REVUE DE STATISTIQUE APPLIQUÉE

F. CRETTAZ DE ROTEN
J.-M. HELBLING

Données manquantes et aberrantes : le quotidien du statisticien analyste de données

Revue de statistique appliquée, tome 44, n° 2 (1996), p. 105-115
<http://www.numdam.org/item?id=RSA_1996__44_2_105_0>

Rev. Statistique Appliquée, 1996, **XLIV** (2), 105-115

DONNÉES MANQUANTES ET ABERRANTES : LE QUOTIDIEN DU STATISTICIEN ANALYSTE DE DONNÉES

F. Crettaz de Roten (1), J.-M. Helbling (2)

(1) Université de Lausanne
IMA-SSP, BFSH2, 1015 Lausanne (Suisse)
(2) Ecole Polytechnique Fédérale de Lausanne
Département de Mathématiques
1015 Lausanne-Ecublens (Suisse)

1. Introduction

Un statisticien, analyste de données, est confronté régulièrement à des données que l'on peut qualifier d'imparfaites ou de «messy data» comme le monde anglophone les appelle. Il s'agit soit de matrices de données incomplètes c'est-à-dire contenant des données manquantes soit de matrices avec des points douteux voire aberrants ou encore de matrices dont les observations ne respectent pas les hypothèses nécessaires à l'application de la technique souhaitée, par exemple, ne suivent pas une loi normale. Dans cet article, nous allons traiter le point de vue analyse de données des deux premiers problèmes simultanément même si, à première vue, il n'y a rien de plus différent qu'une donnée manquante et une donnée aberrante.

Séparément les deux problèmes ont retenu l'attention de nombreux auteurs ces dernières années et la littérature est abondante sur chacun des sujets traités tantôt de façon générale, tantôt dans un contexte particulier comme les séries temporelles ou les plans d'expérience. Pour une vision détaillée de ces sujets, on peut se référer aux livres de Barnett et Lewis (1994) pour les données aberrantes et de Little et Rubin (1987) pour les données manquantes.

Les deux problèmes sont traités très rarement ensemble. Il n'y a que le contexte des sondages qui intègre ces deux problèmes dans la procédure d'édition des données en définissant des règles d'imputation pour les données manquantes et de consistence pour la détection de données aberrantes (Little and Smith, 1987; Giles, 1988). Notre intérêt pour un traitement simultané de ces deux sujets peut s'expliquer de plusieurs façons. Premièrement il est bien connu qu'une méthode de détection d'aberrance consiste à ôter la donnée soupçonnée, c'est-à-dire à créer une donnée manquante, et à comparer les estimations des paramètres obtenues avec les estimations basées sur l'ensemble des données. Deuxièmement, en étudiant et en développant des méthodes d'imputation de données manquantes (Crettaz de Roten, 1993), on a pu constater leur extrême sensibilité à la présence de données aberrantes. Finalement, en recherchant de bons jeux de données multivariées, nous avons remarqué que souvent

les deux problèmes sont imbriqués l'un dans l'autre et qu'il est donc nécessaire d'être conséquent dans le traitement de chacune des difficultés. Nous pensons que les deux problèmes doivent donc être traités par des techniques cohérentes. Les développements théoriques chercheront à mettre en évidence les parallélismes entre les techniques utilisées dans ces deux directions.

Dans cet article, nous allons premièrement rapprocher les deux domaines dans trois contextes à savoir les concepts généraux, la modélisation paramétrique et l'approche heuristique, puis sur la base d'un exemple, nous tenterons d'esquisser des règles pratiques.

2. Concepts généraux

Une difficulté survenant dans les deux domaines est la définition même de donnée manquante ou de donnée aberrante. D'une part, Rubin (1976) a introduit, à l'aide d'un modèle probabiliste, trois types de données manquantes selon les conditions «d'aléatoirité» du processus à l'origine de l'absence de la donnée. D'autre part, une donnée aberrante est assez facilement définissable dans un cadre unidimensionnel puisqu'il s'agit d'un point situé loin dans les queues de la distribution mais il devient pratiquement impossible de donner une définition générale dans une situation multidimensionnelle.

Le traitement des données manquantes ou aberrantes nécessite une approche quasi semblable :

(1) Analyser les données pour détecter les données aberrantes ou le type de données manquantes.

(2) Choisir si l'on va nettoyer les données, par élimination ou par imputation, avant de les traiter par des méthodes standards ou si l'on veut modifier la méthode d'analyse statistique pour tenir compte des impuretés (par l'introduction de poids ou par l'accommodation de l'estimation à la présence de données manquantes).

Quelle que soit l'approche choisie, il nous semble essentiel de garder en mémoire le fait que les données contenaient des non-réponses ou des aberrances. Une solution consiste à conserver le schéma de la répartition des données manquantes et les poids mesurant l'influence de chaque point pour en tenir compte dans l'interprétation des résultats des analyses ultérieures.

Le problème de l'existence de valeurs aberrantes porte en lui le germe des méthodes robustes. Il est donc naturel d'utiliser de telles techniques pour le traitement des aberrances (estimation robuste de la matrice de variance-covariance, régression L1) et plus spécialement pour leur détection (la fiabilité de la détection d'une donnée aberrante par une technique robuste est meilleure) (Rousseeuw et Leroy, 1987). De même, puisque la plupart des méthodes de traitement des données manquantes sont sensibles à la présence de données extrêmes, l'utilisation de variantes robustes a fait ses preuves (imputation à l'aide de modélisation robuste, voir paragraphe 3).

Une autre difficulté commune aux deux problèmes est le passage du traitement d'une seule donnée imparfaite à un groupe de données imparfaites. En présence de

plusieurs données manquantes doit-on utiliser une procédure d'imputation séquentielle ou globale? Dans le cas des données aberrantes, doit-on détecter les aberrances consécutivement ou par groupes? Dans la situation de la figure 1, la plupart des procédures consécutives auront un effet de masque, c'est-à-dire qu'elles ne découvriront pas de points aberrants alors que le groupe de points $\{A, B, C, D\}$ est aberrant. Dans ce cas, l'application de méthodes robustes permet d'améliorer la détection.

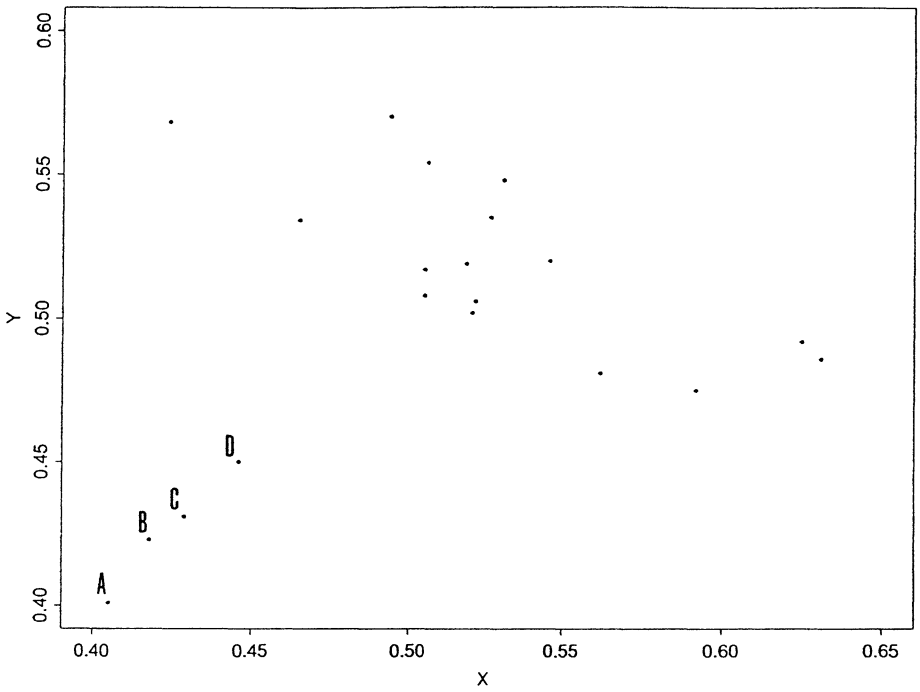

FIGURE 1

D'autre part, l'imputation de plusieurs données manquantes ou la suppression de plusieurs données douteuses a pour conséquence de diminuer la variabilité. Pour y remédier l'imputation multiple propose de remplacer chaque donnée manquante par un ensemble de valeurs générant m matrices de données complétées que l'on analyse chacune par des méthodes standards (Little et Rubin, 1987). Un avantage de cette technique est de pouvoir mesurer la variabilité due à l'imputation, mais l'accroissement des calculs qui en découle limite son utilisation.

Dans le cas des données aberrantes, l'application de méthodes robustes au lieu de la suppression aura un effet comparable. En effet, diminuer la variabilité initiale est souhaitable en présence de données aberrantes car ces dernières créent une variabilité artificielle, mais la suppression est une solution trop radicale pour les partisans des techniques robustes.

3. Modélisation paramétrique

La modélisation des données, le plus souvent par un modèle de distribution normale et/ou par un modèle de régression linéaire est fréquemment utilisée. Dans le cas des données manquantes l'algorithme EM permet d'obtenir une estimation des paramètres et une imputation des données manquantes : l'étape E estime les données manquantes et l'étape M maximise la vraisemblance en vue d'estimer les paramètres (Dempster, Laird et Rubin, 1977). De plus, les écarts importants au modèle peuvent révéler la présence de données aberrantes. L'approche robuste de cette modélisation, proposée par Little (1988), résoud simultanément les deux problèmes en introduisant des poids dans la résolution de l'algorithme EM. Puisque le modèle multinormal du traitement des données manquantes est sensible à la présence d'aberrances, la version robuste de l'algorithme EM est certainement préférable. La modélisation sous la forme de lois normales contaminées fournit un moyen de détection plus sophistiqué.

L'imputation de données manquantes est souvent obtenue par l'application d'un modèle de régression soit en utilisant directement l'espérance conditionnelle estimée (méthode de Buck, 1960) soit en y ajoutant un résidu aléatoire. S'il manque la première composante du dernier individu, cette méthode imputera par :

$$E(X_{1n} \mid X_{2n}, ..., X_{pn}) = \hat{\beta}_1 + \hat{\beta}_2 X_{2n} + ... + \hat{\beta}_p X_{pn}$$

où les $\hat{\beta}_i$ sont les estimateurs des moindres carrés évalués sur les individus sans données manquantes. Si les données suivent une loi normale multivariée, la méthode de Buck et l'algorithme EM donnent la même solution (Little et Rubin, 1987).

En présence de plusieurs prédicteurs possibles, il s'agira de définir le meilleur sous-ensemble pour chaque variable qui possède des données manquantes (Frane, 1977). Certains auteurs utilisent des versions robustes de ces méthodes.

La régression est certainement la technique statistique qui dispose de plus de moyens pour détecter des données aberrantes (Hadi et Simonoff, 1993). Il est par exemple bien connu que l'examen de différents graphiques relatifs aux résidus, met en évidence la présence d'aberrances. De plus, de nombreux auteurs ont proposé des mesures du degré d'influence des points basées sur le modèle de régression. Une procédure fréquemment utilisée est celle qui consiste à calculer les estimateurs des moindres carrés avec et sans la donnée mise en doute puis à évaluer la différence entre les valeurs obtenues.

Régulièrement, de nouvelles méthodes basées sur les développements récents de la statistique sont introduites. Par exemple, les techniques basées sur le rééchantillonnage ont donné lieu à l'utilisation du bootstrap et du jackknife pour l'imputation de données manquantes et la détection d'aberrances. Dans ces cas, la modélisation se fait en se basant sur la fonction de répartition empirique. Cette approche peut être intégrée dans un contexte Bayesien (par exemple dans l'imputation par le bootstrap Bayesien approximé (Little et Rubin, 1987). Le bootstrap offre aussi une approche non-paramétrique pour vérifier la qualité d'un estimateur en présence de données manquantes (Efron, 1993).

La modélisation paramétrique est évidemment un outil statistique puissant mais dans notre contexte, elle est très rigide et accentue l'adéquation des données au

modèle. Par exemple dans une situation de régression, une imputation par la valeur prédite améliore la qualité de l'ajustement sur les données finales, de même que la suppression d'une donnée aberrante à cause d'un résidu trop important.

4. Approche heuristique

De nombreuses techniques ont été développées par des praticiens sous forme de règles heuristiques; l'expérience et certaines simulations ont montré que ces méthodes donnent d'assez bons résultats. On peut citer pour les données manquantes l'imputation par la valeur correspondante d'un individu ayant des caractéristiques semblables, l'imputation par la moyenne ou la médiane et pour les aberrances, l'examen du «boxplot» (données extérieures aux moustaches) ainsi que d'autres représentations graphiques (glyphes, étoiles, faces de Chernoff, diagrammes de Andrews, arbres de classification hiérarchique, Q-Qplot de composantes principales, résidus de régression, etc).

Dans les deux domaines, on utilise abondamment la notion de distances en procédant à une minimisation pour imputer la valeur manquante et en recherchant un maximum pour détecter une valeur aberrante. Pour l'imputation, les principales distances sont la distance euclidienne, la distance de Mahalanobis et différentes distances entre individus. La famille des méthodes de type «hot deck» (utilisation fréquente dans l'édition de questionnaires) se base précisément sur des distances entre individus (Ford, 1983) : dans ce cas, on impute par la valeur de l'un des individus dont la distance à celui qui a une donnée manquante est minimale. Pour la détection, les distances les plus utilisées sont la distance de Mahalanobis, celle de Cook et Weisberg et différentes distances de la forme $(x_i - \overline{x})^t S^b (x_i - \overline{x})$ où $\overline{x}$ et S représentent respectivement le vecteur moyenne empirique et la matrice de variance-covariance estimée et où b peut prendre des valeurs entières positives ou négatives (Beckman et Cook, 1983). Mathématiquement le parallélisme entre les deux domaines se manifeste ainsi :

- supposons que seul le $i^{\text{ème}}$ individu possède des données manquantes; on peut les imputer en minimisant

$$(x_i - \overline{x}_{-i})^t S_{-i}^{-1} (x_i - \overline{x}_{-i})$$

où l'indice $-i$ signifie que l'estimateur est évalué sans la $i^{\text{ème}}$ observation (cette procédure est équivalente à la méthode de Buck);

- supposons que l'on soupçonne l'existence d'une donnée aberrante; on peut la détecter en déterminant l'observation j qui maximise la distance de Mahalanobis

$$(x_i - \overline{x})^t S^{-1} (x_i - \overline{x}).$$

Ce parallélisme dans l'optimisation existe aussi lorsqu'on utilise une technique de modélisation comme la régression. En effet dans ce cas, on impute en minimisant la valeur absolue du résidu et on détecte la donnée douteuse en déterminant le maximum de la valeur absolue des résidus.

En résumé, on peut dire que les approches heuristiques sont certainement le plus fréquemment utilisées.

5. Exemple pratique

Dans les 4 premiers paragraphes, nous avons cherché à mettre en évidence le parallélisme entre ces deux domaines de la statistique multivariée. Nous désirons maintenant faire ressortir des aspects plus pratiques en appliquant certaines des techniques énumérées précédemment sur un exemple.

Dans un article paru en 1988, Daudin, Duby et Trécourt reportent les résultats de mesures faites sur 86 échantillons de lait. Il s'agit des 8 variables suivantes :

$$\vec{X} = \begin{pmatrix} \text{densité} \\ \text{matières grasses (gr/l)} \\ \text{protéines (gr/l)} \\ \text{caséines (gr/l)} \\ \text{particules sèches de fromage mesurées en usine (gr/l)} \\ \text{particules sèches de fromage mesurées en laboratoire (gr/l)} \\ \text{particules sèches de lait (gr/l)} \\ \text{production de fromage (gr/l)} \end{pmatrix}.$$

Nous avons généré aléatoirement 10 données manquantes selon la répartition suivante :

observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
77	×							
78		×						
79			×					
80				×				
83	×	×						
84			×	×				
85		×	×					

Ce jeu de données contient ainsi des données manquantes et vraisemblablement des données aberrantes.

Trois approches ont été étudiées : détecter puis imputer, imputer puis détecter et finalement imputer par des techniques robustes. Pour évaluer la qualité des imputations, nous avons retenus deux critères conformément à la littérature sur le sujet (Gleason et Staelin, 1975) :

- $Q_\alpha = \sqrt{\dfrac{1}{n_{miss}} \sum_{i=1}^{n} \sum_{j=1}^{p} \dfrac{(X_{ij} - X_{ij}^{\alpha})^2}{\hat{\sigma}_j^2}}$ à savoir la racine de la moyenne (calculée par rapport au nombre de valeurs manquantes) des écarts au carré

entre la valeur réelle X_{ij} et la valeur imputée X_{ij}^{α} pondérés par les variances estimées des variables correspondantes,

- $D_{\alpha} = \max_{i,j} \frac{\mid X_{ij} - X_{ij}^{\alpha} \mid}{\hat{\sigma}_j}$ soit le maximum des différences en valeur absolue entre la valeur réelle et la valeur imputée pondérées par les écarts-types estimés des variables correspondantes.

Pour la détection des données douteuses, nous utiliserons :

1. l'analyse en composantes principales (Q-Q plot des deux premières composantes principales) (défaut principal : lien avec la normalité des composantes principales),
2. la classification hiérarchique (groupe de points dont les éléments s'unissent à un bas niveau de distance et restent longtemps avant de s'unir à un autre groupe) (choix du type de classification),
3. la distance de Mahalanobis (effet de masque),
4. le boxplot (détection uniquement univariée),
5. la méthode de Wilks (rapport des déterminants des matrices de données avec et sans le point suspect)(effet de masque),
6. la fonction d'influence de la corrélation vectorielle ρV (Cléroux, Helbling et Ranger, 1990) (deux groupes de variables à définir).

Les méthodes d'imputation retenues estiment la donnée manquante par :

1. la moyenne (défaut principal : distorsion de la distribution),
2. la prédiction fournie par la régression multiple sur toutes les variables disponibles (sous-estimation de la variabilité),
3. la prédiction fournie par la régression simple avec la variable la plus corrélée (sous-estimation de la variabilité).

Dans chaque cas, l'imputation peut se faire en utilisant uniquement les individus n'ayant pas de données manquantes (COMPLETE) ou en exploitant le maximum d'information disponible (ALLVALUE).

Les méthodes robustes imputeront par la moyenne tronquée ($\alpha = 5\%$) ou par la prédiction de la régression simple avec la variable la plus corrélée selon la méthode LMS (Least Median Square) ou celle utilisant la norme L1.

A. Détecter puis imputer

La détection des données aberrantes porte sur tous les individus à l'exception de ceux qui contiennent des valeurs manquantes. Les méthodes 1, 2 et 4 détectent l'ensemble $\{11, 12, 13, 14, 15\}$ alors que les méthodes 3 et 5 détectent $\{41, 44, 47, 75\}$. La figure 2 fait ressortir ces deux types d'aberrances : l'un formant un groupe de points situés dans la queue de la distribution et l'autre formé de points isolés hors du

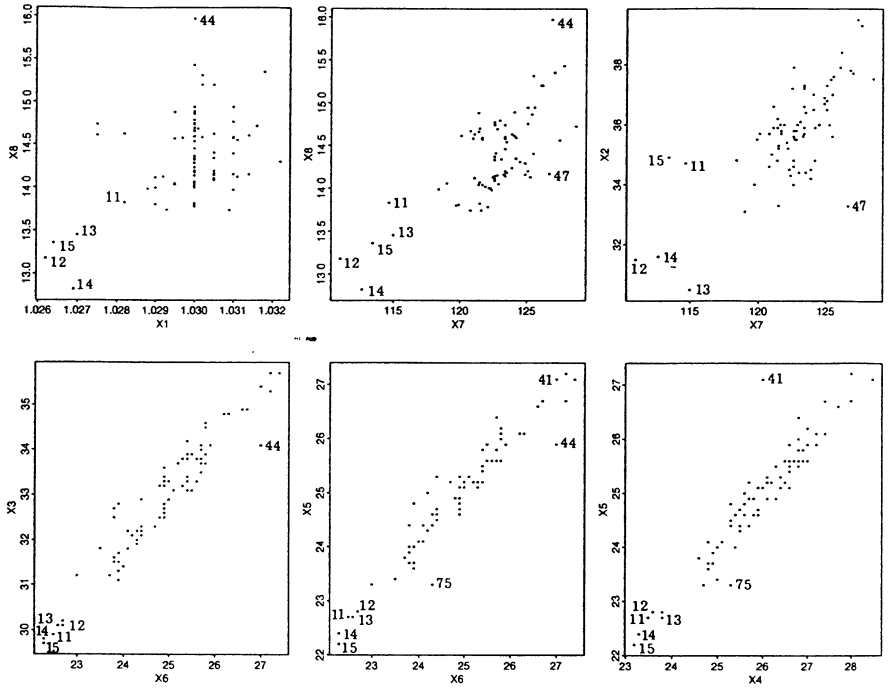

FIGURE 2

nuage. Remarquons que la méthode basée sur ρV révèle comme douteux un ensemble formé de points des deux groupes $\{12, 13, 14, 47\}$.

L'imputation a été effectuée en éliminant les ensembles apparus dans la détection à savoir premièrement sans les observations $\{11, 12, 13, 14, 15, 41, 44, 47, 75\}$ (Groupe 1), deuxièmement sans $\{11, 12, 13, 14, 15\}$ (Groupe 2), troisièmement sans $\{41, 44, 47, 75\}$ (Groupe 3) et finalement sans $\{12, 13, 14, 47\}$ (Groupe 4). Les performances des différentes méthodes sont résumées dans le tableau 1 : (voir page suivante).

Parmi cette grande diversité de résultats, nous constatons qu'il semble optimal d'enlever les aberrances situées dans les queues des distributions marginales et qu'il n'est pas forcément souhaitable de supprimer la réunion des ensembles d'aberrances détectés par les différentes méthodes. Dans notre exemple, la meilleure imputation est obtenue en ôtant les données détectées par la méthode ρV.

B. Imputer puis détecter

Nous avons imputé les neuf données manquantes sur la base des données présentes éventuellement aberrantes puis nous avons appliqué les techniques de détection sur les données complétées c'est-à-dire sur les données présentes et imputées. La qualité des imputations obtenues est présentée dans le tableau 2 :

TABLEAU 1

Imputation $\begin{matrix} Q_\alpha \\ D_\alpha \end{matrix}$	Groupe 1	Groupe 2	Groupe 3	Groupe 4
Moyenne Complète	0.8300 2.4082	0.8329 2.4080	0.8636 2.5820	0.8134 2.4043
Moyenne Allvalue	0.8402 2.4561	0.8315 2.4086	0.8782 2.6215	0.8359 2.4543
Régression simple Complète	0.8702 2.2749	0.8371 2.3313	0.9354 2.2751	0.7535 2.2654
Régression simple Allvalue	0.8135 2.3261	0.8346 2.3212	0.9220 2.3310	0.7642 2.3073
Régression multiple Complète	0.9901 2.2813	0.9721 2.4295	1.0431 2.2136	0.9912 2.3205
Régression multiple Allvalue	0.8715 2.3480	0.9810 2.4319	0.9478 2.2413	0.9065 2.9602

TABLEAU 2

Imputation	$\begin{matrix} Q_\alpha \\ D_\alpha \end{matrix}$
Moyenne Complète	0.8618 2.5731
Moyenne Allvalue	0.8754 2.6112
Régression simple Complète	0.9386 2.3283
Régression simple Allvalue	0.9038 2.3738
Régression multiple Complète	1.0457 2.3735
Régression multiple Allvalue	0.9552 2.3825

La qualité de l'imputation est moins bonne que celle du tableau 1, ce qui concorde avec plusieurs résultats de la littérature (Little, 1988).

L'application par la suite des méthodes de détection aboutit aux mêmes ensembles de données douteuses que ceux de la première approche à une exception près : l'observation 77 imputée par les méthodes Moyenne est détectée comme aberrante par la méthode Mahalanobis (la valeur de chacune des variables observées est parmi les plus faibles de tous les individus et donc son association avec la moyenne imputée créera une aberrance).

C. Imputer par des techniques robustes

L'imputation a été obtenue sur la base des observations des individus n'ayant pas de données manquantes à l'aide de fonctions robustes du logiciel Splus.

TABLEAU 3

Imputation	Q_α	D_α
moyenne tronquée Complète ($\alpha = 5\%$)	0.8530	2.5339
méthode L1 Complète	0.9045	2.2402
méthode LMS Complète	0.8025	2.1692

Cette approche est performante puisque les techniques robustes donnent de meilleurs résultats que les méthodes standards équivalentes avant détection (tableau 2) et une qualité proche de celle de l'imputation après détection (tableau 1). Selon les logiciels à disposition, on optera pour la méthode LMS ou pour la moyenne tronquée.

Intuitivement, les approches A et C aboutissent à des résultats proches, car si une donnée aberrante existe son poids dans les procédures robustes sera négligeable; nous constatons cela aussi dans la pratique.

Globalement, nous pouvons relever les points suivants. Conformément à de nombreux articles, la méthode de régression simple se révèle généralement supérieure aux autres méthodes; notons que dans notre exemple les corrélations simples sont assez fortes (la corrélation de chaque variable ayant des données manquantes avec une autre variable se situe entre 0.5 et 0.95).

Dans le domaine de la détection des aberrances, il est fréquent qu'une donnée soit identifiée comme une aberrance par une méthode et ne pas l'être par une autre. Dès lors, il est important que le statisticien qui trouve une ou plusieurs données douteuses en discute avec l'expérimentateur avant de choisir l'ensemble à écarter pour l'approche A. Si aucune décision ne peut être prise, il est certainement préférable d'imputer par une technique robuste.

Références

BARNETT V. et LEWIS T. 1994. *Outliers in Statistical Data*, Wiley.

BECKMAN R.J. et COOK R.D. 1983. Outlier....'s, *Technometrics*, **25** 119-163.

BUCK S.F. 1960. A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *JRSS B*, **22**, 302-306.

CLÉROUX R. , HELBLING J.-M. ET RANGER N. 1990. Détection d'ensembles de données aberrantes en analyse des données multivariées, RSA, **38**, 5-21.

CRETTAZ de ROTEN F. 1993. *Données manquantes en statistique multivariée : une nouvelle méthode basée sur le coefficient RV*, Thèse no 1111 Ecole Polytechnique Fédérale de Lausanne.

DAUDIN J.J. , DUBY C. et TRÉCOURT P. 1988. Stability of principal component analysis studied by the bootstrap method, *MaOpfSTS*, **19**, 241-258.

DEMPSTER A.P., LAIRD N.M. et RUBIN D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm, *JRSS B*, **39**, 1-38.

EFRON B. 1993. Missing data, imputation and the bootstrap, *JASA*, **89**, 463-480.

FORD B.L. 1983. An overview of hot-deck procedures, *dans «Incomplete data in sample survey»* , **Academic Press**, chapitre 14.

FRANE J.W. 1977. Some simple procedures for handling missing data in multivariate analysis, *Psychometrika*, **41**, 409-415.

GILES P. 1988. A model for generalized edit and imputation of survey data, *The Canadian Journal of Statistics*, **16**, 57-73.

GLEASON T.L. et STAELIN R. 1975. Proposal for handling missing data, *Psychometrika*, **4**, 229-252.

HADI A.S. et SIMONOFF J.S. 1993. Procedures for the identification of multiple outliers in linear models, *JASA*, **88**, 1264-1272.

LITTLE R.J.A 1988. Robust estimation of the mean and the covariance matrix from data with missing values, *Appl Stat*, **37**, 23-38.

LITTLE R.J.A. et RUBIN D.B. 1987. *Statistical Analysis with Missing Data*, Wiley.

LITTLE R.J.A. et SMITH P.J. 1987. Editing and imputation for quantitative survey data, *JASA*, **82**, 58-68.

ROUSSEEUW P.J. et LEROY A.M. 1987. *Robust Regression and Oulier Detection*, Wiley.

RUBIN D.B. 1976. Inference and missing data, *Biometrika*, **63**, 581-592.