Estimation des données manquantes en morphométrie : quelle limite choisir ?

J. Clavel [a], G. Merceron [b], G. Escarguel [a]

[a] Laboratoire de Géologie de Lyon : Terre, Planètes, Environnements
UMR 5276 CNRS, ENS Lyon & Université Lyon 1
Campus de la Doua, 2 rue Raphaël Dubois, 69622 Villeurbanne.
Julien.clavel@univ-lyon1.fr, gilles.escarguel@univ-lyon1.fr

[b] Institut International de Paléoprimatologie Paléontologie Humaine : Evolution et Paléoenvironnements
UMR 7262 CNRS & Université de Poitiers
Bat. 8, 5 rue Albert Turpain, 86022
Gildas.merceron@univ-poitiers.fr

Mots clefs : Imputations multiples, données manquantes, morphométrie, simulations, ordinations, superimpositions procrustes.

Les estimations des dynamiques évolutives et de la diversité passée sont essentiellement basées sur l'étude de la variation morphologique de spécimens fossiles. Malheureusement, les restes fossiles sur lesquels de telles estimations doivent être effectuées sont souvent altérés par les processus post-mortem ou taphonomiques. Une telle perte d'information conduit souvent au retrait de certains spécimens dans les analyses multivariées et exclu de possibles comparaisons contrôlées statistiquement. Afin de contourner ce problème de données manquantes, des méthodes d'imputations sont souvent utilisées pour directement remplacer les valeurs manquantes par des estimations établies sur la partie non altérée du jeu de données. Cependant la proportion de valeurs manquantes dans un jeu de données peut conduire à des estimations significativement biaisées.

Ces dernières années, plusieurs seuils empiriques représentant la proportion maximale de données manquantes, que l'on peut considérer comme acceptable pour l'utilisation de techniques d'imputations, ont été proposés dans la littérature. D'un autre côté, certaines études ont critiqués ces seuils car ils sont souvent spécifiques aux jeux de données utilisés dans les simulations, à la distribution des valeurs manquantes, ou encore aux méthodes d'imputations utilisées, et ne sont donc en aucun cas généralisable.

Alternativement, des méthodes d'imputation multiples (MI) ont été développées pour considérer explicitement l'erreur associée aux estimations. Ces méthodes permettent

d’imputer *m* (>1) fois le même jeu de données via des processus de Monte Carlo. La variabilité obtenue sur ces *m* (>1) tableaux imputés, permet d’évaluer l’erreur associée aux estimations des valeurs manquantes.

Dans cette étude, nous évaluons les performances relatives de sept techniques d’imputations multiples disponibles sur R. Chacune des simulations ont étés effectuées sur un jeu de données morphométriques dégradé artificiellement suivant trois types de biais (aléatoires, anatomiques, et taxinomiques). Les simulations révèlent que les algorithmes FCS (Fully Conditional Specification) et EM (Expectation-Maximization) des packages MICE et Amelia II, produisent les meilleurs compromis statistiques en termes d’erreur systématique et de probabilité de recouvrement pour l’intervalle de confiance à 95%. De plus, les techniques d’imputations multiples apparaissent remarquablement robustes aux transgressions des conditions statistiques qui leur sont propres, comme par exemple la distribution non-aléatoire des données manquantes dans les jeux de données. Ces résultats montrent que les différences observées entre les types de distributions (aléatoire, taxinomiques, anatomiques) sont plus faibles qu’entre les méthodes d’imputations multiples elles-mêmes. Sur la base de ces résultats, plutôt que de proposer une valeur ou un ensemble de valeurs seuils, nous développons une approche qui combine l’utilisation de ces imputations multiples avec la super-imposition Procruste des résultats d’analyses en composantes principales. L’erreur associée à des individus pour lesquels certaines valeurs manquantes ont été imputées, peut être ainsi directement visualisée dans un espace ordonné.