passe de 32,3 à 31,6. Ainsi, l'intervalle de confiance permettant de tester des différences de moyennes sur la base FPD de la Centrale de bilans est très fortement réduit ; il passe de [45 jours, 77 jours] à [50 jours, 52 jours]. Sans nettoyage, la moyenne des délais clients en 1988 est connue avec une **incertitude de 30 jours** ; avec nettoyage et en supprimant 25 observations sur 1908, la précision est de deux jours.

(1) La définition la plus simple du ratio délais clients [Créances clients (ligne bx du bilan) sur chiffres d'affaires TTC (fl + yy)], a été retenue pour pouvoir facilement comparer avec les données SUSE de l'Insee.

Le secteur **des industries des biens d'équipement professionnel (U05A),** fournit un exemple ou le ratio moyen est lui-même affecté de 5 à 6 jours par le nettoyage, aussi bien dans la base FIBEN que dans la base FPD.

Dans le secteur **des Industries des Biens de Consommation courante (U06),** le nettoyage fait baisser la moyenne de ratios de 20 jours pour les données FIBEN, et de près de deux jours pour le FPD, réduisant ainsi fortement l'écart entre les deux sources.

La même comparaison par tranche d'effectifs montre que quel que soit le secteur, ce sont les petites entreprises qui sont le plus sensibles au nettoyage. Le *tableau 2* fournit cette comparaison pour le **secteur U06, industries des biens de consommation courante**. Pour la base FIBEN, la baisse de 20 jours de la moyenne des ratios individuels pour l'ensemble du secteur constaté dans le tableau 1 entre le fichier FIBEN brut et le fichier FIBEN nettoyé est le résultat d'une baisse de 32 jours pour les entreprises de moins de 20 salariés (suppression de 67 entreprises, soit 1,7 %), de 13 jours pour les P.M.E. (suppression de 62 entreprises, soit 0,8 %) et de 18 jours pour les grandes entreprises, cette dernière baisse n'étant due à la suppression que d'une seule entreprise.

3.2. Définition des techniques et des ratios

Le problème du nettoyage d'un échantillon s'est posé dans le cadre de la comparaison des bases de données comptables de la Banque de France avec les données exhaustives SUSE de l'Insee. En effet, étudier la représentativité d'une base A par rapport à une base B avant d'avoir écarté les valeurs apparemment aberrantes des bases A et B n'a pas grand sens. Dans une première étape, les comparaisons de ces deux bases ont été faites en éliminant les points extrêmes avec la méthode de standardisation, que nous avions utilisée dans des travaux précédents (cf. Kremp et Mairesse, 1992, Mairesse et Kremp, 1993), avec comme indicateurs de localisation et de dispersion, la moyenne et l'écart-type. Cette méthode a été appliquée à 7 ratios, sur les données de FIBEN, et parallèlement à l'Insee sur les données SUSE. Elle s'est avérée peu fiable (et n'est pas présentée dans les tableaux), car ses résultats sont trop dépendants de la dispersion des observations dans l'échantillon brut. Si la dispersion est très grande, cette méthode délimitera des bornes très lointaines pour éliminer les points extrêmes. Du fait que la dispersion des écarts types est beaucoup plus importante que la dispersion des moyennes, la première partie de cette recherche s'est portée sur les moyens de réduire l'influence des points aberrants sur l'écart-type (techniques 4, 5 et 6). Ensuite pour des raisons de cohérence, il a semblé logique d'utiliser aussi des estimateurs robustes de localisation si des estimateurs robustes de dispersion sont utilisés (techniques 7 et 8).

A partir des outils et des trois méthodes présentés ci-dessus, huit façons de nettoyer un fichier, appelés par la suite techniques, ont été testées, dont certaines sont une simple

application d'une des méthodes, d'autres sont une combinaison de méthodes ou de méthodes et d'outils.

Les deux premières techniques correspondent à l'application de la méthode expliquée ci-dessus, et attribuée à Tukey. La première montre les conséquences de la suppression de toutes les observations situées à plus de 1.5 écart interquartile du premier et du troisième quartiles. La deuxième technique montre les conséquences de la suppression de toutes les observations situées à plus de 3 écarts interquartiles. Ces deux techniques sont les plus simples à mettre en œuvre.

Les techniques 3 et 4, utilisent la méthode de Belsley et alii. La technique 3 (BKW) donne les résultats de la suppression des valeurs ayant une trop forte influence sur le critère de *COVRATIO*, c'est-à-dire sur le calcul de l'écart-type. La technique 4 (STD/BKW) consiste en l'application de la méthode de standardisation, une fois les points extrêmes, repérés par la technique 3, supprimés.

Les techniques 5 à 8 appliquent la méthode de standardisation avec différents estimateurs de localisation et de dispersion[1] :

Technique 5 : moyenne et écart interquartiles (eiq) ;

Technique 6 : moyenne et pseudo écart-type (psd, intervalle interquartile divisé par 1.349) ;

Technique 7 : moyenne tronquée à 1 % (tm1) et écart interquartile (eiq) ;

Technique 8 : moyenne tronquée à 1 % (tm1) et pseudo écart-type (psd = eiq/1.35) ;

Ces huit techniques ont été testées sur la batterie des sept ratios utilisés lors de l'étude de représentativité des bases de l'Observatoire des entreprises avec les données exhaustives de SUSE de l'insee. Ces sept ratios sont les suivants :

R1 = marge brute d'exploitation = excédent brut d'exploitation/chiffre d'affaires hors taxes

R2 = taux de valeur ajoutée = valeur ajoutée/production

R3 = marge d'autofinancement = capacité d'autofinancement nette/chiffre d'affaires hors taxes

R4 = dettes financières/fonds propres

R5 = fonds propres/total bilan

R6 = délais clients

R7 = délais fournisseurs

3.3. Comparaison des huit techniques sur le ratio délais clients

Pour chaque secteur de l'industrie au niveau NAP15, pour l'année 1988, toutes tailles confondues, le nettoyage a été fait pour chacun de ces sept ratios avec les huit différentes techniques. Ensuite, l'effet de ces différentes techniques a été évalué en regardant les résultats par tranche d'effectifs, en distinguant trois tranches, les moins de 20 salariés, les PME de 20 à 500 salariés, les grandes entreprises de plus de 500 salariés.

La version préliminaire de ce document fournit l'ensemble des comparaisons de ces huit techniques appliquées au ratio délais clients pour les six secteurs[1]. A titre d'illustration le *tableau 3* présente les résultats, toutes tailles confondues, de trois des six secteurs étudiés. La première ligne rappelle les valeurs des différentes statistiques pour l'échantillon brut.

Pour tous les secteurs, la première technique (suppression de toutes les observations situées à plus de 1,5 écart interquartiles du premier et du troisième quartiles) supprime le plus d'observations. Le pourcentage d'entreprises supprimées sur le ratio délais clients varie entre 1,8 % pour le secteur de la construction automobile et 4,6 % pour le secteur des industries des biens d'équipement ménagers.

La deuxième technique (suppression de toutes les observations situées à plus de 3 écarts interquartiles du premier et du troisième quartiles) supprime entre 0,9 % et 2,3 % des observations sur le ratio délais clients, avec une grande homogénéité d'un secteur à l'autre, puisque 5 des 6 secteurs ont un taux entre 0,9 et 1,1 %. Le secteur ayant le taux le plus élevé de 2,3 % est encore une fois le secteur des industries des biens d'équipement ménagers.

La technique 3 de Belsley, qui permet d'identifier les observations ayant une grande influence sur le calcul de l'écart-type n'est pas suffisante pour écarter les points aberrants. Ainsi, dans le secteur des industries de biens de consommation courante, une entreprise ayant un délai client de 1609 jours n'a pas été identifiée comme extrême.

Les techniques 4, 5, et 6 ont comme caractéristique commune l'utilisation d'un indicateur de dispersion robuste tout en conservant l'indicateur de localisation traditionnel qu'est la moyenne. Les techniques 7 et 8, lourdes à mettre en œuvre car SAS ne calcule pas automatiquement de moyenne tronquée, donnent dans la plupart des cas

(1) La méthode de standardisation pour un estimateur de localisation μ et un estimateur de dispersion σ consiste à éliminer les observations X telles que

$$\left| \frac{X - \mu}{\sigma} \right| \geq 4.47.$$

Tableau 3 - Comparaison des huit techniques sur le ratio délais clients fichier FIBEN

Secteur	Base	Nombre	% conservé	Moyenne	Ratio moyen	Écart type	Écart type de la m. empirique	P1	Médiane q2	P99	Max.	Écart inter quartile
U02	FIBEN brut	4876	.	53.7	51.0	222.6	3.2	0	43.7	179.0	15 236	34.1
I.A.A.	1:q3+1.5 eiq	4664	95.7	45.1	48.0	24.3	0.4	0	42.4	106.8	114.4	31.3
	2:q3+3 eiq	4811	98.7	47.8	50.1	28.4	0.4	0	43.3	137.3	164.6	33.2
	3:BKW	4873	99.9	50.0	51.0	35.2	0.5	0	43.7	176.9	539.8	34.0
	4: STD avec BKW	4851	99.5	48.9	50.8	30.8	0.4	0	43.6	158.0	207.1	33.8
	5: STD avec eiq	4850	99.5	48.9	50.8	30.8	0.4	0	43.6	156.5	202.4	33.7
	6: STD avec psd	4812	98.7	47.8	50.1	28.5	0.4	0	43.4	137.4	165.8	33.2
	7: STD avec tm1 et eiq	4847	99.4	48.8	50.8	30.5	0.4	0	43.6	154.3	200.0	33.7
	8: STD avec tm1 et psd	4807	98.6	47.7	50.0	28.2	0.4	0	43.3	136.1	159.0	33.2
U05A Industries biens équipmt professionnel	FIBEN brut	7318	.	89.7	105.7	69.2	0.8	10.4	84.0	245.3	3 571.6	45.3
	1:q3+1.5 eiq	7083	96.8	83.4	96.4	33.0	0.4	9.9	82.8	164.6	175.1	43.6
	2:q3+3 eiq	7239	98.9	85.9	100.8	36.8	0.4	10.2	83.6	197.1	242.9	44.7
	3:BKW	7242	99.0	86.0	100.9	36.9	0.4	10.2	83.6	198.1	244.1	44.7
	4: STD avec BKW	7242	99.0	86.0	100.9	36.9	0.4	10.2	83.6	198.1	244.1	44.7
	5: STD avec eiq	7269	99.3	86.6	105.3	38.4	0.4	10.2	83.7	210.0	288.3	44.9
	6: STD avec psd	7234	98.9	85.8	100.8	36.6	0.4	10.2	83.5	194.8	238.2	44.7
	7: STD avec tm1 et eiq	7269	99.3	86.6	105.3	38.4	0.4	10.2	83.7	210.0	288.3	44.9
	8: STD avec tm1 et psd	7232	98.8	85.7	100.8	36.5	0.4	10.2	83.5	194.5	235.7	44.7
U06 Industries biens consommation courante	FIBEN brut	11737	.	87.3	72.1	983.4	9.1	1.0	65.5	228.2	73680.0	45.0
	1:q3+1.5 eiq	11420	97.3	65.2	70.2	31.2	0.3	0.8	64.5	142.4	155.8	43.3
	2:q3+3 eiq	11616	99.0	67.1	71.3	34.4	0.3	1.0	65.1	171.0	222.6	44.4
	3:BKW	11730	99.9	70.1	72.1	51.7	0.5	1.0	65.4	222.2	1609.4	44.9
	4: STD avec BKW	11685	99.6	68.3	71.7	37.3	0.3	1.0	65.3	195.5	299.5	44.7
	5: STD avec eiq	11677	99.5	68.1	71.7	36.8	0.3	1.0	65.3	194.0	285.7	44.7
	6: STD avec psd	11635	99.1	67.4	71.4	35.0	0.3	1.0	65.2	176.1	235.8	44.5
	7: STD avec tm1 et eiq	11664	99.4	67.9	71.6	36.2	0.3	1.0	65.3	189.1	267.8	44.6
	8: STD avec tm1 et psd	11607	98.9	67.0	71.1	34.2	0.3	1.0	65.1	169.6	216.8	44.4

des résultats très proches des techniques 5 et 6 ; cependant elles sont plus cohérentes que ces dernières, puisqu'elles utilisent aussi un estimateur robuste de localisation. Dans un petit nombre de cas, les techniques 5 et 6 ne détectent pas des valeurs extrêmes, ou au contraire, éliminent toutes les observations (Cf. rôle du choix des ratios, *tableau 4*, secteur des biens d'équipement ménagers).

Le choix entre la technique 7 et la technique 8 dépend de la forme de la distribution de la variable testée. Si il y a de bonnes raisons de croire que cette distribution est très éloignée de la distribution normale, et si une transformation de la variable permettant de rapprocher sa distribution d'une loi normale ne peut être envisagée, alors la technique 8 peut conduire à éliminer trop d'observations.

Les techniques 2 et 8 donnent des résultats extrêmement proches, toutes tailles confondues et par tranche d'effectifs dans le cas du ratio des délais clients. La technique 2 étant beaucoup plus simple à mettre en œuvre, ce peut être une bonne raison pour préférer cette technique si ce résultat se confirme sur d'autres variables et sur les données SUSE de l'Insee.

Enfin, il faut noter le comportement très différent du secteur U05B, qui s'explique en partie par le plus petit nombre d'observations de ce secteur. Même après nettoyage, l'hétérogénéité dans ce secteur reste très forte. Une attention particulière doit lui être portée.

La comparaison par tranche d'effectifs (non reproduite ici), confirme que quelle que soit la technique employée, les entreprises de moins de 20 salariés sont plus touchées par le nettoyage que les entreprises de plus grande taille.

3.4. Le rôle du choix des ratios

La comparaison des huit techniques appliquées à un même ratio a déjà fourni quelques enseignements. Le tableau 4 compare les taux d'acceptation des huit techniques appliquées par secteur à sept ratios. Moins complet que le précédent puisqu'il résume une technique à son taux d'acceptation, il permet d'évaluer la sensibilité des techniques aux différents ratios.

La première constatation est que quelle que soit la technique utilisée, les taux d'acceptation varient beaucoup d'un ratio à l'autre : ils dépendent plus du ratio concerné que du secteur (mis à part le cas du secteur des biens d'équipement ménagers déjà signalé ci-dessus). Le ratio des dettes financières sur fonds propres (R4) est celui pour lequel le plus fort pourcentage d'entreprises est rejeté pour six techniques sur huit (entre 8 % et 11 % suivant les secteurs en appliquant la technique 2).

A l'opposé, le ratio de taux de valeur ajoutée (R2) et le ratio des délais clients (R6) sont pour 6 techniques, ceux pour lesquels le plus faible pourcentage d'entreprises est rejeté (moins de 1 %, pour le ratio R2 en appliquant la technique 2).

La deuxième constatation est que, quel que soit le ratio et quel que soit le secteur, la technique 1 est beaucoup plus sélective que les autres.

La troisième constatation concerne les techniques 3 et 4, qui utilisent la méthode de Belsley et alii. Elles apparaissent fort différentes des autres techniques puisque pour ces deux techniques R3 et R4 n'ont pas les taux de rejet les plus forts. Elles ont un taux de rejet qui varie beaucoup d'un secteur à l'autre pour le même ratio ; ainsi pour le ratio

Tableau 4 - Comparaison des taux d'acceptation selon les huit techniques pour chacun des sept ratios

	1 : [q1-1.5 eiq q3+1.5 eiq]	2 : [q1-3 eiq q3+3 eiq]	3 : Belsley, Kuh, Welsh	4 : standardisation avec BKW	5 : Standardisation avec eiq	6 : Standardisation avec psd	7 : Standardisation avec moyenne tronquée à 1 % et eiq	8 : Standardisation avec moyenne tronquée à 1 % et psd
Secteur des Industries Agro Alimentaires (U02) : 4 876 observations								
R1	88.6	97.4	99.7	99.2	98.9	97.6	98.8	97.7
Marge brute d'exploitation = ebe / chiffre d'affaires								
R2	95.2	99.7	99.8	99.8	99.9	99.7	99.9	99.7
Taux de valeur ajoutée = valeur ajoutée / production								
R3	79.7	92.8	99.7	99.0	96.1	94.0	96.3	93.8
Marge d'autofinancement = capacité d'autofinancement nette / chiffre d'affaires								
R4	81.8	90.6	99.3	97.8	93.5	91.6	93.4	91.4
Dettes financières / fonds propres								
R5	89.0	97.7	97.1	97.1	99.2	98.4	99.2	98.4
Fonds propres / total bilan								
R6	87.7	98.7	99.9	99.5	99.5	98.7	99.4	98.6
Délais clients								
R7	91.6	98.3	99.8	99.2	99.2	98.4	99.1	98.2
Délais fournisseurs								
Secteur des Industries des biens d'équipement ménagers (U05B) : 176 observations								
R1	84.6	94.9	98.9	98.9	0	0	98.3	94.9
R2	92.0	98.3	98.9	98.9	98.3	77.3	98.9	98.9
R3	80.7	92.6	98.9	97.7	0.6	0	96.0	95.5
R4	83.5	89.2	98.9	97.7	4.0	2.3	93.2	90.3
R5	89.8	98.9	97.2	97.2	99.4	99.4	99.4	99.4
R6	88.6	97.7	98.3	98.3	99.4	97.7	99.4	97.7
R7	86.4	98.9	96.6	96.6	100	98.9	100	98.9
Secteur des Industries des biens de consommation courante (U06) : 11 737 observations								
R1	86.1	96.6	99.9	99.7	97.7	93.8	98.5	97.3
R2	89.6	99.6	99.9	99.8	99.7	99.7	99.7	99.7
R3	79.8	92.2	99.9	99.8	96.1	93.7	96.1	93.7
R4	82.5	90.7	99.8	99.0	93.5	91.2	93.7	91.4
R5	90.8	97.8	99.6	99.1	99.2	98.7	99.1	98.7
R6	89.7	99.0	99.9	99.6	99.5	99.1	99.4	98.9
R7	87.5	98.7	99.9	99.6	99.3	98.8	99.2	98.6

R5, le taux de rejet varie entre 0,3 % pour le secteur U06 des industries de biens de consommation courante et 2,9 % pour le secteur U02 des industries agro-alimentaires.

La quatrième constatation est que les techniques 5 et 6, qui n'utilisent pas d'estimateur de localisation robuste, peuvent donner des résultats très bizarres, puisque dans le cas d'un petit secteur avec des valeurs très aberrantes, la technique conduit à éliminer toutes les entreprises (secteur U05B des industries de biens d'équipement ménagers).

Enfin, la dernière constatation est que les techniques 2 et 8, quel que soit le secteur et quel que soit le ratio, donnent des résultats très proches, confirmant les résultats trouvés ci-dessus sur le délais clients.

Le *tableau 5* étudie les phénomènes cumulatifs d'élimination pour les techniques qui apparaissent les plus solides. Pour les six secteurs confondus, la technique 2 appliquée successivement aux sept ratios conserve 83 % des observations, la technique 7 en conserve 89 % et la technique 8, 85 %. Par tranches d'effectifs, ce tableau montre que les plus petites entreprises (moins de 20 salariés) sont les plus atteintes par le nettoyage (79 %, 87 % et 81 % respectivement, contre 87 %, 93 % et 89 % pour les plus de 500 salariés). Il confirme aussi la similitude des résultats obtenus par la technique 2 (trois intervalles interquartile) à l'extérieur de l'intervalle [*q1, q3*]) et la technique 8 (standardisation avec moyenne tronquée à 1 % et pseudo écart-type).

3.5. Comparaison des statistiques après application d'une des trois techniques sur les sept ratios

Les tableaux 6 comparent les trois techniques qui apparaissent les plus solides. La technique 2 élimine les observations à l'extérieur de l'intervalle [q1-3 eiq, q3 + 3 eiq] et est la plus simple à mettre en œuvre. Les techniques 7 et 8 appliquent une méthode de standardisation avec comme estimateurs de localisation une moyenne tronquée à 1 %. Elles diffèrent sur le choix de l'estimateur de dispersion, la première retenant l'intervalle interquartile, la seconde imposant l'écart-type d'une loi normale (le pseudo écart-type). Pour chaque technique, le nettoyage est fait pour les sept ratios ; une observation est éliminée dès qu'elle ne vérifie pas un test pour un des ratios ; ensuite pour chacun des sept ratios, les statistiques sont donc calculées sur le même nombre d'observations pour une technique donnée. Ainsi, le nombre d'observations pour une

(1) Par contre, le tableau 6-6 pour le ratio délais clients n'est pas identique aux tableaux 2 et 3. Dans ces deux tableaux, seules les observations ne vérifiant pas le critère de sélection pour le ratio délais clients ont été éliminées alors que dans le tableau 6.6, les observations ne vérifiant pas le critère de sélection pour un des sept ratios ont été éliminées. Les tableaux 2 et 3 calculent le ratio délais clients sur 98 ou 99% des observations suivant les secteurs, le tableau 6.6 sur 80 à 90% des observations. Ceci n'est pas négligeable, comme le montre la comparaison des résultats pour le secteur des industries de biens d'équipement professionnel. En appliquant la technique 2 à ce seul ratio, le tableau 3 indique un ratio moyen de 100.8, pour 7239 observations. En appliquant cette même technique 2 aux septs ratios, le ratio moyen pour les délais clients, calculé sur 6108 observations (Cf. tableau 6.1). Cela permet de souligner l'importance du choix des ratios pour un nettoyage de l'échantillon.

Tableau 5 - Comparaison pour trois techniques des phénomènes cumulatifs d'élimination pour les sept ratios[1]

	Nombre	Pourcentage d'entreprises supprimées						
Technique 2 :q3+3eiq		**0 fois**	**1 fois**	**2 fois**	**3 fois**	**4 fois**	**5 fois**	**6 fois**
6 secteurs confondus	35855	83,0	12,4	3,0	1,1	0,3	0,1	0,0
I.A.A. (U02)	4876	81,2	14,5	3,0	1,0	0,3	0,0	0,0
Biens intermédiaires (U04)	10959	84,7	11,6	2,5	0,8	0,3	0,1	0,0
Biens de conso. courante (U06)	11737	82,0	12,8	3,4	1,4	0,3	0,1	0,0
Biens d'équipt profess. (U05a)	7318	83,5	11,7	3,3	1,2	0,3	0,1	0,0
Biens d'équipt ménager (U05b)	176	81,3	13,1	2,8	1,1	1,1	0,6	0,0
Construc. auto et MTT (U05c)	789	83,3	12,8	2,7	1,1	0,0	0,0	0,1
Par tranche d'effectifs								
< 20	11798	79,1	15,3	3,5	1,4	0,4	0,2	0,1
20-500	23103	84,9	11,1	2,8	1,0	0,2	0,1	0,0
> 500	954	87,1	9,7	2,5	0,4	0,2	0,0	0,0
Technique 7 : std avec tm1 et eiq		**0 fois**	**1 fois**	**2 fois**	**3 fois**	**4 fois**	**5 fois**	**6 fois**
6 secteurs confondus		89,5	8,6	1,3	0,4	0,1	0,1	0,0
I.A.A. (U02)		88,5	9,7	1,4	0,3	0,1	0,0	0,0
Biens intermédiaires (U04)		90,6	7,9	1,0	0,3	0,1	0,0	0,0
Biens de conso. courante (U06)		89,0	8,8	1,4	0,5	0,1	0,1	0,0
Biens d'équipt profess. (U05a)		89,5	8,5	1,3	0,4	0,2	0,1	0,0
Biens d'équipt ménager (U05b)		89,8	8,5	0,0	0,6	1,1	0,0	0,0
Construc. auto et MTT (U05c)		89,1	9,6	0,9	0,3	0,0	0,0	0,1
Par tranche d'effectifs								
< 20		86,9	10,4	1,7	0,7	0,2	0,1	0,0
20-500		90,7	7,8	1,1	0,3	0,1	0,0	0,0
> 500		93,4	5,7	0,6	0,1	0,2	0,0	0,0
Technique 8 : std avec tm1 et psd		**0 fois**	**1 fois**	**2 fois**	**3 fois**	**4 fois**	**5 fois**	**6 fois**
6 secteurs confondus		84,7	11,8	2,4	0,8	0,2	0,1	0,0
I.A.A. (U02)		83,0	13,3	2,7	0,7	0,3	0,0	0,0
Biens intermédiaires (U04)		86,0	11,0	2,0	0,6	0,2	0,1	0,0
Biens de conso. courante (U06)		83,8	12,4	2,5	1,0	0,2	0,1	0,0
Biens d'équipt profess. (U05a)		85,3	11,1	2,6	0,7	0,2	0,1	0,0
Biens d'équipt ménager (U05b)		82,4	14,2	1,1	1,1	1,1	0,0	0,0
Construc. auto et MTT (U05c)		85,2	12,2	1,8	0,8	0,0	0,0	0,1
Par tranche d'effectif								
< 20		81,0	14,5	2,8	1,2	0,3	0,2	0,0
20-500		86,4	10,6	2,2	0,6	0,1	0,0	0,0
> 500		89,0	8,9	1,8	0,1	0,2	0,0	0,0

(1) La première colonne donne le nombre d'entreprises dans le secteur avant tout nettoyage. La deuxième colonne donne le pourcentage d'entreprises conservées ("éliminées zéro fois"). Les colonnes suivantes donnent le pourcentage d'entreprises éliminées 1 fois (c'est-à-dire pour un seul ratio), deux fois (c'est-à-dire pour deux ratios) ...

technique et son taux d'acceptation correspondant, ne sont donnés que pour le ratio R1 : marge brut d'exploitation. Pour chaque technique ce taux d'acceptation est identique à celui présenté dans le tableau 5[1].

Les résultats sont présentés pour les six secteurs et les sept ratios pour avoir une vue complète des conséquences d'un tel nettoyage. Pour en faciliter tant soit peu la lecture, un certain nombre de chiffres du fichier brut, fortement modifiés par le nettoyage, sont en caractères gras.

La première constatation est que bien que ces trois techniques donnent des résultats différents, les écarts entre les statistiques calculées après application d'une de ces trois techniques sont bien plus faibles que ceux qui séparent ces statistiques de celles calculées sur l'échantillon brut.

La deuxième constatation concerne les écarts types de l'échantillon (s) et de la moyenne empirique ($\sigma_{\bar{X}}$). Quelle que soit la taille du secteur et quel que soit le ratio considéré, l'écart-type de l'échantillon et l'écart-type de la moyenne estimée sont modifiés de façon très importante. Dans quatre des six secteurs, l'écart-type de la moyenne estimée après nettoyage est au moins divisé par 10 (entre 40 et 70 suivant les ratios dans le secteur U06 des industries de biens de consommation courante), ce qui réduit considérablement la longueur des intervalles de confiance. Le nettoyage apparaît un préliminaire indispensable à l'utilisation des tests d'égalité de moyennes.

La troisième constatation est que certains ratios sont plus sensibles que d'autres aux effets du nettoyage. Les statistiques du ratio dettes financières sur fonds propres (R4) sont par exemple fort différentes après nettoyage. La moyenne de ce ratio pour le secteur des industries agro-alimentaires (U02, 4 876 observations) passe de 2,3 avant nettoyage à 1,6 ou 1,7 selon la technique retenue. Pour le secteur des biens d'équipement professionnel (U05A, 7 317 observations), cette moyenne est divisée par trois. Le taux de valeur ajoutée (R2), qui comme l'a montré le tableau 4 est celui qui conduit à éliminer le moins d'observations, connaît une variation de la moyenne de ratios de 10 points dans le secteur des industries de biens de consommation courante (U06, 11 737 observations).

Enfin, les moyennes de ratios ne sont pas les seules statistiques sensibles aux valeurs extrêmes. Pour chacun des sept ratios, on peut trouver un secteur où l'écart entre le ratio moyen calculé sur l'échantillon brut et celui calculé après application de n'importe laquelle des trois techniques est important. Ainsi, le ratio moyen du taux de marge brut d'exploitation (R1) pour le secteur des industries de biens d'équipement ménager (U05B) passe de 3,2 % sur le fichier brut à 8,3 ou 8,4 % pour les fichiers nettoyés. Toujours dans le même secteur, le ratio moyen pour le taux de valeur ajoutée (R2) passe de 26,1 % à 30,1 %. Pour le ratio dettes financières sur fonds propres (R4), le ratio moyen du secteur des industries de biens de consommation courante (U06) passe de 87 % pour le fichier brut à entre 75 et 77 % pour les fichiers nettoyés.

Deux remarques peuvent être faites sur le ratio moyen. D'une part, l'écart entre le ratio moyen calculé sur le fichier brut et ceux calculés sur les fichiers nettoyés est plus important pour les secteurs ayant relativement peu d'observations (secteur des industries de biens d'équipement ménager (U05B, 176 observations), secteur des constructeurs automobiles (U05C, 788 observations) que pour les autres. D'autre part, les écarts sur le ratio moyen sont plus importants quand les statistiques sont calculées par tranche d'effectifs. Ceci s'explique par le fait que toutes tailles confondues, les grandes entreprises dominent dans le calcul du ratio moyen, et, comme il a déjà été souligné, elles appartiennent à la catégorie la moins touchée par un nettoyage.

Que dire sur le choix de la technique elle-même ? Ce tableau montre la proximité des résultats des techniques 2 et 8 avec des taux de rejet entre 15 % et 20 % suivant les secteurs, taux un peu plus élevés que ceux de la technique 7 (entre 10 % et 12 %). Or la technique 8, par définition, a pour conséquence de rapprocher la distribution de l'échantillon de celle d'une loi normale. Si la vraie population pour le ratio étudié a une distribution très éloignée de celle d'une loi normale, les techniques 2 et 8 ont tendance à éliminer trop d'observations.

La dernière colonne des tableaux 6 donne un test rapide de non normalité, puisque, comme il a été dit ci-dessus, si l'intervalle interquartile rapporté à l'écart-type vaut 1,35, la distribution est approximativement normale. Cette dernière colonne fournit donc plusieurs indications. La marge d'autofinancement (R3) est le ratio pour lequel ce rapport est le plus éloigné de 1.35 (entre 0,8 et 1). C'est aussi le ratio (Cf. tableau 4) qui entraîne le plus grand taux de rejet. Ce ratio ne suit sûrement pas une loi approximativement normale et en utilisant des techniques comme la technique 2 ou la technique 8 qui tentent de rapprocher sa distribution d'une loi normale, beaucoup d'observations sont éliminées. Par contre pour des ratios comme les délais clients ou fournisseurs (R6 ou R7), pour lesquels le tableau 4 montre des taux d'acceptation très proches pour les techniques 2, 7 et 8, ce rapport vaut 1.3, confirmant que leur distribution peut être approximée par une loi normale.

Tableaux 6 : Comparaison de trois techniques pour les sept ratios - Année 1988 données Fiben par secteur

Tableau 6-1 : R1 Marge brut d exploitation = excèdent brut d'exploitation / chiffre d'affaires

		nombre	pct obs.	moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/ s
U02 Industries	brut	4876	.	0.050	0.067	**0.391**	0.006	-21.5	-.178	0.023	0.049	0.088	0.282	0.989	0.065	0.2
agricoles	q3+3eiq	3960	81.2	0.063	0.068	0.051	0.001	-0.10	-.028	0.028	0.052	0.088	0.221	0.282	0.060	1.2
alimentaires	STD tm1, eiq	4314	88.5	0.062	0.068	0.056	0.001	-0.21	-.055	0.026	0.051	0.088	0.240	0.347	0.062	1.1
	STD tm1, psd	4047	83.0	0.060	0.065	0.051	0.001	-0.16	-.038	0.027	0.051	0.086	0.217	0.276	0.060	1.2
U04 Industries	brut	10959	.	0.084	0.117	**0.408**	0.004	-34.6	-.189	0.050	0.085	0.132	0.360	0.993	0.082	0.2
des biens	q3+3eiq	9282	84.7	0.100	0.126	0.065	0.001	-0.10	-.022	0.056	0.088	0.133	0.304	0.376	0.076	1.2
intermédiaires	STD tm1, eiq	9934	90.6	0.099	0.121	0.072	0.001	-0.26	-.059	0.054	0.087	0.133	0.328	0.457	0.079	1.1
	STD tm1, psd	9430	86.0	0.098	0.121	0.065	0.001	-0.15	-.046	0.055	0.087	0.131	0.295	0.364	0.076	1.2
U05A Indus.	brut	7318	.	**0.039**	0.101	**1.989**	0.023	-153	-.281	0.036	0.069	0.114	0.319	24.659	0.078	0.0
biens	q3+3eiq	6108	83.5	0.085	0.118	0.063	0.001	-0.12	-.047	0.043	0.074	0.118	0.277	0.345	0.075	1.2
d équipement	STD tm1, eiq	6550	89.5	0.082	0.108	0.070	0.001	-0.27	-.090	0.040	0.072	0.117	0.291	0.418	0.076	1.1
profess.	STD tm1, psd	6241	85.3	0.082	0.110	0.064	0.001	-0.18	-.069	0.041	0.073	0.116	0.273	0.332	0.074	1.2
U05B Indus.	brut	176	.	**-0.987**	**0.032**	**11.331**	0.854	-144	-42.6	0.035	0.071	0.120	0.410	0.411	0.085	0.0
biens	q3+3eiq	143	81.3	0.090	0.084	0.070	0.006	-0.07	-.043	0.045	0.078	0.128	0.340	0.358	0.083	1.2
d équipement	STD tm1, eiq	158	89.8	0.083	0.083	0.086	0.007	-0.31	-.253	0.040	0.074	0.122	0.358	0.411	0.083	1.0
ménagers	STD tm1, psd	145	82.4	0.087	0.084	0.067	0.006	-0.07	-.054	0.045	0.078	0.123	0.294	0.340	0.078	1.2
U05C	brut	789	.	0.075	0.107	0.089	0.003	-1.02	-.111	0.039	0.068	0.108	0.297	0.807	0.069	0.8
construction	q3+3eiq	657	83.3	0.082	0.112	0.054	0.002	-0.06	-.017	0.044	0.074	0.111	0.248	0.302	0.067	1.2
auto, autres	STD tm1, eiq	703	89.1	0.080	0.111	0.059	0.002	-0.14	-.047	0.042	0.070	0.110	0.258	0.356	0.068	1.2
matér. transp.	STD tm1, psd	672	85.2	0.081	0.111	0.054	0.002	-0.07	-.033	0.043	0.072	0.110	0.248	0.302	0.067	1.2
U06 Industies	brut	11737	.	**-0.044**	0.089	**7.536**	0.070	-766	-.271	0.036	0.069	0.111	0.322	0.965	0.075	0.0
biens	q3+3eiq	9626	82.0	0.082	0.095	0.059	0.001	-0.11	-.046	0.044	0.074	0.113	0.263	0.333	0.069	1.2
consommation	STD tm1, eiq	10446	89.0	0.079	0.093	0.067	0.001	-0.25	-.089	0.041	0.072	0.112	0.281	0.404	0.072	1.1
courante	STD tm1, psd	9834	83.8	0.079	0.093	0.060	0.001	-0.17	-.070	0.042	0.072	0.111	0.258	0.320	0.069	1.1

Tableau 6-2 : R2 Taux de Valeur ajoutée = valeur ajoutée / production

		moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	0.248	0.218	0.338	0.005	-16.6	-.039	0.134	0.229	0.348	0.736	0.997	0.214	0.6
agricoles	q3+3eiq	0.249	0.224	0.153	0.002	-0.26	0.000	0.135	0.227	0.338	0.689	0.923	0.203	1.3
alimentaires	STD tm1, eiq	0.252	0.224	0.159	0.002	-0.66	-.011	0.135	0.228	0.343	0.711	0.993	0.208	1.3
	STD tm1, psd	0.248	0.219	0.154	0.002	-0.38	-.013	0.133	0.224	0.336	0.688	0.923	0.203	1.3
U04 Industries	brut	0.420	0.345	**1.022**	0.010	-8.37	0.042	0.305	0.415	0.524	0.785	104.333	0.220	0.2
des biens	q3+3eiq	0.419	0.349	0.152	0.002	-0.09	0.088	0.310	0.417	0.522	0.771	0.993	0.212	1.4
intermédiaires	STD tm1, eiq	0.418	0.346	0.155	0.002	-0.26	0.083	0.308	0.416	0.523	0.776	0.998	0.215	1.4
	STD tm1, psd	0.418	0.346	0.152	0.002	-0.26	0.087	0.309	0.416	0.520	0.771	0.993	0.211	1.4
U05A Indus.	brut	**0.389**	0.388	**1.315**	0.015	-91.2	0.062	0.308	0.410	0.509	0.806	20.447	0.201	0.2
des biens	q3+3eiq	0.419	0.398	0.147	0.002	0.01	0.111	0.316	0.414	0.512	0.795	0.989	0.196	1.3
d équipement	STD tm1, eiq	0.417	0.391	0.148	0.002	-0.19	0.102	0.314	0.413	0.512	0.795	0.989	0.197	1.3
profess.	STD tm1, psd	0.417	0.394	0.148	0.002	-0.19	0.103	0.314	0.413	0.511	0.793	0.989	0.196	1.3
U05B Indus.	brut	**-0.169**	**0.261**	**5.211**	0.393	-60.8	-32.1	0.259	0.360	0.456	0.684	0.832	0.197	0.0
des biens	q3+3eiq	0.374	0.301	0.137	0.011	0.09	0.100	0.277	0.366	0.472	0.684	0.832	0.194	1.4
d équipement	STD tm1, eiq	0.369	0.301	0.145	0.012	-0.28	0.089	0.274	0.368	0.467	0.684	0.832	0.193	1.3
ménagers	STD tm1, psd	0.372	0.301	0.138	0.011	0.09	0.100	0.274	0.365	0.467	0.684	0.832	0.193	1.4
U05C	brut	0.362	0.282	0.141	0.005	-0.51	0.032	0.275	0.359	0.444	0.768	0.877	0.169	1.2
construction	q3+3eiq	0.366	0.284	0.129	0.005	-0.00	0.080	0.277	0.363	0.446	0.721	0.875	0.169	1.3
auto, autres	STD tm1, eiq	0.365	0.284	0.131	0.005	-0.00	0.080	0.277	0.361	0.445	0.721	0.875	0.168	1.3
matér. transp.	STD tm1, psd	0.366	0.284	0.128	0.005	-0.00	0.080	0.278	0.362	0.445	0.721	0.875	0.168	1.3
U06 Industries	brut	**0.309**	0.337	**6.892**	0.064	-721	0.015	0.280	0.391	0.504	0.905	2.491	0.224	0.0
des biens	q3+3eiq	0.406	0.341	0.169	0.002	-0.07	0.072	0.290	0.396	0.504	0.899	0.996	0.214	1.3
consommation	STD tm1, eiq	0.405	0.341	0.171	0.002	-0.11	0.062	0.288	0.395	0.504	0.900	0.996	0.217	1.3
courante	STD tm1, psd	0.405	0.340	0.170	0.002	-0.11	0.068	0.288	0.396	0.504	0.900	0.996	0.216	1.3

Tableau 6.3 : R3 = Marge d autofinancement

		moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	0.018	0.018	**0.728**	0.010	-22.0	-.250	-0.006	0.005	0.025	0.179	35.322	0.031	0.0
agricoles	q3+3eiq	0.013	0.018	0.029	0.000	-0.07	-.059	-0.002	0.007	0.026	0.103	0.118	0.028	1.0
alimentaires	STD tm1, eiq	0.011	0.018	0.035	0.001	-0.13	-.086	-0.004	0.006	0.025	0.119	0.146	0.029	0.8
	STD tm1, psd	0.011	0.016	0.030	0.000	-0.09	-.074	-0.003	0.006	0.024	0.096	0.110	0.028	0.9
U04 Industries	brut	0.023	0.032	**0.729**	0.007	-9.63	-.281	-0.003	0.015	0.043	0.248	68.750	0.046	0.1
des biens	q3+3eiq	0.025	0.040	0.041	0.000	-0.09	-.077	0.001	0.017	0.044	0.150	0.181	0.043	1.0
intermédiaires	STD tm1, eiq	0.023	0.036	0.049	0.000	-0.19	-.117	-0.001	0.016	0.044	0.178	0.224	0.045	0.9
	STD tm1, psd	0.023	0.037	0.043	0.000	-0.13	-.095	0.000	0.016	0.043	0.145	0.172	0.043	1.0
U05A Indus.	brut	**-0.009**	-0.009	**1.214**	0.014	-54.3	-.424	-0.011	0.009	0.037	0.205	47.412	0.048	0.0
des biens	q3+3eiq	0.019	0.015	0.044	0.001	-0.11	-.090	-0.004	0.013	0.040	0.151	0.179	0.043	1.0
d équipement	STD tm1, eiq	0.015	-0.000	0.053	0.001	-0.21	-.152	-0.006	0.012	0.039	0.164	0.221	0.045	0.8
profess.	STD tm1, psd	0.016	0.001	0.047	0.001	-0.15	-.118	-0.005	0.012	0.038	0.144	0.166	0.043	0.9
U05B Indus.	brut	-1.064	**-0.054**	**11.005**	0.830	-138	-49.7	-0.021	0.007	0.036	0.302	0.342	0.057	0.0
des biens	q3+3eiq	0.016	-0.002	0.046	0.004	-0.09	-.085	-0.011	0.014	0.043	0.143	0.153	0.054	1.2
d équipement	STD tm1, eiq	0.010	-0.004	0.055	0.004	-0.18	-.154	-0.013	0.011	0.041	0.143	0.153	0.055	1.0
ménagers	STD tm1, psd	0.013	-0.002	0.051	0.004	-0.15	-.137	-0.013	0.011	0.042	0.143	0.153	0.055	1.1
U05C	brut	0.015	0.029	0.101	0.004	-1.58	-.224	-0.003	0.013	0.039	0.277	0.623	0.041	0.4
construction	q3+3eiq	0.021	0.033	0.035	0.001	-0.08	-.062	0.001	0.016	0.039	0.127	0.151	0.038	1.1
auto, autres	STD tm1, eiq	0.019	0.032	0.043	0.002	-0.15	-.109	-0.001	0.015	0.039	0.149	0.196	0.040	0.9
matér. transp.	STD tm1, psd	0.019	0.033	0.038	0.001	-0.12	-.103	0.000	0.015	0.039	0.127	0.151	0.039	1.0
U06 Industries.	brut	**0.003**	0.014	**17.544**	0.162	-1325	-.408	-0.012	0.007	0.032	0.220	1342.17	0.045	0.0
biens	q3+3eiq	0.015	0.020	0.041	0.000	-0.10	-.085	-0.006	0.010	0.034	0.139	0.167	0.040	1.0
consommation	STD tm1, eiq	0.011	0.018	0.050	0.000	-0.19	-.148	-0.008	0.008	0.034	0.159	0.205	0.042	0.8
courante	STD tm1, psd	0.012	0.018	0.043	0.000	-0.14	-.113	-0.007	0.009	0.033	0.130	0.153	0.040	0.9

Tableau 6-4 : R4 = dettes financières / fonds propres

		pct obs.	moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	.	**2.33**	0.97	**32.93**	0.47	-793	-25.4	0.27	1.08	2.57	37.18	865.51	2.30	0.1
agricoles	q3+3eiq	81.2	1.68	0.92	1.99	0.03	-4.04	-2.71	0.41	1.15	2.39	8.50	9.47	1.98	1.0
alimentaires	STD tm1, eiq	88.5	1.63	0.94	2.44	0.04	-8.35	-5.21	0.34	1.10	2.40	9.76	12.04	2.07	0.8
	STD tm1, psd	83.0	1.63	0.95	2.09	0.03	-5.59	-3.85	0.37	1.12	2.38	8.50	9.49	2.00	1.0
U04 Industries	brut	.	1.39	0.86	**52.62**	0.50	-1574	-14.6	0.23	0.75	1.76	22.36	4708.00	1.52	0.0
des biens	q3+3eiq	84.7	1.15	0.77	1.27	0.01	-2.80	-1.32	0.29	0.77	1.61	5.51	6.33	1.32	1.0
intermédiaires	STD tm1, eiq	90.6	1.16	0.81	1.60	0.02	-5.45	-3.04	0.26	0.75	1.64	7.02	8.13	1.38	0.9
	STD tm1, psd	86.0	1.12	0.78	1.33	0.01	-3.70	-2.24	0.28	0.76	1.60	5.54	6.38	1.32	1.0
U05A Indus.	brut	.	**3.23**	0.91	**93.40**	1.09	-1458	-12.0	0.17	0.66	1.69	24.66	6946.38	1.52	0.0
des biens	q3+3eiq	83.5	1.05	0.69	1.25	0.02	-2.80	-1.36	0.22	0.66	1.50	5.50	6.25	1.28	1.0
d équipement	STD tm1, eiq	89.5	1.07	0.80	1.54	0.02	-5.48	-2.65	0.20	0.66	1.54	6.60	8.12	1.34	0.9
profess.	STD tm1, psd	85.3	1.03	0.73	1.32	0.02	-3.71	-2.24	0.21	0.66	1.51	5.57	6.35	1.30	1.0
U05B Indus.	brut	.	**-7.67**	**1.12**	**215.49**	16.24	-2644	-113	0.14	0.67	1.55	137.17	1054.75	1.42	0.0
des biens	q3+3eiq	81.3	0.90	0.74	1.09	0.09	-2.45	-1.30	0.21	0.61	1.14	4.53	4.92	0.93	0.9
d équipement	STD tm1, eiq	89.8	1.01	0.78	1.51	0.12	-4.55	-4.20	0.21	0.67	1.30	7.15	7.39	1.09	0.7
ménagers	STD tm1, psd	82.4	0.97	0.74	1.15	0.10	-2.45	-1.30	0.22	0.67	1.19	4.53	4.92	0.98	0.8
U05C	brut	.	**1.53**	1.10	**31.47**	1.12	-713	-18.6	0.21	0.65	1.66	31.28	347.27	1.45	0.0
construction	q3+3eiq	83.4	1.03	1.02	1.16	0.05	-2.45	-0.53	0.25	0.65	1.45	5.56	5.89	1.20	1.0
auto, autres	STD tm1, eiq	89.1	1.08	1.07	1.50	0.06	-4.73	-2.78	0.24	0.66	1.51	6.69	7.44	1.26	0.8
matér. transp.	STD tm1, psd	85.2	1.02	1.03	1.22	0.05	-2.94	-2.03	0.25	0.65	1.45	5.64	5.95	1.20	1.0
U06 Industries.	brut	.	0.92	0.87	**94.38**	0.87	-9447	-16.2	0.18	0.76	1.96	29.48	2075.00	1.78	0.0
biens	q3+3eiq	82.0	1.24	0.75	1.50	0.02	-3.38	-1.92	0.26	0.78	1.78	6.42	7.30	1.52	1.0
consommation	STD tm1, eiq	89.0	1.25	0.76	1.91	0.02	-6.48	-3.99	0.23	0.77	1.82	7.99	9.42	1.60	0.8
courante	STD tm1, psd	83.8	1.21	0.77	1.58	0.02	-4.37	-2.69	0.24	0.78	1.78	6.50	7.37	1.54	1.0

Tableau 6-5 : R5 = fonds propres / Total bilan

		moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	0.191	0.289	0.276	0.004	-3.05	-.778	0.082	0.188	0.323	0.781	1.000	0.240	0.9
agricoles	q3+3eiq	0.239	0.292	0.183	0.003	-0.39	-.268	0.123	0.213	0.336	0.754	0.941	0.214	1.2
alimentaires	STD tm1, eiq	0.221	0.289	0.213	0.003	-0.85	-.462	0.110	0.205	0.332	0.762	0.958	0.222	1.0
	STD tm1, psd	0.229	0.284	0.193	0.003	-0.60	-.355	0.118	0.209	0.332	0.749	0.941	0.213	1.1
U04 Industries	brut	**0.223**	0.304	0.369	0.004	-28.9	-.525	0.113	0.217	0.347	0.751	0.995	0.234	0.6
des biens	q3+3eiq	0.263	0.317	0.167	0.002	-0.35	-.140	0.145	0.238	0.359	0.720	0.957	0.214	1.3
intermédiaires	STD tm1, eiq	0.250	0.310	0.187	0.002	-0.80	-.297	0.134	0.231	0.357	0.728	0.995	0.222	1.2
	STD tm1, psd	0.256	0.314	0.176	0.002	-0.54	-.247	0.141	0.235	0.357	0.719	0.957	0.216	1.2
U05A Indus.	brut	**0.201**	0.197	0.399	0.005	-18.5	-.550	0.097	0.197	0.327	0.736	0.948	0.230	0.6
des biens	q3+3eiq	0.250	0.238	0.169	0.002	-0.36	-.167	0.132	0.221	0.342	0.722	0.945	0.210	1.2
d équipement	STD tm1, eiq	0.237	0.205	0.186	0.002	-0.76	-.312	0.123	0.213	0.336	0.725	0.945	0.214	1.1
profess.	STD tm1, psd	0.243	0.219	0.176	0.002	-0.55	-.249	0.129	0.217	0.338	0.720	0.945	0.209	1.2
U05B Indus.	brut	0.232	**0.249**	0.213	0.016	-0.90	-.337	0.125	0.226	0.346	0.727	0.830	0.221	1.0
des biens	q3+3eiq	0.274	0.283	0.168	0.014	-0.27	-.123	0.153	0.250	0.368	0.663	0.684	0.216	1.3
d équipement	STD tm1, eiq	0.259	0.279	0.182	0.014	-0.27	-.212	0.144	0.236	0.353	0.684	0.830	0.209	1.2
ménagers	STD tm1, psd	0.272	0.283	0.167	0.014	-0.27	-.123	0.152	0.249	0.368	0.663	0.684	0.216	1.3
U05C	brut	0.222	0.220	0.215	0.008	-1.48	-.512	0.105	0.206	0.339	0.757	0.947	0.234	1.1
construction	q3+3eiq	0.257	0.230	0.155	0.006	-0.27	-.095	0.144	0.236	0.358	0.645	0.801	0.214	1.4
auto, autres	STD tm1, eiq	0.244	0.225	0.173	0.007	-0.67	-.220	0.135	0.227	0.348	0.645	0.815	0.213	1.2
matér. transp.	STD tm1, psd	0.250	0.229	0.167	0.006	-0.54	-.220	0.139	0.232	0.351	0.645	0.801	0.212	1.3
U06 Industries.	brut	**0.187**	**0.289**	0.778	0.007	-67.4	-.837	0.089	0.196	0.337	0.752	0.955	0.248	0.3
biens	q3+3eiq	0.252	0.306	0.182	0.002	-0.41	-.234	0.128	0.223	0.354	0.729	0.955	0.226	1.2
consommation	STD tm1, eiq	0.235	0.306	0.206	0.002	-0.90	-.395	0.116	0.214	0.349	0.733	0.955	0.233	1.1
courante	STD tm1, psd	0.243	0.302	0.192	0.002	-0.61	-.327	0.124	0.219	0.350	0.728	0.955	0.227	1.2

Tableau 6-6 : R6 = Délais clients

		moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	53.7	51.0	**222.6**	3.2	0.0	0.0	29.3	43.7	63.4	179.0	15236.1	34.1	0.2
agricoles	q3+3eiq	47.6	50.2	27.0	0.4	0.0	0.3	29.9	43.7	61.6	134.3	164.6	31.8	1.2
alimentaires	STD tm1, eiq	48.2	50.9	29.0	0.4	0.0	0.1	29.5	43.6	62.3	146.4	200.0	32.9	1.1
	STD tm1, psd	47.6	49.8	27.0	0.4	0.0	0.3	29.8	43.6	61.7	134.2	158.1	31.8	1.2
U04 Industries	brut	**83.3**	79.1	**193.3**	1.8	0.0	8.5	58.2	77.5	97.6	207.0	17235.0	39.4	0.2
des biens	q3+3eiq	78.5	78.7	29.4	0.3	0.0	11.4	59.3	77.9	96.6	158.2	214.5	37.3	1.3
intermédiaires	STD tm1, eiq	78.5	78.6	30.6	0.3	0.0	10.7	58.8	77.6	96.9	163.6	251.3	38.1	1.2
	STD tm1, psd	78.1	78.6	29.4	0.3	0.0	10.8	58.9	77.6	96.5	157.2	207.3	37.6	1.3
U05A Indus.	brut	89.7	105.7	69.2	0.8	0.0	10.4	61.9	84.0	107.2	245.3	3571.6	45.3	0.7
des biens	q3+3eiq	84.8	96.8[21]	34.6	0.4	0.0	12.1	62.1	83.1	104.6	184.6	239.7	42.5	1.2
d équipement	STD tm1, eiq	85.8	104.8	36.6	0.5	0.0	11.7	62.1	83.4	105.5	198.5	288.3	43.4	1.2
profess.	STD tm1, psd	84.7	99.2	34.7	0.4	0.0	11.1	61.9	83.0	104.7	184.0	235.7	42.8	1.2
U05B Indus.	brut	86.9	77.2	79.1	6.0	0.0	1.1	53.2	78.2	102.1	290.2	938.7	48.9	0.6
des biens	q3+3eiq	74.9	76.1	33.9	2.8	0.0	1.1	51.5	77.4	97.0	168.7	201.5	45.5	1.3
d équipement	STD tm1, eiq	81.5	76.8	45.3	3.6	0.0	1.1	52.7	78.4	101.6	268.6	290.2	48.9	1.1
ménagers	STD tm1, psd	75.2	76.1	33.7	2.8	0.0	1.1	51.6	77.4	97.0	167.0	168.7	45.4	1.3
U05C	brut	69.1	45.0	41.8	1.5	0.1	5.4	45.6	66.0	86.9	216.7	626.2	41.3	1.0
construction	q3+3eiq	66.7	43.2	29.0	1.1	0.1	4.8	46.7	66.8	86.0	137.4	173.1	39.3	1.4
auto, autres	STD tm1, eiq	67.5	44.2	30.1	1.1	0.1	5.4	46.7	67.0	87.7	138.4	216.7	41.1	1.4
malér. transp.	STD tm1, psd	66.8	43.2	29.1	1.1	0.1	4.8	46.7	66.7	86.5	137.4	173.1	39.8	1.4
U06 Industries.	brut	87.3	72.1	**983.4**	9.1	0.0	1.0	43.5	65.5	88.4	228.2	73680.0	45.0	0.0
biens	q3+3eiq	66.7	71.1	32.0	0.3	0.0	2.0	44.6	65.5	86.6	157.3	222.6	42.0	1.3
consommation	STD tm1, eiq	67.4	71.5	34.3	0.3	0.0	1.6	44.1	65.4	87.2	171.0	266.7	43.1	1.3
courante	STD tm1, psd	66.4	71.0	32.0	0.3	0.0	1.6	44.3	65.3	86.4	156.2	213.4	42.2	1.3

Tableau 6-7 : R7 = Délais fournisseurs

		moyenne	ratio moyen	écart type s	écart type de la moyenne	minimum	p1	q1	q2	q3	p99	maximum	eiq	eiq/s
U02 Industries	brut	**66.4**	51.3	**220.1**	3.2	0.0	4.4	33.3	50.8	74.4	231.7	10281.8	41.1	0.2
agricoles	q3+3eiq	54.7	50.4	31.5	0.5	0.0	5.2	32.2	48.5	70.2	163.0	196.6	38.0	1.2
alimentaires	STD tm1, eiq	56.5	51.0	34.2	0.5	0.0	5.1	32.6	49.7	72.1	176.7	234.2	39.5	1.2
	STD tm1, psd	55.1	50.1	31.6	0.5	0.0	4.7	32.3	48.8	70.6	163.0	194.6	38.3	1.2
U04 Industries	brut	96.2	76.6	**84.6**	0.8	0.0	19.6	70.9	90.4	112.2	237.2	6690.0	41.3	0.5
des biens	q3+3eiq	91.7	75.7	31.9	0.3	0.0	21.2	70.7	89.6	109.8	185.9	235.0	39.1	1.2
Intermédiaires	STD tm1, eiq	92.6	75.8	33.9	0.3	0.0	20.4	70.7	89.9	110.6	199.5	277.8	39.9	1.2
	STD tm1, psd	91.8	75.6	31.9	0.3	0.0	21.0	70.7	89.7	110.0	185.9	230.2	39.2	1.2
U05A Indus.	brut	**102.7**	99.6	**135.5**	1.6	0.0	22.0	74.1	95.9	118.8	256.2	10838.6	44.8	0.3
des biens	q3+3eiq	96.8	92.5	35.0	0.4	0.0	22.5	73.9	94.9	116.3	196.5	253.0	42.4	1.2
d équipement	STD tm1, eiq	97.9	98.7	37.1	0.5	0.0	22.1	73.9	95.3	117.1	216.0	295.1	43.2	1.2
profess.	STD tm1, psd	96.9	94.3	35.0	0.4	0.0	22.6	73.9	95.0	116.5	198.2	247.4	42.6	1.2
U05B Indus.	brut	94.6	76.5	**39.9**	3.0	20.7	22.1	67.3	88.7	107.9	229.9	264.3	40.6	1.0
des biens	q3+3eiq	91.1	74.3	36.8	3.1	20.7	22.1	65.1	86.2	107.3	204.4	204.6	42.2	1.1
d équipement	STD tm1, eiq	93.9	74.9	39.8	3.2	20.7	22.1	66.6	87.2	108.1	204.6	264.3	41.4	1.0
ménagers	STD tm1, psd	91.5	74.4	36.7	3.0	20.7	22.1	66.6	86.6	107.3	204.4	204.6	40.6	1.1
U05C	brut	93.7	77.6	**89.0**	3.2	0.0	18.3	69.1	87.6	108.1	248.1	2178.2	38.9	0.4
construction	q3+3eiq	87.3	77.7	30.3	1.2	2.0	20.1	68.9	85.8	106.4	161.9	215.5	37.5	1.2
auto, autres	STD tm1, eiq	87.7	77.8	31.3	1.2	2.0	20.1	68.9	86.1	106.6	167.7	248.1	37.7	1.2
matér. transp.	STD tm1, psd	87.4	77.7	30.6	1.2	2.0	20.1	69.1	86.0	106.6	167.7	215.5	37.5	1.2
U06	Brut	**97.6**	74.6	**1129.1**	10.4	0.0	10.6	58.5	79.8	103.0	256.4	122045	44.5	0.0
biens de	q3+3eiq	80.4	72.2	34.7	0.4	0.0	11.0	57.3	78.1	99.6	183.4	235.9	42.3	1.2
consommation	STD tm1, eiq	82.0	73.3	37.2	0.4	0.0	11.0	57.9	78.8	100.7	203.5	282.6	42.9	1.2
courante	STD tm1, psd	80.6	72.3	34.7	0.3	0.0	11.1	57.5	78.3	99.8	183.1	230.8	42.3	1.2

Conclusion

La construction d'une base de données "nettoyée" suppose d'une part un consensus sur la méthode de nettoyage à appliquer et d'autre part un consensus sur les ratios sur lesquels cette méthode devrait être appliquée. L'objet du présent papier se concentre sur le premier point, c'est-à-dire la comparaison de différentes techniques de nettoyage dans une optique de cohérence transversale d'un fichier. Après avoir rappelé certains concepts et outils nécessaires, ce travail décrit différentes méthodes d'identification de valeurs extrêmes puis teste huit techniques définies à partir de ces trois méthodes et appliquées sur la base FIBEN.

Ces comparaisons de techniques sur différents ratios et le travail effectué en parallèle avec le Département des statistiques d'entreprises de l'Insee montrent que le choix des ratios sur lesquels un nettoyage devrait être appliqué est aussi crucial, et devrait faire l'objet d'une autre étude[1]. Certains des ratios étudiés ici sont sûrement trop fragiles et conduisent à éliminer trop d'observations (ratios faisant intervenir les dettes par exemple) ; par contre des ratios faisant intervenir les effectifs (valeur ajoutée par tête, capital par tête) et qui n'ont pas été utilisés dans ce travail sont de bons ratios pour repérer des points aberrants.

S'il n'existe probablement pas une solution unique qui ferait le consensus de tous les chercheurs-utilisateurs, il y a des pratiques plus ou moins dangereuses. Quelle que soit la technique appliquée parmi les trois finalement retenues, on voit que sur les sept ratios, elle met à l'abri d'un certain nombre d'erreurs. Ces techniques reposent sur des critères qui laissent le moins de place possible à l'arbitraire. Leurs utilisations facilitent les comparaisons inter-temporelles ou inter-fichiers. Elles permettent de plus à l'utilisateur de données comme au lecteur de connaître de façon précise les outils utilisés et les raisons pour lesquelles des observations peuvent être écartées.

(1)Cette étude est en cours au Service de Méthodologie de la Direction des Entreprises de la Banque de France.

BIBLIOGRAPHIE

BELSLEY David, KUH Edwin, et WELSH Roy, *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, John WILEY, New York, 1980.

DORMONT Brigitte, *Modèles de demande de travail : une comparaison France - R.F.A. sur données de panels*, Thèse pour le Doctorat de 3ème cycle en Economie Mathématique et Econométrie, Université de Paris I, 1983.

EMERSON John, HOAGLIN David, « Stem-and-Leaf Displays », p. 7-32, dans Hoaglin et alii (1983), Chapter 1.

GOODALL Colin, « M-Estimators of Location : an Outline of the Theory », p. 339-401, dans Hoagliet et alii (1983), Chapter 11.

GOULD William, HADI Ali, « Identifying Multiple Outliers », *Stata Technical Bulletin,* STB11, p. 28-32, January 1993.

HADI Ali, « Identifying Multiple Outliers in Multivariate Data », *Journal of Royal Statiscal Society*, B 54 (3), p. 761-771, 1992.

HAMILTON Lawrence, *Statistics With Stata,* Brooks/Cole Publishing Company, 1990.

HAMILTON Lawrence, « How Robust is Robust Regression », *Stata Technical Bulletin,* STB2, p. 21-25, 1991.

HAMILTON Lawrence, « Resistant Normality Check and Outlier Identification », *Stata Technical Bulletin*, STB3, p. 16-18, 1991.

HOAGLIN David, MOSTELLER Frederick et TUKEY John, (Eds), *Understanding Robust and Exploratory Data Analysis*, John WILEY, New York, 1983.

HOAGLIN David, MOSTELLER Frederick et TUKEY John, (Eds), *Exploring Data Tables, Trends and Shapes*, John WILEY, New York, 1985.

IGLEWICZ Boris, « Robust Scale Estimators and Confidence Intervals for Location », p. 404-433, dans Hoaglin et alii (1983), Chapter 12.

KRASKER William, KUH Edwin et WELSH Roy, « Estimation for Dirty Data and Flawed Models », dans *Handbook of Econometrics*, Volume 1, p. 651-696, GRILICHES Zvi et INTRILIGATOR Michael (Eds), 1983.

KREMP Elizabeth, « La question du nettoyage des données », Document interne D93/01, Centrale de Bilans, Banque de France, mars 1993.

KREMP Elizabeth et MAIRESSE Jacques, « Dispersion and Heterogeneity of Firm Performances in Nine French Service Industries, 1984-1987 », dans Griliches (Zvi) (ed.), *Output Measurement in the Service Sectors*, Chicago, University Press of Chicago, p. 461-489, 1992.

LI Guoying, « Robust Regression », p. 281-340, dans Hoaglin et alii (1983), Chapter 8.

MAIRESSE Jacques et KREMP Elizabeth, « A look at Productivity at the Firm Level in Eight French Service Industries », *The Journal of Productivity Analysis*, 4, p. 211-234, 1993.

MUDHOLKAR Anil, « A Construction and Appraisal of Pooled Trimmed-t Statistics », *Communications in Statistics, Theory and Methods*, 20 (4), p. 1345-1359, 1991.

SACHS Lothar, *Applied Statistics, A Handbook of Techniques*, Springer-Verlag, New York, 1984.

SAS, *SAS/STAT User's Guide* – Version 6 – 4^{e} édition, volume 2, 1990.

SPSS, *Base System User's Guide*, Release 5.0, 1992.

ROSENBERG James, GASKO Miriam, « Comparing Location Estimators : Trimmed Means, Medians and Trimean », p. 297-336, dans Hoaglin et alii (1983), Chapter 10.

TUKEY John, *Explorating Data Analysis*, John WILEY, New York, 1977.

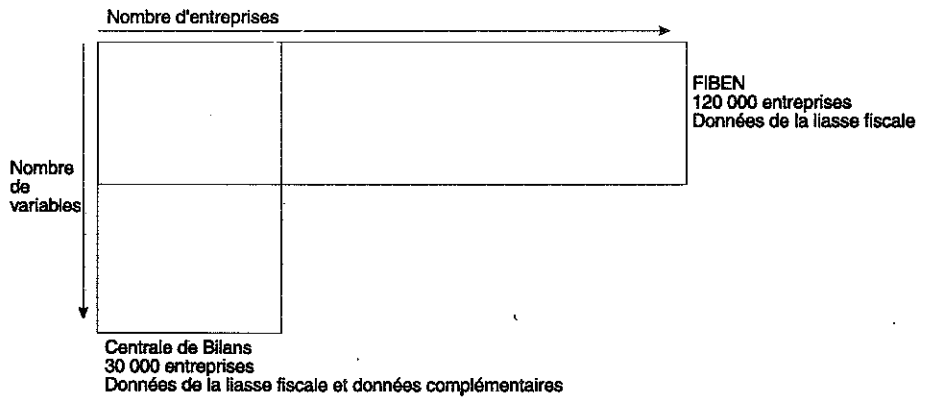
WONNACOTT Thomas et WONNACOTT Ronald, *Statistique*, Economica, 4ème édition, 1991.

ANNEXE 1

Description des bases de données comptables de la Banque de France

FIBEN est un fichier de renseignements créé et géré par la Banque de France pour répondre à ses propres besoins et à ceux des établissements de crédit. Il recense des informations de diverse nature (données descriptives et comptables) sur 1 300 000 entreprises et compte environ 120 000 bilans annuels (liasse fiscale). Ces bilans représentent 96 % des sociétés anonymes et 65 % des SARL existantes en France. Ces données sont conservées 5 ans.

La base de données de la Centrale de Bilans (FPD) comprend des informations descriptives et comptables sur les entreprises adhérentes à la Centrale de Bilans. Les informations recueillies dépendent d'un acte volontaire d'adhésion de la part de l'entreprise, et des relations que la Banque entretient avec ces adhérents. Outre les renseignements nécessaires à son identification et la liasse fiscale (renseignements communs au Fichier FIBEN), le dossier de collecte comprend des feuillets complémentaires portant le détail de certains postes du bilan, de l'endettement et des flux inter-exercices. Ces données sont conservées en ligne sur une période minimum de 10 ans. 30 000 entreprises sont actuellement adhérentes à la Centrale. Du fait de la demande de feuillets complémentaires, de la relation privilégiée entre l'entreprise et la succursale de la Banque de France qui saisit ces données, et des nombreux contrôles de cohérence (environ 400), la base de la Centrale est à la fois plus riche et plus fiable.



ANNEXE 2

Définition des ratios utilisés dans cette étude

Les ratios utilisés dans cette étude ont été définis lors de l'étude de la représentativité des bases de l'Observatoire des Entreprises par rapport aux données exhaustives de SUSE de l'INSEE. Les lignes suivantes donnent leur définition à partir des postes comptables de la liasse fiscale.

R1 = marge brute d'exploitation = excédent brut d'exploitation / chiffre d'affaires hors taxes
= [(FL + FM + FN) – (FU + FS + FV + FT + FW) + FO – (FY + FZ + FX)] / FL

R2 = taux de valeur ajoutée = valeur ajoutée / production
= [(FL + FM + FN) – (FU + FS + FV + FT + FW)] / (FL + FM + FN)

R3 = marge d'autofinancement = capacité d'autofinancement nette/chiffre d'affaires hors taxes
= [GG +(GH-GI) +(GJ + GK +GL-GR) + (GO-GT) +(HA + HB-HE-HF)–HK-FP] / FL

R4 = dettes financières / fonds propres
= (DS + DT + DU + DV) / (DL – DK – AA)

R5 = fonds propres / total bilan
= (DL – DK – AA) / EE

R6 = délais clients
= 360 * BX / (FL + YY)

R7 = délais fournisseurs
= 360 * DX / (FS + FU + FW + FZ)

ANNEXE 3

Tableau récapitulatif des huit techniques

	1 q1 - 1.5 eiq, q3 + 1.5 eiq	2 q1 - 3eiq, q3+3 eiq	3 BKW	4 STD sur BKW	5 STD sur moyenne et EIQ	6 STD sur moyenne et PSD	7 STD sur moyenne tronquée à 1 % et EIQ	8 STD sur moyenne tronquée à 1 % et PSD
Outils								
Moyenne				⊗	⊗	⊗		
Moyenne tronquée à 1 %							⊗	⊗
Écart-type			⊗	⊗				
Écart interquartile (EIQ)	⊗	⊗			⊗		⊗	
Pseudo écart-type (PSD)						⊗		⊗
Méthodes								
Box plot de Tukey	⊗	⊗						
Covratio de Belsley-Kuh-Welsh			⊗	⊗				
Standardisation de la distribution				⊗	⊗	⊗	⊗	⊗