Management of missing data in databases: the multiple imputation method in XLSTAT

NJAMEN KENGDO, Arsène Aurélien

Université de Dschang, Faculté des Sciences Economiques et de Gestion

8 January 2016

Gestion des données manquantes dans les bases de données : la méthode d'imputation multiple sous XLSTAT

NJAMEN KENGDO Arsène Aurélien
Doctorant au Département de Sciences Économiques, Université de Dschang ;
arsenekengdo@yahoo.fr

Résumé *: L'objectif principal de ce papier est d'évaluer la robustesse de la méthode de gestion de données manquantes, dite d'imputation multiple, dans les séries de données secondaires en sciences sociales. Nous utilisons une simulation à partir des données observées pour voir la portée de la méthode d'imputation multiple. Les résultats montrent une proche similitude entre les données observées et les données imputées.*

Mots-clés : Données manquantes, imputation multiple

Classification JEL : C15, C82.

Abstract *: The objective of this study is to evaluate the robustness of the missing data management method, called multiple imputation, in the series of secondary data in social sciences. We use a simulation using data observed to see the scope of the multiple imputation method. Results show a close similarity between the observed data and imputed data.*

I. Introduction

Les recherches en Sciences Sociales, en général et en sciences économiques en particulier, sont confrontées à un problème majeur lié à l'absence d'observation dans des séries de données sur une ou plusieurs années. Face à cette difficulté, bon nombre de chercheurs abandonnent le plus souvent la variable concernée au profit d'une autre variable susceptible de représenter la même réalité. Certains, dans un cas extrême, abandonnent complètement la thématique traitée. Mais de plus en plus, les chercheurs en sciences sociales s'attellent à utiliser différentes méthodes pour combler les données manquantes (moyenne mobile, moyenne géométrique, moyenne arithmétique, etc.). Mais ces méthodes, que nous qualifions de « traditionnelles », posent des biais considérables car les adeptes de ceux-ci ne se préoccupent pas de la raison pour laquelle la donnée manque. Face à ces limites, les méthodes d'imputation se développent. Elles consistent à remplacer la donnée manquante par une valeur plausible obtenue à partir des informations disponibles. L'objectif de cette étude est de simuler la méthode d'imputation multiple sur une série de données observées. Pour ce faire, nous allons présenter les méthodes d'imputation, la méthode de simulation utilisée et les résultats obtenus.

II. Présentation des méthodes d'imputation

Il existe trois types de données manquantes (Allison, 2001) : les données manquantes complètement aléatoirement, les données manquantes aléatoirement et les données manquantes non aléatoirement.

Ainsi, si une donnée manquante ne dépend d'aucune variable observable et d'aucun paramètre non observable, alors elles sont complètement aléatoirement. Le fait qu'une donnée manque est alors considéré comme dû au hasard. Dans ce cas, les analyses effectuées sont non biaisées.

En outre, si une donnée manquante est liée à la valeur d'une variable externe, mais pas aux valeurs de la variable ayant des données manquantes, alors les données manquent aléatoirement : c'est le cas le plus classique.

Par ailleurs, si les données manquent pour une raison particulière, alors les données manquent non aléatoirement. Un exemple simple est le cas des questions filtrées (certaines questions ne concernent que certaines personnes dans un questionnaire, les autres personnes sont manquantes).

Dans le cadre des recherches en sciences économiques, nous faisons face à des données qui manquent aléatoirement. Différentes méthodes existent pour gérer ce problème (Glasson-Cicognani et Berchtold, 2010). La première solution consiste à exclure du fichier de données tous les individus ayant au moins une donnée manquante. Ce qui permet d'effectuer des analyses sur des cas où toutes les données sont valides. Une autre solution est l'imputation simple qui consiste à remplacer chaque donnée manquante par une valeur plausible. Par exemple, remplacer toutes les données manquantes par la moyenne calculée sur les données réellement observées. D'autres méthodes d'imputation simple sont également disponibles comme l'imputation par le plus proche voisin (remplace les données manquantes par les valeurs provenant d'individus similaires pour lesquels toute l'information est observée), l'imputation par régression (remplace les données manquantes par des valeurs prédites selon un modèle de régression).

Mais de sérieux problèmes sont relevés dans l'application de ces méthodes[1]. Face à ces manquements, Glasson-Cicognani et Berchtold (2010) montrent que la méthode basée sur l'imputation multiple est globalement meilleure.

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 des 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 $m = 5$.

La méthode retenue dans cette étude pour combler les données manquantes, notamment celle de l'imputation multiple, utilise un algorithme d'imputation multiple basé sur les chaînes de Markov (Van Buuren, 2007). L'algorithme fonctionne de la manière suivante :

[1] Confère Schafer et Graham (2002) pour plus de détails.

- 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 demande l'utilisateur. La valeur moyenne de chaque donnée manquante imputée est utilisée.

III. Méthode de simulation

Afin de tester empiriquement la particularité de la méthode d'imputation multiple, nous procédons à une simulation à partir d'une série de donnée observée. Les données utilisées font référence au PIB par habitant du Congo sur la période 1980-2012. La procédure est la suivante : Dans la série de données, nous supprimons les observations sur la période 1995-2000. Puis par la suite, nous utilisons la méthode d'imputation multiple pour voir si les données imputées se rapprochent des données observées. A ce sujet, Niass et al (2013) montre que pour une dispersion de moins de 20 % entre les écarts-types sur données observées et imputées, la méthode d'imputation multiple produit des estimations pratiquement sans biais. Le logiciel utilisé est XLSTAT v5.03 fonctionnant sous EXCEL.

IV. Résultats

La série de données se présente dans le tableau suivant. La simulation porte sur la période 1995-2000.

Tableau 1 : Résultat de la simulation par le logiciel XLSTAT

Année	Données observées	Données imputées
1980	546,143537	
1981	463,328067	
1982	491,637064	
1983	386,329085	
1984	268,699759	
1985	239,516413	
1986	262,224838	

Données Simulées

1987	241,379727	
1988	271,160597	
1989	267,489569	
1990	267,820289	
1991	250,681572	
1992	217,492852	
1993	272,759308	
1994	142,965455	
1995	**134,327546**	**276,5244**
1996	**133,838283**	**301,9797**
1997	**138,181988**	**242,8173**
1998	**138,244243**	**280,5491**
1999	**102,666198**	**327,1441**
2000	**406,567703**	**282,6178**
2001	154,424858	
2002	176,263601	
2003	175,341546	
2004	196,189983	
2005	221,449689	
2006	257,17378	
2007	286,14472	
2008	326,527953	
2009	301,93252	
2010	330,003266	
2011	372,768149	
2012	417,818378	
2013	453,673871	

Source : Auteur à partir des données de la Banque mondiale (WDI, 2014)

- **Statistiques descriptives (Avant traitement) :**

Observations	Obs. avec données manquantes	Obs. sans données manquantes	Min	Max	Moyenne	Ecart-type
33	6	27	142,9655	491,6371	**285,67**	**93,6019**

Source : Auteur à partir de XLSTAT

- **Statistiques descriptives (Après traitement) :**

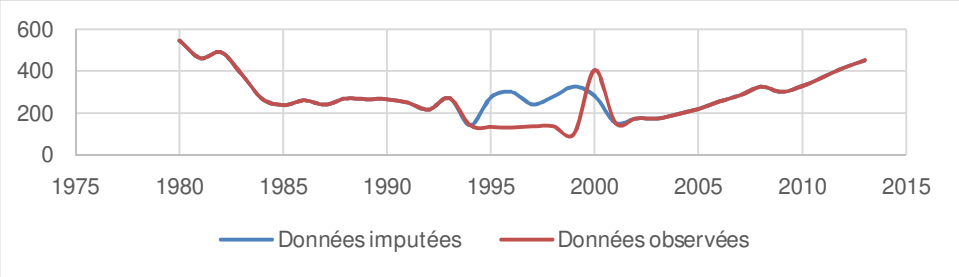
Observations	Obs. avec données manquantes	Obs. sans données manquantes	Min	Max	Moyenne	Ecart-type
33	0	33	142,9655	491,6371	**285,60**	**85,098**

Source : Auteur à partir de XLSTAT

A la fin de la simulation (Apres traitement), on observe que la moyenne des observations varie de 7 dixièmes. L'écart-type passe de 93.60 à 85.09. Ainsi, on observe une variation de 9.09 % entre l'écart-type observé et l'écart-type sur données imputées. Ce résultat

est conforme avec celui obtenu par Niass et al (2013) car la dispersion entre les écarts-types est inférieure à 20 %. Par ailleurs, il est important de noter le fait que la moyenne de la série de données simulée (285.60) diffère très peu de celle des données observées (285.67). Globalement, la méthode d'imputation multiple montre qu'il n'excite pas de différence significative entre les données observées et les données imputées. Graphiquement, cela se traduit par la figure suivante.

Figure 1 : Analyse de la dispersion entre données observées et données imputées



Source : Auteur à partir des données de la Banque mondiale (WDI, 2014) et XLSTAT

V. Conclusion

En définitive, on observe une très proche similitude entre les données observées et les données imputées car la dispersion au niveau de l'écart-type est acceptable, car inférieure à la limite de 20 % définit par Niass et al (2013). La méthode d'imputation multiple apparaît donc comme une solution pour faire face aux problèmes de données manquantes dans les séries de données secondaires en sciences économiques.

Bibliographie :

Glasson-Cicognani et Berchtold, (2010), « Imputation des données manquantes : comparaison de différentes approches », 42èmes journées de Statistique, Marseille, France.

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.

Rubin, (1987), « Multiple imputation for nonresponse in surveys », New York: John Wiley.

Schafer et Graham, (2002), « Missing Data: our view of the state of the art », *Psychological Methods*, n°7, vol 2, pp.147-177.

Schafer, J. L. (1997), « Analysis of Incomplete Multivariate Data ». London: Chapman and Hall.

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.