Gestion Des Donnees Manquantes Dans Les Bases De Donnees En Sciences Sociales : Algorithme Nipals Ou Imputation Multiple?

Njamen Kengdo Arsène Aurélien
Doctorant PhD, Département d'Analyse et Politiques Economiques,
Université de Dschang (Cameroun)
Kwatcho Kengdo Steve
Global Change Ecology, Bayreuth Center of Ecology and Environmental Research. University of Bayreuth (Germany).

doi: 10.19044/esj.2016.v12n35p390 URL:http://dx.doi.org/10.19044/esj.2016.v12n35p390

Abstract

The main objective of this paper is to assess the robustness of imputation methods to fill up the series of secondary data in social sciences. The methodology used, especially that of mean imputation, multiple imputation and NIPALS algorithm, is based on a simulation using observed data. Results show a close similarity between the observed data and the data obtained by multiple imputation, mean imputation and NIPALS algorithm. The results also suggest that multiple imputation provides values substantially similar to observed data.

Keywords: Missing data, multiple imputation, NIPALS Algorithm

Résumé:

L'objectif principal de ce papier est d'évaluer la robustesse des méthodes d'imputation pour combler les séries de données secondaires en sciences sociales. La méthodologie utilisée, notamment celle de l'imputation par la moyenne, l'imputation multiple et l'imputation par l'algorithme NIPALS, se base sur une simulation à partir des données observées. Les résultats montrent une proche similitude entre les données observées et les données obtenues par l'imputation multiple, l'imputation par la moyenne et l'algorithme NIPALS. Les résultats suggèrent également que c'est l'imputation multiple qui fournit des valeurs sensiblement proche de celles observées.

Mots-clés : Données manquantes, imputation multiple, Algorithme NIPALS

Introduction

Les recherches en Sciences Sociales, en général et en Sciences Economiques en particulier, sont confrontées à un problème majeur lié à l'absence d'observation dans des séries de données sur une ou plusieurs années. Ces observations manquantes sont le plus souvent fréquentes dans les données des pays en voie de développement où le processus de collecte est souvent mis à mal par certaines contraintes (absence d'organisme fiable de collecte de données, financement, accessibilité à l'information, insécurité et instabilité politique). Cela constitue un problème important dans le contexte de l'extraction des connaissances à partir de bases de données (Preda et al. 2005).

Il faut gérer les données manquantes avec beaucoup de délicatesse afin d'éviter de détériorer les performances des procédures économétriques et la réalité économique, sociale et dans une certaine mesure politique, que ces données traduisent. Au cours des dernières décennies, les recherches sur cette problématique se sont multipliées. De même, les logiciels d'analyse statistique commencent à y proposer des solutions (Celeux, 1988 ; Hox, 1999).

Dans le processus de gestion de données manquantes, il est mis en exergue trois stratégies principales (Preda et al. 2005). La première stratégie consiste à éliminer la ou les variables qui présentent des données manquantes. Ici, bon nombre de chercheurs abandonnent la variable qui présente des valeurs manquantes au profit d'une autre variable « *proxy* » susceptible de représenter la même réalité. Cette stratégie est largement critiquée, car présente deux limites majeures. D'une part, la perte d'information ainsi obtenue peut être considérable si de nombreuses variables ont des données manquantes sur différents « *individus* ». D'autre part, cette stratégie risque d'introduire un biais si le processus qui conduit aux valeurs manquantes n'est pas complétement aléatoire (Little et Rubin, 1987).

La seconde stratégie consiste à utiliser une méthode propre à l'algorithme de traitement de données utilisé (Breiman et al. 1984). Un exemple est l'algorithme CART (Classification And Regression Trees), qui s'attelle à construire un arbre de décision binaire en classifiant un ensemble d'enregistrements. Cet arbre fournit un modèle pour classer de nouveaux échantillons sur la base d'un critère de segmentation qui repose sur l'indice de Gini. On peut également citer l'algorithme ID3 et C4.5, qui sont des dérivés de CART. Les méthodes propres à cette stratégie supposent implicitement un mécanisme de données manquantes complètement aléatoire, mais présentent un inconvénient lié au fait qu'elles n'ont pas été évaluées, pour la plupart d'entre elles (Preda et al. 2005).

Face aux limites associées aux deux précédentes stratégies, une troisième stratégie qui est celle de l'imputation est envisagée (Allison, 2000). Elle consiste à remplacer la donnée manquante par une valeur plausible obtenue à partir des observations disponibles. Rubin (1987), Allison (2000), Preda et al. (2005), montrent que cette méthode est bien adaptée au processus d'extraction des connaissances à partir de bases de données, puisque la base complétée peut être analysée par toutes les procédures économétriques et statistiques.

Il existe plusieurs méthodes d'imputation. Les plus simples sont l'imputation par la moyenne, la médiane et le mode (Schafer, 1999). On note également la méthode d'imputation basée sur la régression (Horton et Lipsitz, 2001), les méthodes basées sur les procédures de classification (Benali et Escofier, 1987) et les règles d'association (Ragel, 1999), la méthode d'imputation par l'algorithme NIPALS (Tenenhaus, 1998) et l'impution multiple (Rubin, 1987). Preda et al. (2005) montrent que l'imputation multiple et l'imputation par l'algorithme NIPALS, sont plus adaptées aux situations où les données manquantes sont observables sur plusieurs variables de la base de données et peuvent facilement être implémentées à l'aide de logiciels usuels.

Le problème qui se pose est qu'avec cette multiplicité des méthodes d'imputation, les auteurs dans leurs études adoptent différentes approches. Cette étude se propose alors de trouver un consensus en ce qui concerne cette problématique.

Ainsi, sur la base de cette classification, on peut se poser la question relative à la méthode d'imputation adaptée pour combler les bases de données en Sciences Sociales.

L'objectif de ce papier est d'évaluer la robustesse des méthodes d'imputation. Pour ce faire, nous allons procéder à une analyse comparative entre la méthode d'imputation multiple, l'imputation par l'algorithme NIPALS, et l'imputation par le moyenne.

Ce papier est organisé de la manière suivante. Après l'introduction, La section 1 présente les méthodes d'imputation. La section 2 expose la méthodologie qui débouche sur les résultats en section 3. Enfin, la section 4 porte sur la conclusion et les recommandations.

Presentation des méthodes d'imputation

Dans cette section, nous allons présenter succinctement la méthode d'imputation par la moyenne, l'algorithme NIPALS, la méthode d'imputation multiple et les axiomes relatives aux méthodes d'imputation. Mais avant d'y arriver, nous procédons à la clarification du concept de « donnée manquante ».

Le concept de « donnée manquante »

En statistique, on parle de donnée manquante lorsqu'on n'a pas d'observations sur une variable donnée pour un individu donnée (Meyer, 2006). Un indicateur important de la qualité des données est alors la fraction de données manquantes (Leeuw et al. 2008). Il existe trois types de données manquantes : les données manquantes complètement aléatoires (MCAR[15]), les données manquantes aléatoires (MAR[16]) et les données manquantes non aléatoires (MNAR[17]) (Allison, 2001).

- Les données sont manquantes complètement aléatoirement si la probabilité qu'une observation soit manquante ne dépend pas des mesures observées ou non observées. En termes mathématiques, cela s'écrit :

$$P(r \mid X_{obs}, X_{miss}) = P(r)$$

Où r représente la réponse, X_{obs}, X_{miss} respectivement les valeurs observées et manquantes.

- Les données sont manquantes aléatoirement si, connaissant les données observées, le mécanisme de non réponse ne dépend pas des données non observées. Mathématiquement, on écrit :

$$P(r \mid X_{obs}, X_{miss}) = P(r \mid X_{obs})$$

On remarque que dans ce cas, on peut mener des analyses en utilisant uniquement l'information observée.

- Le cas des données manquantes non aléatoirement correspond à un mécanisme de non réponse. Ce qui signifie que même en tenant compte de toute l'information, les raisons pour lesquelles les données sont manquantes dépendent des données manquantes elles-mêmes. Ici, pour obtenir des estimations valides, un modèle complété sachant le mécanisme de réponse est nécessaire.

Néanmoins, une limite importante apparait à ce niveau. On ne peut généralement pas dire, à partir des données, quelle est la raison pour laquelle la donnée manque (MCAR, MAR ou MNAR)

Dans le cadre de cette étude, nous nous limitons au cas des données manquantes aléatoirement (MAR). Une distinction doit être faite à ce niveau entre « saut de donnée » et donnée manquante aléatoire. En effet, le « saut de donnée » intervient lorsque l'information pour une variable n'est pas disponible sur une ou plusieurs années, ceci du fait qu'elle n'est pas observable ou bien que le répondant ne révèle pas l'information (Leeuw et al. 2008). Le « saut de donnée » s'inscrit alors dans le cadre de données non manquantes aléatoirement.

[15] En anglais : Missing Completely At Random
[16] Missing At Random
[17] Missing Not At Random

Le problème de la gestion des données manquantes est un vaste sujet. Ces données ne peuvent pas être ignorées lors d'une analyse statistique et économétrique. On peut dès lors retirer les variables ou les individus qui présentent des valeurs manquantes ou imputer des valeurs aux données manquantes. A propos de cette dernière, différentes méthodes existent pour gérer ce problème (Glasson-Cicognani et Berchtold, 2010).

L'imputation par la Moyenne

L'imputation par la moyenne est la méthode la plus utilisée par les économistes. Elle consiste à remplacer la donnée manquante par une valeur moyenne tirée des observations disponibles ayant le même jeu de caractéristiques prédéterminées (Schafer, 1999). Celle-ci est largement critiquée ces dernières années car les utilisateurs ne se préoccupent pas de la raison pour laquelle la donnée manque. Par ailleurs, cette méthode conduit à une réduction systématique de la dispersion de chacune des variables et risque de briser d'éventuelles relations multidimensionnelles sous-jacentes entre les variables.

L'imputation par l'algorithme NIPALS

L'algorithme NIPALS (Nonlinear Iterate PArtial Least Squares) est une méthode itérative, proche de la régression PLS (Partial Least Squares)[18], utilisée pour estimer les éléments d'une Analyse en Composantes Principales (ACP) d'un vecteur aléatoire de dimension finie. Il permet d'estimer les paramètres d'un modèle non linéaire à l'aide d'une suite de régressions simples entre les données et une partie des paramètres. Cet algorithme est adapté à l'imputation des données manquantes par Tenenhaus (1998) en l'appliquant sur les données pour obtenir un modèle ACP. Ce modèle ACP est ensuite utilisé pour prédire les données manquantes.

Pour ce faire, on dispose d'individus *i* sur un nombre *p* de variables quantitatives.

$$\text{Soit } Y = (Y_1, \dots, Y_p) \text{ tel que } \forall_i \in 1 \dots p \; ; E(Y_i) = 0$$

L'expansion de Y, en termes de composantes principales et de facteurs principaux, est donnée par l'expression suivante.

$$Y = \sum_{h=1}^{q} \xi_h \, u_h \; \dots\dots\dots\dots\dots\dots (1)$$

Où $q = \dim L_2(Y)$ et $\{\xi_h\}_{h=1\dots q}$, sont des composantes principales et $\{u_h\}_{h=1\dots q}$ les vecteurs principaux de l'ACP de Y.

[18] La régression PLS ou Moindres Carrés Partiels est inventée en 1983 par Svante Wold et Herman Wold. Elle maximise la variance des prédicteurs (X_i)=X et maximise la corrélation entre X et la variable à expliquer Y. Cet algorithme emprunte sa démarche à la fois à l'Analyse en Composantes Principales (ACP) et à la régression par les Moindres Carrés Ordinaires (MCO). Il cherche des composantes appelées variables latentes, liées à X et à Y, servant à exprimer la régression de Y sur ces variables et enfin de Y sur X.

Pour chaque variable Y_i, on a : $Y_i = \sum_{h=1}^{q} \xi_h \, u_h(i)$ …… (2)

L'idée étant que pour chaque h, $u_h(i)$ représente la pente de la régression linéaire de Y_i sur la composante ξ_h.

Le but de l'algorithme NIPALS est de permettre d'obtenir $\{\hat{\xi}_h\}_{h=1\ldots q}$ et $\{\hat{u}_h\}_{h=1\ldots q}$, les approximations de $\{\xi_h\}_{h=1\ldots q}$ et $\{u_h\}_{h=1\ldots q}$. L'algorithme fonctionne de la manière suivante.

Algorithme NIPALS

---

1 - $Y^O = Y$

2 - pour $h = 1 \ldots q$ **faire**

(a) $\xi_h = Y_1^{h-1}$

(b) **tant que** u_h n'a pas convergé **faire**

i - pour $j = 1 \ldots q$ **faire**

$$u_h(j) = \frac{\sum_{i:\, y_{ji} \, et \, \xi_h(i) \, existent} \; y_{ji}^{h-1} \xi_h(i)}{\sum_{i:\, y_{ji} \, et \, \xi_h(i) \, existent} \; {\xi_h}^2(i)}$$

ii - normaliser u_h à 1

iii - pour $i = 1 \ldots n$ **faire**

$$\xi_h(i) = \frac{\sum_{j: y_{ij} \, existe} \; y_{ij}^{h-1} u_h(j)}{\sum_{j: y_{ij} \, existe} \; {u_h}^2(j)}$$

(c) $Y^h = Y^{h-1} - \xi_h u_{h'}$

Source : Tenenhaus (1998)

Selon Tenenhaus (1998), l'idée forte de l'algorithme NIPALS réside dans l'interprétation des étapes 2-(b)-i et 2-(b)-iii : on calcule à chaque fois les pentes des droites des moindres carrés, passant par l'origine, des nuages de points sur les données disponibles. Cela se matérialise dans les expressions suivantes.

$\{(\xi_h(i), y_{ij}^{h-1})_{i=1\ldots n} \, ; où \, (\xi_h(i) \, et \, y_{ij} \, existent\}$ ……. (3)

$\{u_h(j), y_{ji}^{h-1})_{j=1\ldots p} \, ; et \, (\xi_h(i) \, et \, y_{ji} \, existe\}$ …… (4)

On pourra alors approximer les données manquantes par :

$(\hat{y}_{ij})_{manquante} = \sum_{h=1}^{q} \hat{\xi}_h(i) \, \hat{u}_h(j)$ …….. (5)

Mais il faut noter que la convergence de l'algorithme est assurée lorsqu'il n'y a pas trop de données manquantes. Il permet en particulier de faire une Analyse en Composantes Principales en présence de valeurs manquantes. Cet algorithme s'intéresse au calcul du vecteur propre d'une matrice associé à la plus grande valeur propre d'un jeu de données. La sous-section suivante présente la méthode d'imputation multiple.

L'imputation multiple

La méthode par imputation multiple a été proposée pour la première fois par Rubin (1978), puis développée par Rubin (1987) et repris par Schafer (1997). Elle consiste à remplacer une valeur manquante par *m* valeurs plausibles au sens d'un modèle statistique (m > 1). Rubin (1987) décrit la méthode comme une succession de trois étapes : tout d'abord on attribue des valeurs aux données manquantes en utilisant un modèle aléatoire adapté. Ensuite, on répète *m* fois l'étape précédente afin d'obtenir les *m* tableaux de données complétées. Enfin, on analyse ces *m* tableaux en utilisant une méthode statistique standard d'analyse de données complétées matérialisée par la formule suivante :

$\beta_i^* = \frac{1}{m}\sum_{j=1}^{m} \beta_{i,j}^*$, avec β les données complétées

Plus le nombre *m* d'imputation est grand, plus les estimations sont précises. Mais Selon Rubin (1987), en pratique, à partir d'un nombre d'imputation relativement faible, on obtient de bons résultats ; notamment pour *m* = 5.

Cette méthode, qui utilise un algorithme d'imputation basé sur les chaînes de Markov, peut nécessiter jusqu'à 1000 itérations (Van Buuren, 2007). L'algorithme fonctionne de la manière suivante :

- Des valeurs initiales pour les données manquantes sont obtenues en tirant aléatoirement des valeurs sur une loi normale de moyenne et variance égale à la moyenne et à la variance obtenues sur les données disponibles.
- Pour chaque variable du jeu de données ayant des données manquantes, une méthode d'imputation basée sur l'échantillonnage dans une distribution et le modèle des Moindres Carrés Ordinaires est appliquée. Le modèle utilisé est un modèle de régression ayant la variable étudiée comme variable dépendante et les autres variables du jeu de données comme variables indépendantes. Des valeurs aléatoires tirées sur des lois définies sont utilisées pour apporter une part aléatoire au modèle. Les valeurs imputées sont obtenues à partir du modèle estimé.

Ces deux étapes sont répétées autant de fois (*m* fois) que le requiert l'utilisateur. La valeur moyenne de chaque donnée manquante imputée est utilisée.

Axiomes relatives aux méthodes d'imputation

L'appréhension des données manquantes est un problème délicat (Niass et al. 2013). Il est nécessaire de poser des axiomes de départ afin de délimiter l'environnement de notre étude.

Axiome 1 : cette méthode n'est applicable que sur des données de sources secondaires, de nature quantitative. En effet, les simulations sur

données qualitatives donnent des résultats qui s'écartent le plus souvent de la réalité.

Axiome 2 : la méthode est applicable pour une série contenant au plus 50% de données manquantes, pour produire des résultats acceptables (Kouadjo et al. 2013).

Axiome 3 : on considère des variables quantitatives pour lesquelles les données manquent aléatoirement.

Axiome 4 : l'imputation est appliquée sur des séries de données dont les valeurs sont positives.

La section suivante évoque la méthodologie utilisée pour atteindre l'objectif principal de notre étude.

Méthodologie

Afin de tester empiriquement la robustesse des méthodes d'imputation, nous procédons à une simulation à partir d'une série de données observées. Ces données font référence au PIB par habitant du cameoun sur la periode 1990-2015. ces données sont obetenues à partir de la base de données de la banque mondiale (WDI, 2015)

La procédure est la suivante : dans la série de données, nous supprimons les observations sur cinq ans (1996-2000). Par la suite, nous utilisons la méthode d'imputation par la moyenne, l'algorithme NIPALS et celle de l'imputation multiple pour voir si les données imputées se rapprochent des données observées. Les critères de décision portent sur la moyenne des observations, la dispersion des estimateurs[19] et l'écart entre les données observées et les données imputées. Le logiciel utilisé est XLSTAT v5.03. Les résultats sont présentés dans la section suivante.

Résultats

Nous allons tout d'abord présenter les résultats relatifs à chaque méthode d'analyse, ensuite faire une analyse comparative des différentes approches pour décliner la méthode la plus appropriée pour combler les données qui manquent.

Résultats relatifs aux différentes méthodes

Le tableau suivant recense les résultats obtenus après imputation par la moyenne, l'imputation multiple et l'algorithme NIPALS.

[19] Cette dispersion est égale à $\frac{ecart\ type\ (avant\ traitement) - ecart\ type\ (après\ traitement)}{ecart\ type\ (avant\ traitement)}$

Tableau 1: résultats des simulations

Années	PIB/Hab observé	imputation par la moyenne	imputation multiple	Algorithme NIPALS
1990	923,88	923,88	923,88	923,88
1991	1000,32	1000,32	1000,32	1000,32
1992	890,56	890,56	890,56	890,56
1993	1027,56	1027,56	1027,56	1027,56
1994	680,63	680,63	680,63	680,63
1995	626,95	626,95	626,95	626,95
1996	679,76	**925,48**	**827,84**	**909,04**
1997	668,97	**925,78**	**781,23**	**1061,95**
1998	637,36	**926,69**	**847,05**	**1022,81**
1999	675,92	**925,59**	**840,19**	**1051,73**
2000	583,09	**928,24**	**935,31**	**1069,07**
2001	589,20	589,20	589,20	589,20
2002	648,39	648,39	648,39	648,39
2003	791,10	791,10	791,10	791,10
2004	892,89	892,89	892,89	892,89
2005	915,09	915,09	915,09	915,09
2006	965,36	965,36	965,36	965,36
2007	1070,95	1070,95	1070,95	1070,95
2008	1191,70	1191,70	1191,70	1191,70
2009	1164,71	1164,71	1164,71	1164,71
2010	1147,24	1147,24	1147,24	1147,24

Source : Auteur à partir de XLSTAT

Sur la base de ces résultats, nous pouvons faire les commentaires suivants concernant chaque méthode d'imputation. Ces commentaires portent principalement sur l'analyse de la moyenne et de la dispersion des écarts type.

La méthode d'imputation par la moyenne donne une moyenne des observations de l'ordre de 911.7215, de loin superierue à la moyenne des données observées (846.26). La dispersion au niveau de l'écart type est égale à 14.97%.

L'imputation multiple donne des résultats intéressant, avec une moyenne des observations de 891.71, sensiblement proche de la moyenne sur données observées. Au niveau de la dispersion entre écarts type observé et imputé, on a une dispersion de 13.14%.

Concernant l'imputation par l'algorithme NIPALS, on observe une moyenne des observations de 935.86, très superieure à la moyenne sur données observées. Paradoxalement, la dispersion au niveau des écarts type donne une valeur de 10.16%.

Globalement, les résultats ci-dessus confirment le bien fondé des méthodes d'imputation car on observe une dispersion des écarts type inferieure à la limite de 20% (Niass et al. 2013).

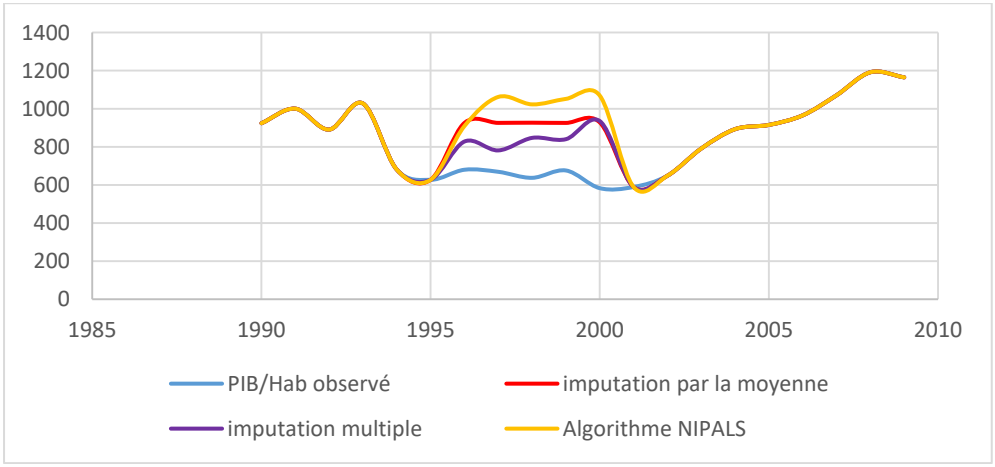
Mais il apparait une contradiction au niveau des critères de décision. En effet, sur la base des travaux de Preda et al. (2005), on s'attendait à une corrélation entre la moyenne des observations et la dispersion des estimateurs, ce qui n'est pas le cas. Pour preuve, les résultats de l'algorithme NIPALS donne une moyenne des observations de loin supérieure à celles des autres méthodes, mais une dispersion inferieure à celles des deux autres méthodes. Ceci montre que le critère de 20% définit par Niass et al. (2013) est une condition nécessaire, mais pas suffisante. La sous-section suivante se concentre sur une analyse comparative graphique des différentes méthodes pour voir la quelle produit des données proches de celles observées.

Analyse comparative des différentes méthodes d'imputation

Cette sous-section se concentre principalement sur une analyse graphique entre données observées et données complétées.

Le graphique suivant permettent d'apprécier la différence entre imputation par la moyenne, algorithme NIPALS et imputation mutiple.

Figure 1 : représentation graphique de la différence entre imputation multiple et algorithme NIPALS



Source : Auteurs

Ce graphique montre une très proche similitude entre les données observées et les données obtenues par l'imputation multiple, malgré le déphase observé. L'analyse graphique vient ainsi, en complément de l'analyse des dispersions des écarts type, comme une condition suffisante.

De manière générale, on conclut que c'est l'imputation multiple qui produit des données proches de celles observées.

Conclusion et recommandations

Cette étude nous a permis d'analyser la robustesse de la méthode d'imputation multiple, imputation par la moyenne et de l'algorithme

NIPALS pour combler les valeurs manquantes dans les bases de données. Les résultats suggèrent que ces méthodes sont valides pour combler les données manquantes, car satisfont le critère de Niass et al. (2013). Mais à la différence de l'imputation par l'algorithme NIPALS et l'imputation par la moyenne, c'est l'imputation multiple qui fournit des valeurs très proche des données observées.

Dans l'optique de rendre les données plus représentatives de l'environnement économiques, les recommandations suivantes sont formulées à l'intention des chercheurs et des organismes de collecte de données.

Il s'agit pour les organismes de collecte de données de mettre à la disponibilité de la communauté scientifique, des documents qui expliquent les raisons de l'absence des valeurs dans les bases de données. La conséquence pourrait être une déformation de l'information traduite par les données dans le cas où on comblerait des valeurs manquantes alors que ces dernières ne sont pas observables dans la réalité. Les institutions de collecte de données doivent également organiser des séminaires de gestion des données manquantes dans les bases qu'elles mettent à la disposition du public.

Certes l'avancée technologique est à l'origine de la multiplication des logiciels et des méthodes de gestion des données manquantes, mais les chercheurs doivent s'assurer de la robustesse des méthodes que ces différents logiciels présentent afin de mieux affiner la qualité de l'information extraite des données. Par ailleurs, le recourt à une quelconque méthode de gestion de données manquantes doit être clairement justifié, afin de ne pas biaiser les analyses.

References:

1. Allison P. D. (2000). « Multiple imputation for missing data : A cautionaary tale », *Sociological Methods Research*, vol 03, n°28, pp.301-309.
2. Benali H., Escofier B. (1987). « Nouvelle etape de traitement des données manquantes en analyse factorielle des correspondances multiples dans le système portable d'analyse de données », Rapports Tecniques n°85, Institut National de Recherche en Informatique et en Automatique (INRIA), France.
3. Breiman L., Friedman J. H., Ohlsenn R. A., Stone C. J. (1984). « Classification and regression trees », Belmont Wadsworth.
4. Celeux G. (1988). « Le traitement des données manquantes dans le logiciel SICLA », Rapport Techniques n°102, Institut National de Recherche en Informatique et en Automatique (INRIA), France.

5. Glasson-Cicognani et Berchtold. (2010). « Imputation des données manquantes : comparaison de différentes approches », 42èmes journées de Statistique, Marseille, France.
6. Horton N. J., Lipsitz S. R. (2001). « Multiple imputation in practice: comparison of software packages for regression models with missing variables », *Statistical Computing Sofware Reviews*, The American Statistician, vol 3, n°55.
7. Hox J. (1999). « A review of current software for handling missing data », *Kwantitatieve Methoden*, n°62, pp.123-138.
8. Kouadjo J. M., Kouakou J. A., Kouamé D. K. (2013). « Methodologie d'obtention d'une base de données imputées », *The African Statistical Journal*, vol 16, pp.61-80.
9. Leeuw, Edith D. de, and Joop Hox. (2008). «Missing Data », *Encyclopedia of Survey Research Methods*, retrieved from : http://sage-ereference.com/survey/Article_n298.html
10. Little R. J., Rubin D. B. (1987). « Statistical analysis with missing data », Wiley, New York.
11. Meyer N. (2006). « Les données manquantes en statistique », Séminaire de Statistique, Laboratoire de Biostatistique - Faculté de Médecine Dep. Santé Publique CHU – STRASBOURG, 7 novembre 2006
12. Niass Omy, Touré Aissatou, Diongue Abdou et Dabye Souleymane. (2013). « Gestion des données manquantes dans les études séro-épidémiologiques », Laboratoire d'Etudes et de Recherches en Statistiques et Développement (LERSTAD), Sénégal.
13. Preda C., Duhamel A., Picavet M., Kechadi T. (2005). « Gestion des données manquantes dans les grandes bases de données en santé », Journée Francophones d'Informatique Medicale, Lille 12-13 mai.
14. Ragel A. (1999). « MVC- A preprocessing method to deal with missing values », Knowledge Based Systems, vol 5, n°12, pp.285-291.
15. Rubin. (1987). « Multiple imputation for nonresponse in surveys », New York: John Wiley.
16. Schafer et Graham. (2002). « Missing Data: our view of the state of the art », *Psychological Methods*, n°7, vol 2, pp.147-177.
17. Schafer, J. L. (1997). « Analysis of Incomplete Multivariate Data ». London: Chapman and Hall.
18. Schafer, J. L. (1999). « Imputation procedures for missing data », University of Pennsylvania, USA.
19. Tenenhaus M. (1998). *La régression PLS : théorie et pratique*, Editions Technip, Paris.

20. Van Buuren. S. (2007). « Multiple imputation of discrete and continuous data by fully conditional specification ». *Statistical Methods in Medical Research*, n°16, pp.219 – 242.

Annexes

- **Statistiques descriptives avant simulations**

Variable	Obs	Obs. avec données manquantes	Min	Max	Moyenne	Ecart-type
PIB/Hab	21	5	583,09	1191,70	846,26	204,00

Source : XLSTAT

- **Imputation par la moyenne : Statistiques descriptives Après simulation**

Variable	Obs	Obs. avec données manquantes	Min	Max	Moyenne	Ecart-type
PIB/Hab	21	0	589,20	1191,70	911,72	173,45

Source : XLSTAT

- **Imputation multiple : Statistiques descriptives Après simulation**

Variable	Obs	Obs. avec données manquantes	Min	Max	Moyenne	Ecart-type
PIB/Hab	21	0	589,20	1191,70	891,71	177,18

Source : XLSTAT

- **Algorithme NIPALS : Statistiques descriptives Après simulation**

Variable	Obs	Obs. avec données manquantes	Min	Max	Moyenne	Ecart-type
PIB/Hab	21	0	589,20	1191,70	935,86	183,27

Source : XLSTAT