Comparaison de méthode d'imputation multiple pour données multi-niveaux systématiquement et sporadiquement manquantes

V. Audigier [123], *I. White* [4], *S. Jolani* [5], *T. Debray* [6], *M. Quartagno*[7], *S. van Buuren*[8] *M. Resche-Rigon* [123]

Courriel : vincent.audigier@inserm.fr

Résumé La méta-analyse sur données individuelles est souvent considérée comme étant la méthode de référence pour les revues systématiques. Le principe de la méta-analyse est de considérer plusieurs études, partageant une même variable à expliquer, pour obtenir une inférence, par rapport à cette variable, plus précise que celle qui aurait pu être obtenue avec une seule de ces études. Cependant, les études diffèrent généralement par le recueil de données effectué, il est alors fréquent que les facteurs de confusions considérés ne soient pas les mêmes d'une étude à l'autre. En conséquence, par la concaténation de ces études, des données systématiquement manquantes, *i.e.* manquantes pour tous les individus d'une même étude, peuvent être introduites. De plus, au sein d'une même étude, il est fréquent que certaines variables soient incomplètes pour certains individus. Ces données sont quant à elles qualifiées de sporadiquement manquantes.

L'imputation multiple est un moyen classique de gérer le problème des données manquantes. Le modèle d'imputation utilisé peut être un modèle joint explicite, spécifiant une distribution pour l'ensemble des variables, ou il peut être implicite, spécifiant seulement les distributions conditionnelles de chaque variable. On parle respectivement d'imputation par modèle joint ou d'imputation par équations enchaînées. Le choix d'un modèle d'imputation est une étape majeure, mais il peut constituer une tâche difficile à réaliser *a priori*.

Nous étudions ici les méthodes d'imputation multiple permettant de gérer les données systématiquement manquantes dans des données multi-niveaux, telles que les méta-analyses, dans le contexte de données mixtes (quantitatives et binaires). L'étude comparative porte sur : l'imputation par modèle joint pour des données en clusters, proposée par Quartagno and Carpenter (2015), l'imputation par équations enchaînées utilisant des modèles mixtes généralisés, proposée par Jolani et al. (2015), l'imputation par équations enchaînées utilisant une procédure d'estimation par méta-analyse en deux étapes (Resche-Rigon

[1]INSERM, UMR 1153, Equipe ECSTRA, Hôpital Saint-Louis, Paris

[2]Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, Paris, France

[3]Université Paris Diderot - Paris 7, Sorbonne Paris Cité, UMR-S 1153, Paris, France

[4]MRC Biostatistics Unit, Cambridge Institute of Public Health, U.K

[5]Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht,Pays-Bas

[6]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Pays-Bas

[7]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Londre, Royaume-Unis

[8]Department of Statistics, TNO Prevention and Health, Leiden, Pays-Bas

and White, 2016) et l'imputation par équations enchaînées par forêts aléatoires (Doove et al., 2014).

Les avantages et inconvénients de chaque méthode sont tout d'abord discutés au travers d'une étude par simulation et des recommandations pratiques en sont déduites. Ensuite, les méthodes d'imputation multiples sont appliquées aux données GREAT (Great Network, 2013), une méta-analyse sur données individuelles dans le cadre de maladies cardio-vasculaires, constituées de 12 études observationnelles contenant des données systématiquement et sporadiquement manquantes.

Mots-clés : Méta-analyse sur données individuelles, Modèles mixtes, Données multi-niveaux, Données manquantes, Imputation multiple, Equations enchainées

References

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

Great Network (2013). Managing acute heart failure in the ed - case studies from the acute heart failure academy. http://www.greatnetwork.org.

Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S., and Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine*, 34(11):1841–1863.

Quartagno, M. and Carpenter, J. . (2015). Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Stat Med*.

Resche-Rigon, M. and White, I. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *smmr*. in revision.