MÉTHODES D'IMPUTATION DE VALEURS ABERRANTES POUR DES DONNÉES D'ENQUÊTES

Ruilin REN

Ecole Nationale de la Statistique et de l'Analyse de l'Information
Laboratoire de Statistique d'Enquête, Campus de Ker Lann

1. Introduction

Cet article a pour objet de présenter des méthodes d'imputation de valeurs aberrantes pour des données d'enquêtes. L'imputation des valeurs aberrantes fait partie importante du processus de data editing, surtout pour des données issues des enquêtes auprès des populations industrielles et économiques. Pourtant, ce sujet est très peu porté dans la recherche car les statisticiens d'enquête appliquent les méthodes d'imputation pour des données manquantes à l'imputation des valeurs aberrantes. Ceci est équivalent de traiter les valeurs aberrantes comme des valeurs manquantes. Néanmoins, une valeur manquante et une valeur aberrante sont statistiquement différentes. Une valeur manquante sur une variable est le résultat de n'avoir pas pu collecter l'information sur une unité enquêtée à cause d'une certaine raison. Donc nous n'avons pas d'information sur la caractéristique à enquêter pour cette unité. Une valeur manquante constitue un 'trou' dans le fichier de données ceci peut intriguer la défaillance de certain logiciel statistique qui n'accepte pas de valeur manquante. Tandis que une valeur aberrante est une observation ayant grand écart par rapport à sa valeur prévue. Elle est une vraie valeur et est correctement observée (nous distinguons une valeur aberrante d'une valeur erronée). Elle est aberrante à cause des erreurs dans la base de sondage ou des changements brusques de la population. Donc une valeur aberrante porte l'information sur l'unité enquêtée et n'introduit pas de problème pour l'utilisation de logiciel statistique. Une valeur aberrante peut mal conduire les analyses statistiques si elle n'est pas correctement traitée. Le processus du data editing devrais produire un fichier de données purifiées et complètes et prêt à utiliser par le grand public en utilisant des méthodes et des logicielles standard.

La recherche sur les traitements de valeurs aberrantes est portée exclusivement sur des méthodes d'estimation résistant aux valeurs aberrantes, voir *Ren et Chambers (2001a, 2001b)*. Ces méthodes sont souvent très sophistiques pour une utilisation par le grand public, donc les statisticiens d'enquête sont obligés de délivrer un fichier de données purifiées qui sont convenables à l'utilisation des méthodes standard. Cet objectif peut être atteindre par l'imputation pour donner à une valeur aberrante une valeur imputée, une valeur normale ou moins aberrante. Dans cet article, nous allons étudier des méthodes d'imputation classiques qui sont utilisées initialement pour l'imputation de valeurs manquantes, et nous allons modifier ces méthodes pour leur adapter à l'imputation de valeurs

aberrantes. Ils s'agissent de l'imputation par régression et l'imputation par le plus proche voisin. Les résultats présentés dans cet article sont encore préliminaires. Ils sont pour objectif d'intriguer la recherche sur ce sujet important. Nous présentons aussi une méthode étudiée par *Ren et Chambers (2002a, 2002b)*, il s'agit de l'imputation par le calage inverse. Les méthodes étudiées dans cet article sont évaluées par des données issues d'une enquête auprès des entreprises.

2. Valeur aberrante et estimation résistante

Supposons une population finie de taille N qui contient des valeurs aberrante représentative (*Ren et Chambers 2001a, 2001b*) sur une variable y. Par valeurs aberrantes représentatives, nous désignons des observations ayant grand écart par rapport à leur valeur prévue ou à leur moyenne, qui peuvent se présenter dans l'échantillon et également dans la partie de la population non observée. En représentant la population par une sous population normale U_0 et une sous population aberrante U_a, la population finie U s'écrit :

$$U = U_0 + U_a$$

Supposons un échantillon s tiré de la population par un plan de sondage quelconque et contenant des valeurs aberrantes qui peut représenter par un sous échantillon normal s_0 et un sous échantillon aberrant s_a :

$$s = s_0 + s_a$$

L'objectif principal de l'enquête est de produire un estimateur du total ou de la moyenne pour une variable y à partir de l'échantillon. L'estimateur du total le plus utilisé est l'estimateur de Horvitz-Thompson qui est sans biais pour le total mais qui est non résistant aux valeurs aberrantes :

$$\hat{t}_p = \sum_{k \in s} d_k y_k = \sum_{k \in s_0} d_k y_k + \sum_{k \in s_a} d_k y_k$$

où d_k est le poids de sondage associé à l'individu k, $k \in s$. Des poids normaux associés à des valeurs aberrantes peuvent torturer complètement l'estimateur du total. Par exemple, un petit nombre de valeurs aberrantes associées avec leur poids peut représenter un pourcentage important de la somme pondérée. Dans des cas extrêmes, une seule valeur extrêmement grande associée avec un grand poids peut représenter jusqu'à 70% de la somme pondérée, et peut résulter une sur estimation du total. Pour produire un estimateur résistant aux valeurs aberrantes, nous pouvons modifier soit les valeurs aberrantes ou leur poids associé.

Un estimateur résistant par la modification des valeurs aberrantes s'écrit :

$$\hat{t}_{mv} = \sum_{k \in s} d_k y_k^*$$

où $\left\{ y_k^*, k \in s \right\}$ est un ensemble d'observations modifiées. Notons que les modifications ne sont pas limitées aux valeurs aberrantes. Pour compenser le biais introduit par les modifications des valeurs aberrantes, des modifications peuvent atteindre aussi aux valeurs normales.

Un estimateur résistant par la modification des poids s'écrit :

$$\hat{t}_{mp} = \sum_{k \in s} d_k^* y_k$$

où $\{d_k^*, k \in s\}$ est un ensemble de poids modifiés. Comme dans le cas de modification de valeurs, les modifications de poids ne sont pas limitées aux poids associés aux valeurs aberrantes. Pour des détailles, voir *Ren et Chambers (2001a, 2001b).*

Les méthodes pour modifier les valeurs aberrantes pour obtenir des estimateurs résistants sont souvent sophistiquées et les modifications n'interviennent qu'à l'étape de l'estimation, en laissant le fichier de données sans toucher. Elles ne sont pas élaborées pour l'utilisation par le grand public. D'autre part, elles ne sont pas convenables pour créer un fichier de données purifiées car la modification de l'ensemble de données observées risque de détruire ou de torturer la distribution de données ou la corrélation avec d'autres variables. Pour délivrer un fichier nettoyé et prêt à l'emploi par le grand public, les modifications doivent être limitées aux valeurs aberrantes, et que les données modifiées permettent d'appliquer des méthodes classiques et de produire des résultats raisonnables.

3. Imputation par régression

Il est évident que les méthodes d'imputation classiques pour les valeurs manquantes s'appliquent à l'imputation des valeurs aberrantes en traitant ces dernières comme des valeurs manquantes. Nous commençons par l'imputation par la régression. Supposons une variable auxiliaire x (uni-variée ou multi-variée) liant la variable d'enquête y par un modèle linéaire :

$$y_k = b'x_k + e_k,\ k \in U$$

où $\{e_k\}$ sont les résidus de régression, $E(e_k|x_k) = 0$, $Var(e_k|x_k) = s^2 v^2(x_k)$, $v(x) > 0$ est une fonction connue ; b est le coefficient de régression inconnu. Soit s un échantillon contenant des valeurs aberrantes dans un sous échantillon s_a, $s = s_0 + s_a$. En traitant les valeurs aberrantes comme des valeurs manquantes, une estimation du b à partir de l'échantillon non aberrant s'écrit :

$$\hat{b}_{s_0} = \frac{\sum_{k \in s_0} d_k \frac{x_k y_k}{v^2(x_k)}}{\sum_{k \in s_0} d_k \frac{x_k' x_k}{v^2(x_k)}}$$

$\hat{b}_{s_0}$ est un estimateur résistant car il ne dépend pas de valeurs aberrantes. L'imputation classique par régression pour les valeurs aberrantes s'écrit alors :

$$y_k^* = \hat{b}_{s_0}' x_k, \quad k \in s_a$$

Cette imputation n'a pas fait référence de la valeur aberrante qui est une valeur correctement observée. En effet, la valeur imputée par régression y_k^* pour une valeur aberrante y_k est son expérience conditionnelle sous l'hypothèse que y_k soit une valeur normale :

$$y_k^* = \hat{b}_{s_0}' x_k = E\left(y_k | x_j, j \in s \text{ et } y_k \text{ est une valeur normale}\right),\ k \in s_a$$

Ceci est normal lorsque nous traitons les valeurs aberrantes comme des valeurs manquantes parce que une valeur manquante ne porte pas d'information sur elle-même, nous ne pouvons pas lui supposer une valeur différente de sa moyenne. Si nous regardons l'estimation du total après imputation qui est naturellement l'estimateur par régression :

$$\hat{t}_{lr} = \sum_{k \in s} d_k y_k^* + \hat{b}_{s_0}' \left(t_x - \sum_{k \in s} d_k x_k\right) = \sum_{k \in s_0} d_k y_k + \hat{b}_{s_0}' \left(t_x - \sum_{k \in s_0} d_k x_k\right)$$

où $t_x = \sum_{k \in U} x_k$ est le total connu à priori de la variable x. Ceci est équivalent de ne pas prendre en compte les valeurs aberrantes dans l'estimation du total. Lorsque les valeurs aberrantes sont en plupart des valeurs extrêmement grandes, cette estimation aura beaucoup sous estimé le total. Cette méthode d'imputation nécessite une adaptation pour que les valeurs aberrantes soient référées dans l'imputation et par conséquent prises en compte dans l'estimation du total.
Une adaptation simple est de rajouter un terme de correction dans l'imputation ci-dessus :

$$y_k^* = \hat{b}'_{s_0} x_k + d_k , \quad k \in s_a$$

où d_k est un terme de correction qui peut être déterminant ou aléatoire. Une correction déterminant consiste à augmenter y_k^* par une quantité positive $z_{1-a/2} \hat{s}_{s_0} v(x_k)$ lorsqu'il s'agit d'une valeur aberrante qui se trouve largement au dessus de la ligne de régression, et par une quantité négative $-z_{1-a/2} \hat{s}_{s_0} v(x_k)$ lorsqu'il s'agit d'une valeur aberrante qui se trouve largement au dessous de la ligne de régression, voir la figure 3.1. Nous pouvons représenter la quantité par une formule compacte :

$$d_k = Sign\left(y_k - \hat{b}'_{s_0} x_k\right) z_{1-a/2} \hat{s}_{s_0} v(x_k) , \quad k \in s_a$$

où $z_{1-a/2}$ est la valeur critique d'une variable $N(0, 1)$; $\hat{s}_{s_0}$ est la partie inconnue de l'écart type du résidu estimée sur échantillon non aberrant :

$$\hat{s}_{s_0}^2 = \frac{\sum_{k \in s_0} d_k (e_k - \bar{e})^2}{\sum_{k \in s_0} d_k},$$

avec $\bar{e} = \frac{\sum_{k \in s_0} d_k e_k}{\sum_{k \in s_0} d_k}$, $e_k = \frac{y_k - \hat{b}'_{s_0} x_k}{v(x_k)}$, $k \in s_0$. Cette modification a un sens intuitif vu dans la figure 3.1 ci-dessous : au lieu de ramener une valeur aberrante sur la ligne de régression, on le ramène au bord de la région confidentielle de régression. C'est-à-dire, nous modifions au minimum une valeur aberrante pour qu'il soit une valeur acceptable.

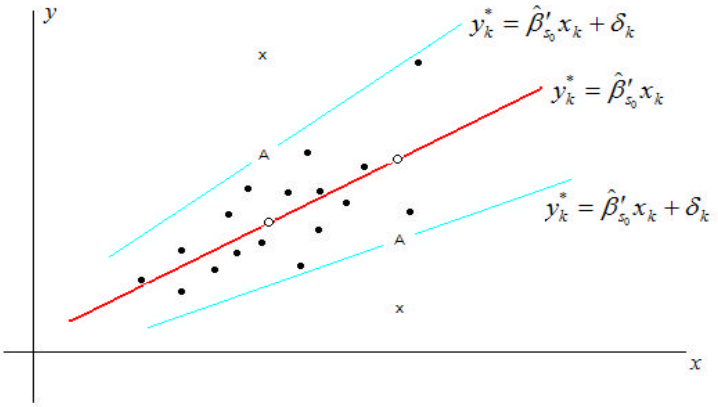


Figure 3.1 Imputation par régression modifiée
Légendes : x - valeurs aberrantes ; o - valeurs imputées par régression ; A - valeurs imputées par régression modifiée

Une correction aléatoire consiste à augmenter y_k^* par une quantité positive et aléatoire $|z_k| \hat{s}_{s_0} v(x_k)$ lorsqu'il s'agit d'une valeur aberrante qui se trouve largement au dessus de la ligne de régression, et par une quantité négative et aléatoire $-|z_k| \hat{s}_{s_0} v(x_k)$ lorsqu'il s'agit d'une valeur aberrante qui se trouve largement au dessous de la ligne de régression. C'est-à-dire par une quantité :

$$d_k = Sign\left(y_k - \hat{b}'_{s_0} x_k\right)\left|z_k\right|\hat{s}_{s_0} v(x_k), \quad k \in s_a$$

où $\{z_k, k \in s_a\}$ est un échantillon *iid* tiré dans une loi $N(0,1)$. Donc il est facile à calculer la valeur moyenne de $|z_k|$, qui vaut $E(|z_k|) = 2/\sqrt{2p}$. De ce fait, la correction aléatoire modifie moins l'imputation par régression par rapport à la correction déterminante, par exemple, $z_{1-a/2} = 1{,}96$ lorsque $a = 0{,}05$.

Si nous regardons l'estimateur du total après imputation par l'estimateur par régression, nous avons :

$$\begin{aligned}\hat{t}^*_{lr} &= \sum_{k\in s} d_k y^*_k + \hat{b}'_{s_0}\left(t_x - \sum_{k\in s} d_k x_k\right) \\ &= \sum_{k\in s_0} d_k y_k + \hat{b}'_{s_0}\left(t_x - \sum_{k\in s_0} d_k x_k\right) + \sum_{k\in s_a} d_k d_k\end{aligned}$$

Le terme supplémentaire $\sum_{k\in s_a} d_k d_k$ correspond à une correction pour l'imputation par régression classique. Lorsque la plupart des valeurs aberrantes sont des valeurs extrêmement grandes, ce terme de correction est positif, qui fait une compensation de biais introduit par la correction ou l'imputation de valeurs aberrantes. Voir les résultats numériques dans le tableau 6.3 où nous avons observé des estimations ayant le terme de correction positif.

4. Imputation par le plus proche voisin

Pour une valeur aberrante donnée y_{k_0}, supposons connue une variable auxiliaire x, l'imputation par le plus proche voisin au sens classique, c'est-à-dire, traiter une valeur aberrante comme une valeur manquante, consiste à chercher parmi les non aberrantes une unité k' telle qu'elle minimise une certaine distance entre l'unité k_0 et l'unité k' :

$$k' = \text{Arg}\underset{k\in s_0}{Min}\left\{d\left(x_k, x_{k_0}\right)\right\}$$

où d est une mesure de distance, par exemple, la mesure usuelle $d\left(x_k, x_{k_0}\right) = \left|x_k - x_{k_0}\right|$. La valeur imputée pour y_{k_0} est $y^*_{k_0} = y_{k'}$. Comme dans le cas de l'imputation par régression classique, la valeur aberrante elle-même n'est pas prise en compte dans la recherche de son plus proche voisin. Mais le plus proche voisin est peut-être préférable à la régression classique parce qu'il est ressemble l'imputation par régression modifiée lorsque la taille de l'échantillon est importante et que les observations sur la variable x sont denses. Dans ce cas, nous aurons $x_{k_0} \cong x_{k'}$ et par conséquent :

$$y^*_{k_0} = y_{k'} = b' x_{k'} + e_{k'} \cong \hat{b}'_{s_0} x_{k_0} + e_{k'}$$

où $e_{k'}$ est la partie correspondante du d_k dans l'imputation par régression modifiée. Cette explication est prouvée par les résultats numériques dans le tableau 6.3, où l'estimation du total par la régression modifiée est très proche de celle par le plus proche voisin.

Pourtant, la valeur aberrante n'est pas prise en compte dans la recherche de son plus proche voisin. Une simple modification pour que la valeur aberrante soit prise en compte dans la recherche de voisin consiste à utiliser une mesure de distance $d\left[\left(x_k, y_k\right), \left(x_{k_0}, y_{k_0}\right)\right]$ qui mesure la distance entre les doublons $\left(x_k, y_k\right)$ et $\left(x_{k_0}, y_{k_0}\right)$. Le plus proche voisin de l'unité k_0 est celui qui minimise la distance :

$$k' = \text{Arg}\underset{k\in s_0}{Min}\left\{d\left[\left(x_k, y_k\right), \left(x_{k_0}, y_{k_0}\right)\right]\right\}$$

La valeur imputée pour y_{k_0} est $y^*_{k_0} = y_{k'}$, où k' est le plus proche voisin de k_0.

Une mesure de distance usuelle est :

$$d\left[(x_k, y_k),(x_{k_0}, y_{k_0})\right] = \left[(y_k - y_{k_0})^2 + (x_k - x_{k_0})' (x_k - x_{k_0}) \right]^{\frac{1}{2}}$$

En référant la figure 4.1 et le tableau 6.4 on constate que les valeurs imputées par le plus proche voisin modifié sont plus proches de leur valeur originale et ont une forte corrélation avec ces dernières. Le coefficient de corrélation entre les valeurs imputées et leur valeur originale est un critère d'évaluation de l'imputation. Nous voulons que les valeurs imputées respectent au maximum la réalité.

Comme le plus proche voisin classique, le plus proche voisin modifié subit de même inconvénient, nous pouvons avoir des valeurs imputées non valides, c'est-à-dire des valeurs imputées qui ne passent pas de l'editing. Dans ce cas, il faut faire une seconde fois de l'imputation pour les valeurs imputées non valides, ou même une troisième fois jusqu'à toutes les valeurs imputées sont validées. Cet inconvénient peut être amélioré par une modification de la mesure de distance. Nous pouvons utiliser une mesure de distance pondérée, *le plus proche voisin pondéré* :

$$d\left[(x_k, y_k),(x_{k_0}, y_{k_0})\right] = \left[a\,(y_k - y_{k_0})^2 + (1-a)(x_k - x_{k_0})' (x_k - x_{k_0}) \right]^{\frac{1}{2}}$$

où $0 \le a < 1$ est une pondération choisie par le statisticien qui représente le niveau de l'importance que le statisticien met sur la valeur aberrante, qui peut être une pondération unique pour toutes les valeurs aberrantes ou une pondération spécifique pour chaque valeur aberrante. Par exemple, dans le dernier cas, nous pouvons choisir $a_{k_0} = d_{k_0} / (d_{k_0} + d_{k'})$. Nous retrouvons le plus proche voisin au sens classique lorsque $a = 0$. Nous retrouvons le plus proche voisin modifié lorsque $a = 0{,}5$. La valeur imputée pour y_{k_0} est toujours $y^*_{k_0} = y_{k'}$.

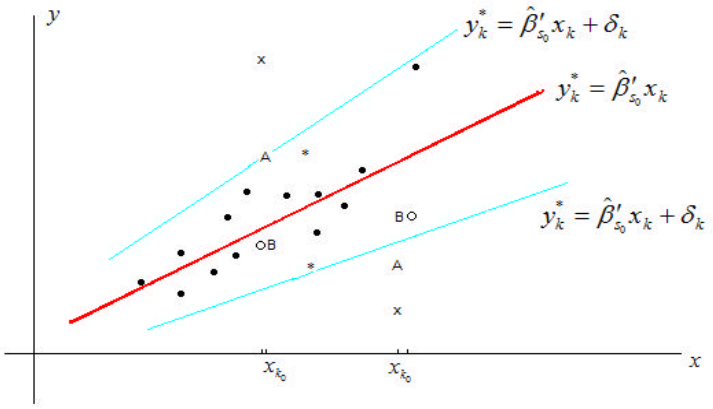


Figure 4.1 Imputation par le plus proche voisin
Légendes : x - valeurs aberrantes ; o - le plus proche voisin ; * - le plus proche voisin modifié
B - valeurs imputées par le plus proche voisin ; A - valeurs imputées par le plus proche voisin modifié

En effet, la recherche du plus proche voisin n'est pas obligée d'être limitée parmi les non aberrantes car une valeur y_{k_0} est aberrante associée avec x_{k_0} peut être devenue non aberrante associée avec $x_{k'}$

même si $y_{k'}$ est aberrante, et peut être encore aberrante même si $y_{k'}$ est non aberrante. Mais une chose qui apparaît sûr c'est que le processus de l'imputation converge plus vite lorsque nous limitons la recherche du plus proche voisin parmi les non aberrants. Un exemple le plus défavorable pour la recherche du plus proche voisin parmi toutes les observations est que deux valeurs aberrantes sont les plus proches voisins l'une pour l'autre ainsi que le processus de l'imputation ne converge pas.

Remarque : Il est évident que le plus proche voisin d'une valeur aberrante n'est pas nécessairement unique. Lorsque une valeur aberrante possède plusieurs voisins également proches, il faut choisir un seul pour procéder l'imputation. Le principe est de choisir celui qui produit une valeur imputée valide. Dans le cas où toutes les valeurs imputées sont valides ou aucune valeur imputée n'est valide, le choix entre eux n'a plus d'importance. Nous pouvons choisir un par hasard, par exemple.

5. Imputation par le calage inverse

Ren et Chambers (2002a) proposent une méthode d'imputation basée sur un estimateur du total $\hat{t}_y$ obtenu par une méthode résistante. L'objectif de l'imputation est de modifier ou imputer les valeurs aberrantes y_k, $k \in s_a$, par des valeurs y_k^*, $k \in s_a$ normales ou moins aberrantes telles que :

$$\hat{t}_y^*\left(y_k^* \middle| k \in s\right) = \hat{t}_y$$

où $\hat{t}_y^*$ est un estimateur classique. Donc l'idée est de donner des valeurs imputées pour les valeurs aberrantes telles que l'estimation du total obtenue par une méthode résistante peut être retrouvée par une méthode classique. Par exemple, lorsque $\hat{t}_y^*$ est un estimateur pondéré :

$$\hat{t}_y^* = \sum_{k \in s} w_k y_k^* = \sum_{k \in s_0} w_k y_k + \sum_{k \in s_a} w_k y_k^* = \hat{t}_y$$

où $\{w_k, k \in s\}$ est un ensemble de poids. Cela implique que la contribution à l'estimation du total par les valeurs non aberrantes, notée $\hat{t}_{1y}$, et celle par les valeurs aberrantes, notée $\hat{t}_{2y}$, seront :

$$\hat{t}_{1y} = \sum_{i \in s_0} w_i y_i \,, \quad \hat{t}_{2y} = \sum_{i \in s_a} w_i y_i^*$$

Notons que $\hat{t}_{1y}$ est connu, donc $\hat{t}_{2y}$ peut être calculé par :

$$\hat{t}_{2y} = \hat{t}_y - \hat{t}_{1y}, \quad (\text{supposons } \hat{t}_{2y} > 0)$$

L'objectif de l'imputation est donc d'imputer des valeurs y_k^*, $k \in s_a$ telles que :

$$\sum_{k \in s_a} w_k y_k^* = \hat{t}_{2y}$$

Comme les valeurs aberrantes sont des valeurs vraies, nous ne voulons pas que une valeur imputée soit trop éloignée de sa valeur vraie. C'est typiquement un problème de calage sur marge. Les résultats dans *Deville et Sarndäl (1992), Deville, Sarndäl et Sautory (1993)* nous donnent immédiatement les valeurs imputées :

$$y_k^* = y_k F_k(w_k l), \; k \in s_a$$

où F_k est la fonction de calage, $F_k(0) = 1$ et $F_k'(0) = q_k$; l est une constante à déterminer par :

$$\sum_{k \in s_a} w_k y_k F_k (w_k l) = \hat{t}_{2y} .$$

Par exemple, lorsque la mesure de distance est donnée par :

$$r\left(y^*, y\right) = \sum\nolimits_{k \in s_a} \left(y_k^* - y_k\right)^2 / 2 q_k y_k$$

où $q_k > 0, k \in s_a$ sont des poids de calage choisis par statisticien (elle correspond à la méthode linéaire dans *Deville et Särndal (1992)*). Le calage donne les valeurs imputées :

$$y_k^* = y_k + q_k w_k y_k \left(\sum_{k \in s_a} q_k w_k^2 y_k \right)^{-1} \left(\hat{t}_{2y} - \sum_{k \in s_a} w_k y_k \right), \; k \in s_a$$

Lorsque les poids de calage sont choisis $q_k = w_k^{-1}, k \in s_a$, nous avons :

$$y_k^* = y_k \frac{\hat{t}_{2y}}{\sum_{k \in s_a} w_k y_k}, \; k \in s_a$$

Dans la construction de l'imputation par le calage inverse, la variable y joue le rôle de la variable de poids, la variable de poids w joue le rôle d'une variable auxiliaire, d'où vient le nom 'calage inverse'. Les avantages principaux de cette méthode sont que l'estimation du total résistant aux valeurs aberrantes peut être retrouvée par un estimateur classique en utilisant les données imputées ainsi que la variance de ce dernier est identique à celle de l'estimateur résistant ; le programme *CALMAR* (*Sautory, 1993*) peut être utilisé pour accomplir le calage. Il y a un inconvénient aussi : les valeurs imputées ne sont pas garanties d'être valides. Une solution pour surmonter cet inconvénient consiste à réaliser un calage conditionnel. Lorsqu'il existe deux types de valeur aberrante, valeur aberrante extrêmement grande et valeur aberrante extrêmement petite, il faut les imputer séparément. Il consiste à déterminer d'abord les contributions à l'estimation du total de chaque type de valeur aberrante, puis procéder l'imputation par calage inverse indépendamment.

Remarque : Dans l'imputation par le calage inverse, le calcul de la contribution à l'estimation du total de chaque type de valeur aberrante joue un rôle essentiel. Nous supposons que ces calculs sont fiables. La recherche des méthodes d'estimations résistantes ne fait pas objet de cet article. Pour ceux que cela intéresse, nous leur proposons de consulter l'article de *Ren et Chambers (2001b).*

6. Validations numériques

Dans cette section, nous utilisons des données issues d'une enquête auprès des entreprises sur un secteur spécifique contenant des valeurs aberrantes pour tester les méthodes proposées dans cet article. Nous considérerons deux variables d'intérêt : le chiffre d'affaires (*turnover*) et le total d'achat (*purtot*). Nous disposons également une variables auxiliaire qui est le chiffre d'affaires enregistré (*turnreg*) dont le total est connu à priori : $t_x = 211732739$. Dans ce fichier de donnée, nous avons 6099 observations. Les valeurs aberrantes sont identifiées par une procédure étudiée par *Hentges (2001)*. Les estimations résistantes de totaux sont produites par l'estimateur résistant basé sur un modèle linéaire (*Chambers, 1986*). Dans le tableau 6.1 nous présentons les nombres de valeurs aberrantes pour chaque variable, les estimations des totaux obtenues par l'estimation par régression non résistante et les estimations résistantes de *Chambers (1986).*

Nous présentons dans la figure 6.1 des représentations graphiques des chiffres d'affaires par rapport aux chiffres d'affaires enregistrés en échelle de logarithme. Les graphiques montrent que nous avons une variable auxiliaire qui explique très bien la variable d'enquête en échelle de logarithme. En fait, la variable auxiliaire explique très bien la variable d'enquête en échelle originale aussi, mais nous ne présentons que les graphiques en échelle de logarithme pour simplifier les présentations. Les graphiques montrent que parmi les valeurs imputées il y a quelques observations qui sont encore 'aberrantes' pour les imputations par le plus proche voisin, le plus proche voisin modifié et par le calage inverse. En effet, ils agissent de sept valeurs les plus aberrantes. Leur valeur originale sont beaucoup modifiées par l'imputation, comme montrées dans le tableau 6.2. pour le plus proche voisin modifié.

Dans la figure 6.2, nous présentons des graphiques de valeurs imputées par rapport à leur valeur originale pour les valeurs aberrantes en échelle de logarithme pour le chiffre d'affaires. Les graphiques et les résultats dans le tableau 6.4 montrent que les valeurs imputées sont fortement corrélées à leur valeur originale pour l'imputation par le calage inverse et l'imputation par le plus proche voisin.

Dans le tableau 6.3, nous présentons les résultats de l'estimation du total avant et après imputation, par l'estimateur basé sur un modèle linéaire pour l'estimateur résistant avant imputation, et par l'estimateur par régression classique après imputation. Les résultats montrent que les estimations par régression après imputation pour les méthodes de calage inverse et de régression sont proches, et elles sont proches de l'estimation résistante avant imputation. Les résultats sont proches pour la méthode de régression modifiée et la méthode de plus proche voisin. Les résultats obtenus par le plus proche voisin modifié montrent les plus faibles différences par rapport aux estimations non résistantes avant imputation. Nous espérons de plus recherche sur cette méthode.

En conclusion, les résultats numériques et les autres recherches menées par l'auteur de cet article montrent que toutes les méthodes étudiées dans cet article marchent bien pour l'imputation de valeurs aberrantes pour des données d'enquêtes. Elles consistent au moins des méthodes compétitives par rapport aux méthodes initialement utilisées pour l'imputation de valeurs manquantes.

Tableau 6.1. Nombre de valeurs aberrantes et estimation du total par l'estimateur basé sur un modèle linéaire

	Nombre de valeurs Aberrantes	*Estimation non-résistante du total*	*Estimation résistante du total*
turnover	106	269545407	252938704
purtot	111	192575028	180732418

Figure 6.1. Chiffre d'affaires et chiffre d'affaires imputé par rapport au chiffre d'affaires enregistré en échelle de logarithme

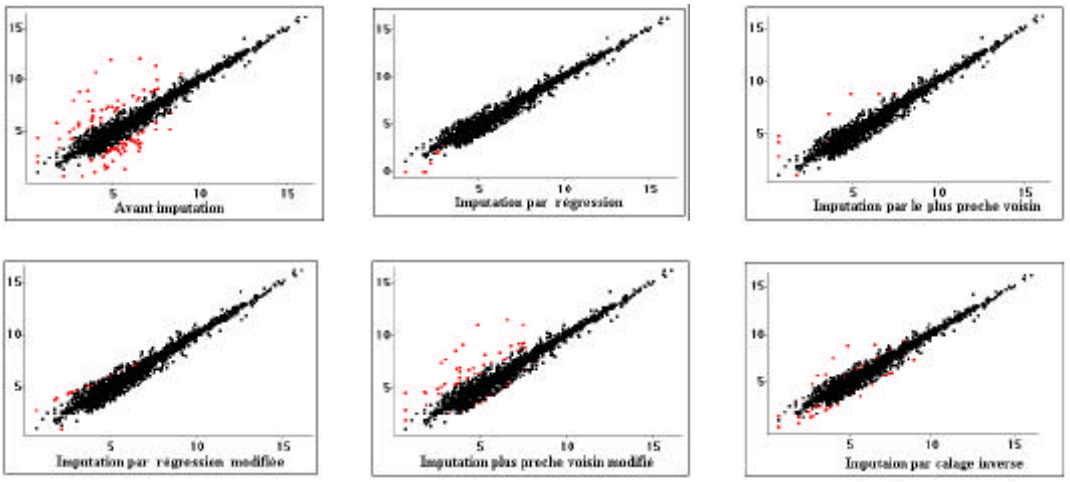
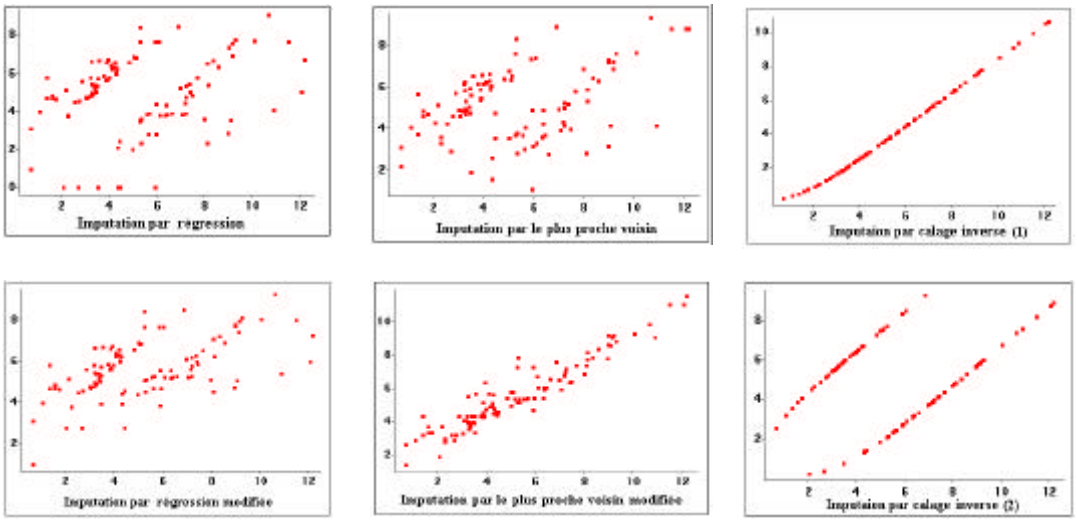

Avant imputation
Imputation par régression
Imputation par le plus proche voisin
Imputation par régression modifiée
Imputation plus proche voisin modifié
Imputaion par calage inverse

Figure 6.2. Chiffre d'affaires imputé par rapport au vrai
chiffre d'affaires en échelle de logarithme



(1) imputation des grandes et des petites valeurs aberrantes ensembles ; (2) imputation séparément

Tableau 6.2. Les sept plus grosses valeurs aberrantes pour le chiffre d'affaires
avant imputation et après imputation par le plus proche voisin modifié

Valeur originale	209683	186399	103956	56000	9084	8176	3380
Valeur imputée	118879	67231	67231	9738	5683	2544	1699

Tableau 6.3. Estimation du total avant et après imputation par estimateur basé sur un modèle linéaire

	Estimation Résistante	*Estimation Classique Après Imputation*				
	Avant Imputation	*Calage Inverse*	*Régression*	*Proche voisin*	*Régression modifiée*	*Proche voisin modifié*
turnover	252938707	252808484	252225060	253479240	253005961	259245185
purtot	180732418	180670463	180483772	181352764	181098312	185556428

Tableau 6.4. Valeur moyenne des valeurs aberrantes, de leur imputation et leur coefficient de corrélation

	Valeur originale	*Calage Inverse*	*Régression*	*Proche voisin*	*Régression modifiée*	*Proche voisin modifié*
turnover	1456	325 (1,00)* 325 (0,42)	214 (0,07)	217 (0,33)	273 (0,14)	547 (0,87)
purtot	763	182 (1,00) 182 (0,08)	173 (0,03)	156 (0,09)	209 (0,05)	320 (0,89)

* Les chiffes dans les parenthèses sont les coefficients de corrélation entre les vraies valeurs aberrantes et leur imputation. Pour l'imputation par calage inverse, la première ligne correspond de l'imputation ensemble des grandes et des petites valeurs aberrantes. La seconde ligne correspond de l'imputation séparément de deux types

de valeurs aberrantes. Pourtant, le coefficient de corrélation est sous évalué dans ce dernier cas, où il faut calculer le coefficient de corrélation pour les deux types de valeurs aberrantes séparément.

Bibliographie

-CHAMBERS, R. L. : Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069 (1986)

-HENTGES, A. L. : Robust multivariate outlier detection methods. Euredit project report. (2001).

-REN, R. et CHAMBERS, R. L. : Unbiased outlier resistant estimation for finite populations. Euredit Project Report . (2001a)

-REN, R. et CHAMBERS, R. L. : Studies on outlier robust estimators. Euredit Project Report . (2001b)

-REN, R. et CHAMBERS, R. L. : Outlier imputation by reverse calibration. Euredit Project Report (2002a)

-REN, R. et CHAMBERS, R. L. : Méthodes d'estimation résistant aux valeurs aberrantes et méthodes d'imputation pour des données d'enquêtes. *Troisième Journées de Sondages*, Autrans, Grenoble. (2002b) (à apparaître dans *Enquêtes et Sondages*)

-DEVILLE, J.C. et SÄRNDAL, C. E. : Calibration estimators in survey sampling, Journal of the American Statistical Association, Vol. 87, p 376-382 (1992)

-DEVILLE, J. C. SÄRNDAL, C. E. et SAUTORY, O. : Generalized raking procedures in survey sampling, Journal of the American Statistical Association, Vol. 88, pp 1013-1020 (1993)

-SAUTORY, O. : La macro CALMAR: Redressement d'un échantillon par calage sur marges. Document de travail série méthodologie, F9310, Insee (1993).