Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes

Alyssa Imbert, Nathalie Vialaneix

Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes

Title: Exploring, handling, imputing and evaluating missing data in statistical analyses: a review of existing approaches

Alyssa Imbert[1] et Nathalie Vialaneix[1]

Résumé : Le problème des données manquantes est intimement lié à l'analyse statistique, au fait de collecter et préparer les données pour l'analyse statistique. Nous proposons ici une revue des approches permettant de diagnostiquer et d'imputer les données manquantes, ainsi que de contrôler les conséquences de l'imputation dans les analyses statistiques. Nous décrivons également les implémentations disponibles, dans des packages R, des diverses approches décrites.

Abstract: Missing data is strongly connected to statistics that is concerned with the collect and pre-processing of data. In this article, we review the different methods that can be used to diagnose and impute missing data. We also present approaches aiming at evaluating the impact of imputation on subsequent analyses. Finally, we describe available implementations, in R packages, of the presented methods.

Mots-clés : données manquantes, imputation
Keywords: missing data, imputation
Classification AMS 2000 : 62-07, 62Nxx

1. Introduction

L'apparition de données manquantes est intimement liée à l'analyse statistique, au fait de collecter et préparer les données pour l'analyse statistique et elle a des origines multiples. Les données manquantes peuvent être la conséquence de non réponses (en sondages), de problèmes expérimentaux divers (en biologie), d'une mauvaise saisie de l'information ou de données aberrantes que l'on supprime après la première analyse exploratoire, ... La donnée manquante est parfois partielle [2] (pour un individu donné, seules quelques valeurs sont manquantes) ou bien totale [3] (toutes les variables d'un individu donné sont non observées).

L'objectif des méthodes permettant de traiter les données manquantes est multiple : il peut s'agir d'estimer les valeurs manquantes elles-mêmes, pour reconstituer une vision réaliste des données. Toutefois, dans de nombreux cas, les données contenant des valeurs manquantes sont utilisées pour des analyses statistiques de natures diverses : estimation d'un paramètre de la population dont sont tirées les données, analyses exploratoires (types ACP), modèles prédictifs...

[1] MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France.
E-mail : alyssa.imbert@inra.fr and E-mail : nathalie.vialaneix@inra.fr

[2] *item non-response* en anglais.

[3] *unit non-response* en anglais.

 ISSN: 2102-6238

Dans ces divers cas, la manière d'aborder les données manquantes, en utilisant uniquement l'information disponible ou bien en tentant de reconstituer les données manquantes (imputation), doit tenir compte de l'objectif lui-même, afin de limiter la perte de précision dans les méthodes de prédiction ou bien les biais d'estimation dans les méthodes d'inférence.

Schafer (1997), Allison (2001), Little et Rubin (2002), Schafer et Graham (2002), Gelman et Hill (2007), Baraldi et Enders (2010), van Buuren (2012) et Carpenter et Kenward (2013) constituent les principaux ouvrages de référence sur les données manquantes. L'objectif de cet article est de proposer au lecteur une vision générale des divers problèmes liés aux données manquantes et des principales stratégies qui peuvent être mises en œuvre pour tenir compte de leur présence dans les analyses statistiques.

L'article est organisé comme suit : la section d'introduction présente les notations et la typologie usuelle des données manquantes. La section 2 présente les approches utilisant uniquement les données observées (c'est-à-dire, les méthodes qui ne recourent pas à l'imputation des données manquantes). La section 3 présente les approches de modélisation jointe principalement utilisées dans les problèmes d'inférence statistique. La section 4 présente les méthodes d'imputation simple qui permettent d'obtenir un tableau de données complet. La section 5, quant à elle, décrit les diverses approches permettant d'évaluer la qualité de l'imputation ou l'incertitude liée à l'imputation ou à la présence de valeurs manquantes dans les résultats de l'analyse statistique. Enfin, la section 6 décrit les approches plus spécifiquement dédiées au cas le plus complexe, celui dans lequel les données sont manquantes MNAR (c'est-à-dire, manquantes de manière non aléatoire). En complément, compte tenu de l'impact croissant de l'utilisation du logiciel R dans l'analyse statistique, nous nous attacherons, quand cela est possible, à présenter des packages dans lesquels les diverses méthodes décrites dans cette revue sont implémentées.

1.1. Notations

Soit un vecteur $Y = (Y_1, ..., Y_p)$ de p variables aléatoires numériques ou catégorielles. On notera y_{ij} l'observation de la variable Y_j pour un individu $i \in \{1, \dots, n\}$, $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$ le vecteur des observations des p variables de Y et $\mathbf{Y}$ la matrice des observations $(y_{ij})_{i=1,\dots,n,\, j=1,\dots,p}$ dont les lignes sont des observations i.i.d. de Y. Pour simplifier, on confondra la notation de la variable aléatoire Y_j et de son observation $Y_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix}$ sur les n individus.

On définit aussi la *matrice indicatrice des valeurs manquantes*, $\mathbf{R}$, dont les valeurs, $(r_{ij})_{i=1,\dots,n,\, j=1,\dots,p}$, sont :

$$r_{ij} = \begin{cases} 1 & \text{si } y_{ij} \text{ est observée} \\ 0 & \text{sinon} \end{cases}$$

et on note R la variable aléatoire associée. De manière similaire, Y_{obs} et Y_{miss} correspondent (respectivement) aux parties observées et manquantes de Y de telle sorte que $Y = RY_{\text{obs}} + (1 - R)Y_{\text{miss}}$.

Le *mécanisme de génération des données manquantes* est défini comme étant la distribution conditionnelle de R sachant Y, $f(R \mid Y)$ (Little et Rubin, 2002). Ce mécanisme peut éventuellement

 ISSN: 2102-6238

dépendre de paramètres, notés ψ. Également, dans certains cas, des covariables $(X_j)_{j=1,\ldots,q}$ sont complètement observées sur tous les individus (on note alors x_{ij} l'observation de la covariable j pour l'individu i et X les variables aléatoires correspondantes). Dans ces cas plus complexes, le mécanisme de génération des données manquantes est alors noté $f(R \mid Y, X; \psi)$ ou $f(R|Y; \psi)$.

Enfin, quelques-unes des notions de cette revue seront illustrées sur des données de questionnaire, présentes dans le package R **naniar** et qui concernent une enquête annuelle produite en 2009 par le Behavioral Risk Factor Surveillance System (BRFSS)[4] destinée à évaluer les comportements à risque dans la population adulte aux États-Unis. Le jeu de données contient la mesure de 34 variables (État de résidence, sexe, âge, statut marital, grossesse, tabagisme...) pour 245 adultes de 18 ans et plus. Ces données contiennent un total de 1186 valeurs manquantes.

1.2. Répartition des données manquantes

Pour décider de l'approche la plus judicieuse pour prendre en compte les valeurs manquantes dans l'analyse (suppression d'individus ou de variables, correction manuelle, imputation par prédiction, ...), il est recommandé de réaliser une analyse exploratoire permettant de comprendre la distribution des valeurs manquantes dans le jeu de données. Little et Rubin (2002) définissent trois types de répartition des données manquantes, illustrés par la figure 1 :

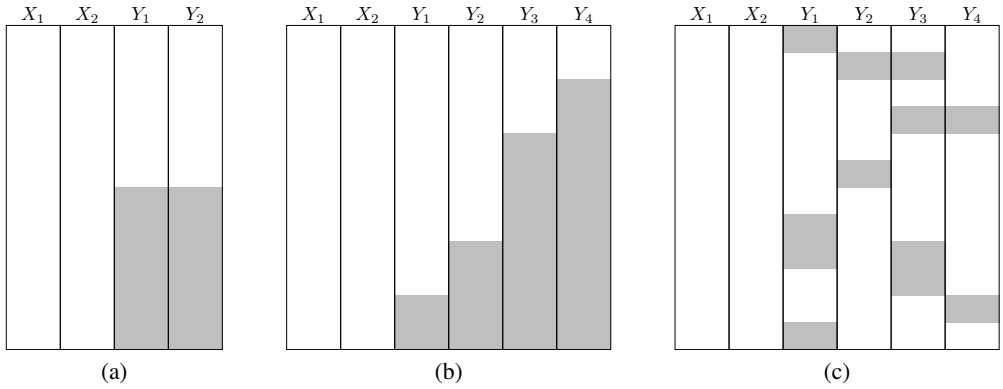

FIGURE 1. *Répartition des données manquantes, (a) univariée, (b) monotone et (c) sans structure. Les zones grisées indiquent la position des données manquantes.*

— la structure des valeurs manquantes est *univariée* (figure 1(a)) si les mêmes individus ont des valeurs manquantes pour les mêmes $d < p$ variables ;

— les valeurs manquantes sont *monotones* (figure 1(b)) si les variables peuvent être ordonnées de telle sorte que, lorsque l'observation y_{ij} est manquante pour la variable Y_j, alors toutes les variables suivantes pour ce même individu, $\{y_{ik}\}_{k>j}$, sont aussi manquantes. Ce cas est fréquemment rencontré dans les études longitudinales, particulièrement en épidémiologie (il peut correspondre, par exemple, à la sortie de l'étude d'un individu : on parle alors de données censurées) ;

[4] https://www.cdc.gov/brfss/annual_data/annual_2009.htm

 ISSN: 2102-6238

— les valeurs manquantes sont *sans structure* (voir figure 1(c)), si elles sont réparties sans structure particulière dans le jeu de données.

En outre, la quantité de données manquantes peut être définie de manière variée selon que l'on considère une proportion de manquants par rapport aux individus (lignes), aux variables (colonnes) ou bien aux valeurs elles-mêmes (entrées du tableau).

Comme souligné par Templ *et al.* (2012) et Tierney *et al.* (2015), comprendre la répartition des valeurs manquantes dans le jeu de données permet d'adapter la stratégie de traitement de celles-ci, qu'il s'agisse d'exclure des variables ou individus (qui contiennent une fréquence de manquants trop importante), de collecter de nouvelles données, d'estimer ou de remplacer les valeurs manquantes (imputation). Pour aborder cette question, le package R **mi** (Su *et al.*, 2011) identifie les motifs identiques de valeurs manquantes entre paires de variables à la création du tableau de données avec la fonction `missing_data.frame` (voir figure 2).

```
NOTE: The following pairs of variables appear to have the same missingness pattern.
 Please verify whether they are in fact logically distinct variables.
      [,1]           [,2]
 [1,] "diet_fruit"   "diet_salad"
 [2,] "diet_fruit"   "diet_potato"
 [3,] "diet_fruit"   "diet_carrot"
 [4,] "diet_fruit"   "diet_vegetable"
 [5,] "diet_fruit"   "diet_juice"
 [6,] "diet_salad"   "diet_potato"
 [7,] "diet_salad"   "diet_carrot"
 [8,] "diet_salad"   "diet_vegetable"
 [9,] "diet_salad"   "diet_juice"
[10,] "diet_potato"  "diet_carrot"
[11,] "diet_potato"  "diet_vegetable"
[12,] "diet_potato"  "diet_juice"
[13,] "diet_carrot"  "diet_vegetable"
[14,] "diet_carrot"  "diet_juice"
[15,] "diet_vegetable" "diet_juice"
```

FIGURE 2. *Message concernant les motifs de valeurs manquantes identiques entre diverses variables tel que fourni par le package* **mi**.

Une autre manière standard d'explorer la répartition et la structure des valeurs manquantes est d'avoir recours à des graphiques diagnostiques, qui peuvent s'avérer particulièrement efficaces en raison de la capacité de l'œil humain à détecter facilement des motifs (Tierney *et al.*, 2015). Le package R **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016) permet ce type d'analyse exploratoire et peut aider à identifier le mécanisme de génération des données manquantes (voir section suivante) ainsi qu'à déceler des anomalies ou des erreurs dans les données imputées (voir section 5.1). **VIM** contient, en outre, quelques méthodes d'imputation des données que nous décrirons dans les sections suivantes. Enfin, **VIM** peut être facilement utilisé au travers de l'interface graphique **VIMGUI**. Sur l'exemple décrit brièvement en section 1.1, la figure 3 montre le type de graphiques disponibles dans ce package : la répartition du nombre de valeurs manquantes par variable est visualisée par un diagramme en barres, les motifs et fréquences de ces motifs sont visualisés par un diagramme en grille et la relation entre les niveaux de valeurs

 ISSN: 2102-6238

des variables et les valeurs manquantes est disponible sous la forme d'un graphique en matrice (ordonné, dans cet exemple, selon la variable « age », en troisième colonne).

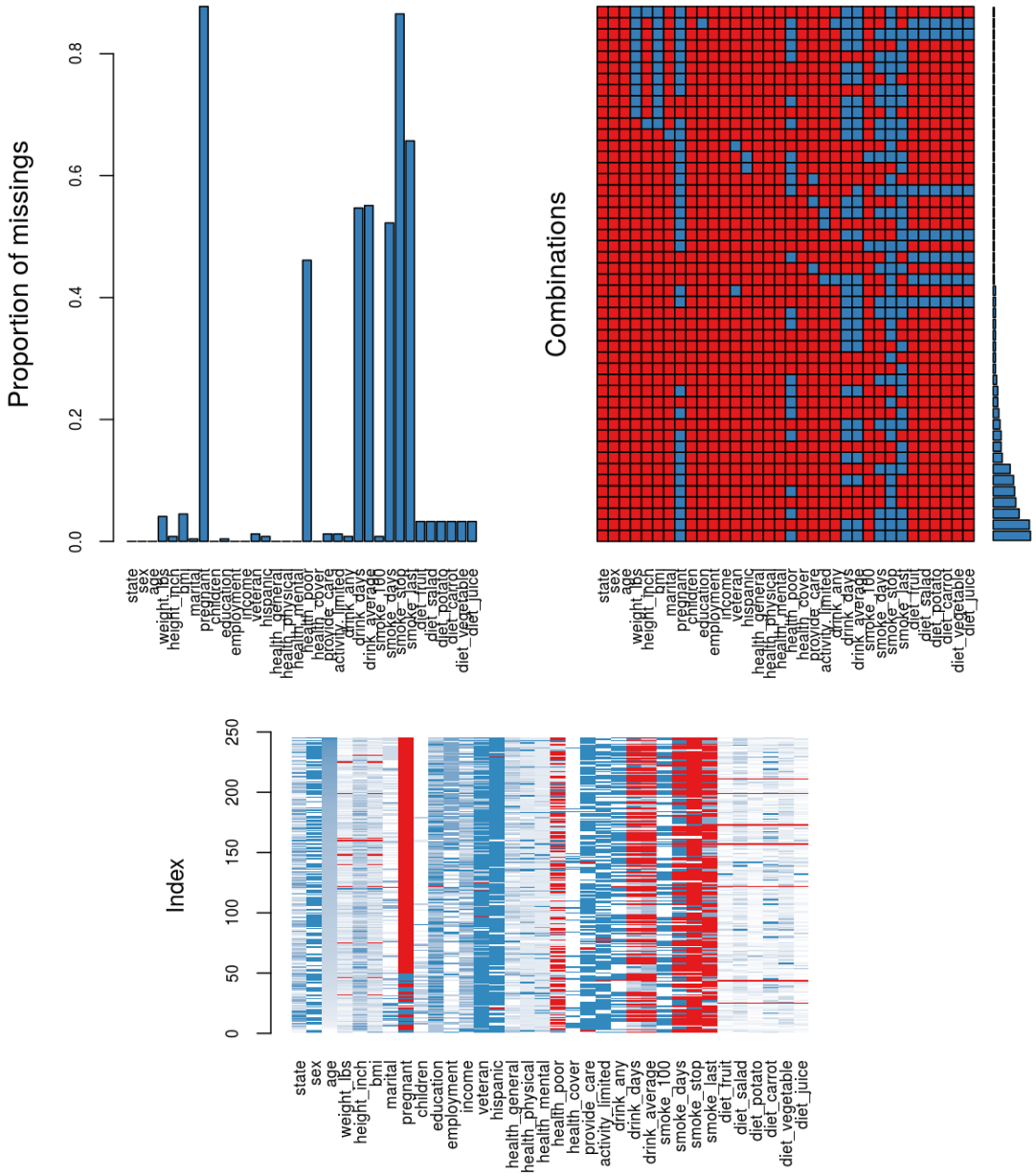

FIGURE 3. *Graphiques de visualisation de la distribution des valeurs manquantes disponibles dans* **VIM**. *En haut à gauche : diagramme en barres du nombre de valeurs manquantes par variable. En haut à droite : diagramme en grille des motifs et fréquences de ces motifs. En bas : diagramme de la répartition des valeurs manquantes (en rouge) dans la distribution des valeurs de chaque variable (en niveaux de bleu) dans lequel les individus sont ordonnés selon la valeur de la variable « age ».*

 ISSN: 2102-6238

De manière similaire, le package **naniar** est dédié à la manipulation et la visualisation des données manquantes selon les principes développés dans la collection de packages « tidyverse »[5]. Parmi les graphiques disponibles dans ce package, on trouve un graphique en matrice permettant de visualiser la répartition des manquants et très similaire à celui du package **visdat** de visualisation de données. On trouve également un graphique en bâtons permettant de visualiser le nombre de valeurs manquantes par variable.

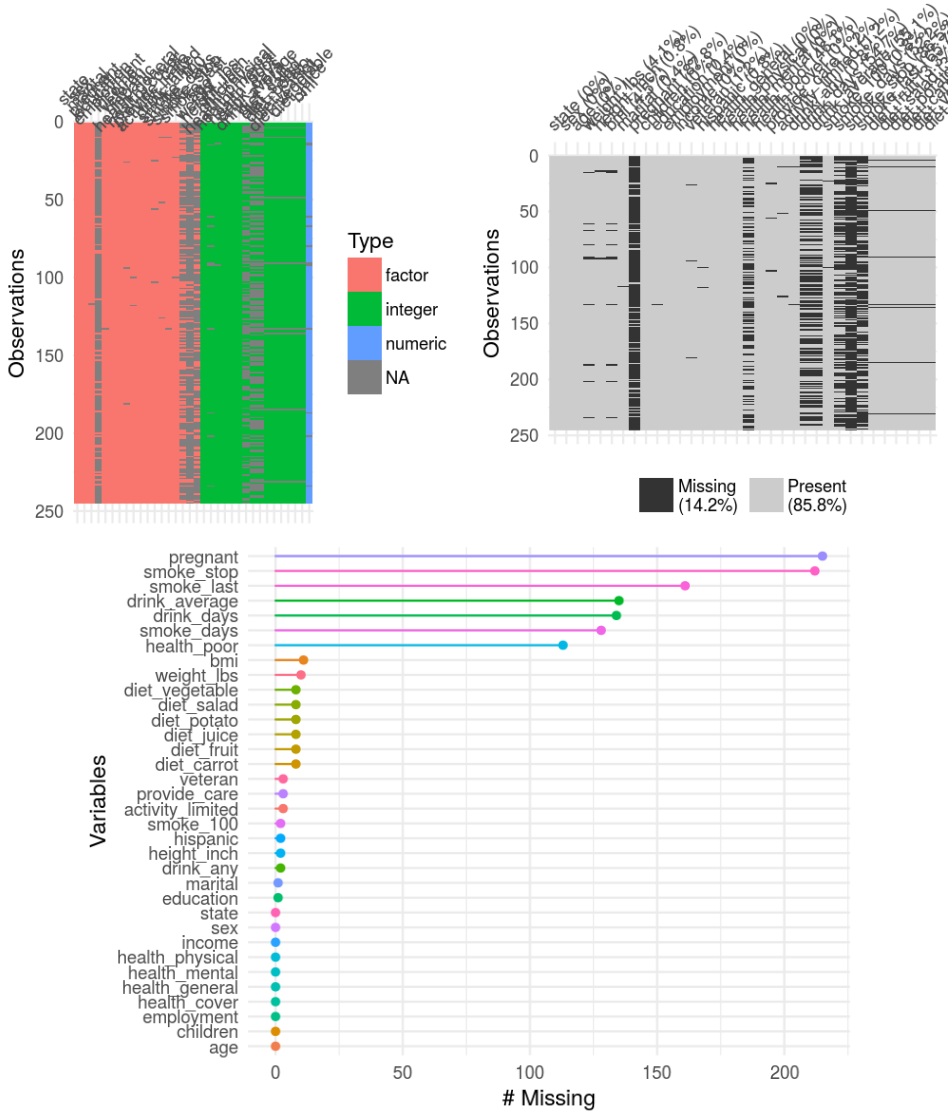

FIGURE 4. *Graphiques de visualisation de la distribution des valeurs manquantes disponibles dans* **visdat** *(en haut à droite) et dans* **naniar**. *En haut : diagrammes de la répartition des valeurs manquantes. En bas : diagramme en bâtons du nombre de manquants par variable.*

[5] https://www.tidyverse.org/

 ISSN: 2102-6238

Dans l'exemple des figures 3 et 4, on peut, par exemple, identifier de manière immédiate que, si la plupart des variables sont renseignées pour presque tous les individus, quelques variables ont une forte proportion de valeurs manquantes (parmi lesquelles la variable indiquant si la personne est enceinte, « pregnant » ou la variable précisant la fréquence à laquelle la personne fume, « smoke_day »). Ces variables sont souvent manquantes simultanément. On observe également un groupe de variables qui sont manquantes de manière simultanée sur la droite des graphiques et qui correspondent aux variables décrivant les habitudes alimentaires, « diet_... », comme déjà identifié par le message de la figure 2 (ce sous-groupe présente donc une structure univariée). De même, le sous-groupe relatif aux habitudes de consommation d'alcool, « drink_... » a une structure monotone. Enfin, les valeurs manquantes de la variable « pregnant » sont clairement liées à la variable « age » (les personnes les plus âgées de l'échantillon ayant systématiquement un statut manquant pour la variable « pregnant »). Comme nous le verrons dans la section suivante, ces observations simples donnent des indices sur la nature du mécanisme des données manquantes et orientent l'utilisateur vers des manières de prendre en charge l'information manquante.

Enfin, notons que, si les deux packages précédents proposent des visualisations statiques de la répartition des données manquantes, Templ *et al.* (2012) soulignent le très grand intérêt pratique, pour déceler des problèmes de collectes de données ou des motifs dans la distributions des valeurs manquantes, des représentations interactives. Le logiciel GGobi [6] (Cook et Swayne, 2007), accessible dans R via le package **rggobi**, permet une telle visualisation. Des exemples d'utilisation des fonctionnalités d'interactivité, sous forme de vidéos, sont disponibles sur le site web associé au livre `http://www.ggobi.org/book/`. Elles illustrent, par exemple, comment le fait de pouvoir lier des graphiques différents à la souris permet d'explorer la distribution des valeurs manquantes ou bien comment visualiser les effets de l'imputation sur la distribution des variables.

1.3. Mécanisme de génération des données manquantes

Au-delà du simple aspect descriptif de la répartition des données manquantes, il est souvent nécessaire d'appréhender la loi de probabilité à l'origine des données manquantes (càd le mécanisme de génération des données manquantes). La connaissance de ce mécanisme (ou plutôt de son type) est, en effet, une hypothèse standard des garanties théoriques qui existent pour certaines méthodes qui prennent en compte les valeurs manquantes, comme nous le verrons dans les sections suivantes.

1.3.1. Typologie générale

Little et Rubin (2002) définissent une typologie générale des données manquantes en trois catégories qui dépendent de la relation statistique entre les données et le mécanisme de génération des données manquantes. Les définitions suivantes sont données dans le cas où il n'y a pas de covariables complètement observées, X, pour alléger les notations, mais s'étendent de manière triviale au cas où elles sont présentes.

[6] `http://www.ggobi.org/`

 ISSN: 2102-6238

— **Données manquantes complètement aléatoirement ou MCAR** [7]
Les données sont *manquantes complètement aléatoirement* si la probabilité d'absence est la même pour toutes les observations. Cette probabilité ne dépend que des paramètres extérieurs indépendants de cette variable. De manière formelle, ce cas est défini par :

$$f(R|Y,X;\psi) = f(R;\psi).$$

Dans ce cas-ci, les données manquantes sont nécessairement sans structure. Un exemple typique de données MCAR est le cas où une personne oublie par accident de répondre à une question lors d'une enquête. Les données manquantes des variables présentes au centre du tableau de la figure 4 (en haut à droite) pourraient être de ce type (par exemple, les variables niveau d'éducation, « education » et statut vis-à-vis du service militaire « veteran ») : elles présentent peu de manquants, pour lesquels on ne décèle, de manière visible, aucune relation avec les valeurs ou le statut des autres variables.

— **Données manquantes aléatoirement ou MAR** [8]
Le cas des données manquantes complètement aléatoirement est rare : si la probabilité d'absence est liée à une ou plusieurs variables observées, les données manquantes sont dites *données manquantes aléatoirement*. De manière formelle, ce cas est défini par :

$$f(R|Y,X;\psi) = f(R|Y_{\text{obs}},X;\psi).$$

Dans l'exemple introduit dans la figure 3 (bas), le couple $(\text{age},\text{pregnant})$ pourrait constituer un exemple de données MAR : les valeurs manquantes de la variable « pregnant » sont liées de manière visible à la variable « age » de l'individu, qui est complètement observée.

— **Données manquantes non aléatoirement ou MNAR** [9]
Enfin, le dernier cas est de données *manquantes de façon non aléatoire* se présente lorsque la probabilité d'absence d'une variable dépend de la variable elle-même ou d'autres variables non observées. De manière formelle, ce cas est défini par :

$$f(R|Y,X;\psi) = f(R|Y_{\text{obs}},Y_{\text{miss}},X;\psi).$$

Ce type de données manquantes est plus complexe à traiter. Il peut être abordé par analyse de sensibilité (voir section 6 pour des détails sur le traitement spécifique de ce type de données manquantes). Un exemple typique de ce type de données manquantes est le cas de questions sensibles dans un questionnaire où le niveau de non-réponse dépend de la réponse elle-même. Dans les données de l'exemple précédent, on peut suspecter, par exemple, une plus grande propension des gros fumeurs ou des gros consommateurs d'alcool à ne pas répondre (variables « smoke_... » et « drink_... »).

Notons que les exemples donnés ne sont fondés que sur des hypothèses liées à l'observation de la distribution des variables. Dans le cas des variables $(\text{age},\text{pregnant})$, on peut aussi imaginer que les données sont MNAR si le statut de la variable « pregnant » est lui-même lié à la présence

[7] *Missing Completely At Random* en anglais.
[8] *Missing At Random* en anglais.
[9] *Missing Not At Random*

 ISSN: 2102-6238

de manquants sur cette variable (les valeurs négatives de « pregnant » étant, par exemple, plus fréquemment non collectées) et que l'observation d'un lien entre âge et statut manquant de « pregnant » est lié à une dépendance (qui existe de manière évidente) entre ces deux variables. Même dans le cas de la variable « education », il est impossible de distinguer une potentielle absence MCAR du cas où toutes les valeurs manquantes de cette variable correspondent, par exemple, à une même modalité de la variable (« *n'est jamais allé à l'école ou seulement à l'école maternelle* », par exemple), qui correspondrait à un cas MNAR.

1.3.2. Pourquoi s'intéresser aux valeurs manquantes ?

Une approche naïve, en présence de données manquantes, est d'analyser les données en utilisant uniquement les observations disponibles. Prenons, par exemple, le cas simple de l'inférence statistique, dans lequel on chercherait à estimer l'espérance de Y_1, $\mu_1 = \mathbb{E}(Y_1)$. Dans ce cas, l'estimateur habituel de μ_1 est $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}$ qui est sans biais ($\mathbb{E}(\hat{\mu}_1) = \mu_1$) mais n'est pas nécessairement observé (si certaines valeurs de la variable Y_1 sont manquantes). Remplacer cet estimateur par $\widetilde{\mu_1} = \frac{1}{n_1}\sum_{i=1}^{n} r_{i1} y_{i1}$ avec $n_1 = \sum_{i=1}^{n} r_{i1}$ le nombre de valeurs observées pour Y_1 a des conséquences variées selon le type de mécanisme des données manquantes :

— si les données manquantes sont MCAR, R et Y sont indépendantes et $\widetilde{\mu}_1$ est donc aussi un estimateur sans biais de μ_1. Toutefois, cet estimateur est obtenu avec $n_1 < n$ observations et il en résulte une perte de précision de l'intervalle de confiance autour de μ_1 ou (dans le cas de tests statistiques) une perte de puissance ;

— si les données manquantes sont MAR ou MNAR, R et Y ne sont plus indépendantes. Cela peut être le cas, par exemple, si l'observation de Y_1 est liée à la variable Y_2 comme suit :

$$R_1 = \begin{cases} 0 & \text{si } Y_2 \leq a \\ 1 & \text{sinon.} \end{cases}$$

pour un $a \in \mathbb{R}$, fixé. Dans ce cas,

$$\mathbb{E}(\widetilde{\mu}_1) = \mathbb{E}\left(Y_1 \mathbf{1}_{\{Y_2 > a\}}\right)$$

ce qui résulte en un biais de $\mathbb{E}\left(Y_1 \mathbf{1}_{\{Y_2>a\}}\right)$ dans l'estimation de μ_1. La différence entre le cas MAR et le cas MNAR réside dans la dépendance de R aux données non observées. Dans l'exemple précédent, si Y_2 est complètement observée, le mécanisme de génération des données est MAR.

1.3.3. Identification et utilisation de la typologie des valeurs manquantes

Il est donc important de connaître le type de données manquantes pour éviter les erreurs conduisant à des biais d'analyse dans leur prise en compte. Un test statistique, permettant de tester l'hypothèse selon laquelle les données manquantes sont MCAR contre MAR, est décrit dans Little (1988). Il est fondé sur une statistique de test qui suit une loi du χ^2. Le test fait l'hypothèse d'une distribution gaussienne, $\mathcal{N}(\mu, \Sigma)$ de Y et son principe est de grouper les individus en K sous-groupes de profils de valeurs manquantes distincts, $\mathscr{C}_k$ ($k = 1, \ldots, K$). Si les données

 ISSN: 2102-6238

manquantes sont MCAR, la statistique de test proposée et fondée sur le calcul des moyennes et variances conditionnelles aux K groupes de profils, a une distribution asymptotique suivant une loi du χ^2.

Ce test est implémenté dans la fonction `LittleMCAR` du package R **BaylorEdPsych**. S'il permet de tester l'hypothèse MCAR, il n'indique pas, en revanche, quelles variables ne sont pas MCAR. Comme le test est fondé sur une distribution asymptotique, son efficacité est fortement conditionnée à la taille de l'échantillon. Lorsque le nombre d'individus est trop faible ou que l'hypothèse de distribution gaussienne n'est pas réaliste, Jamshidian et Jalal (2010) ont proposé un test non paramétrique. Ce test est disponible dans le package R **missMech** (Jamshidian *et al.*, 2014). En revanche, comme souligné dans van Buuren (2012), il n'existe pas de test de l'hypothèse MAR contre l'hypothèse MNAR car l'information qui serait nécessaire pour réaliser un tel test est, justement, l'information manquante.

Par ailleurs, lorsque les données sont manquantes MAR, Rubin (1976) décrit les conditions minimales requises qui permettent d'ignorer le processus de génération des données manquantes dans l'inférence statistique (le processus de génération des données manquantes est alors dit « ignorable »). Pour cela, les données doivent être manquantes aléatoirement (cas MAR et MCAR) et les paramètres régissant le mécanisme de génération des données manquantes et des données doivent être « distinguables » : cela signifie que les paramètres du modèle de génération des données, ϕ, peuvent s'écrire $\phi = (\psi, \theta)$ où ψ désigne les paramètres qui régissent la distribution de R et où θ sont les paramètres qui régissent celle de Y. Ces paramètres sont distinguables lorsqu'ils vivent dans des espaces en produits cartésiens. Dans ce cas, lorsque les données manquantes sont MAR, il est possible de factoriser la densité des données observées de la façon suivante :

$$f(Y_{\text{obs}}, R; \theta, \psi) = f(R|Y_{\text{obs}}; \psi) \times \int f(Y; \theta) dY_{\text{miss}} = f(R|Y_{\text{obs}}; \psi) f(Y_{\text{obs}}; \theta), \tag{1}$$

et la vraisemblance des données observées est donc proportionnelle à la vraisemblance ignorant le mécanisme à l'origine des données manquantes $\mathscr{L}(\theta|Y_{\text{obs}})$:

$$\mathscr{L}(\theta, \psi|Y_{\text{obs}}, R) \propto \mathscr{L}(\theta|Y_{\text{obs}}).$$

En présence d'un mécanisme ignorable, Rubin (1976) montre qu'il n'est donc plus nécessaire de modéliser la distribution du mécanisme à l'origine des données manquantes pour estimer θ. Ce type d'approche est à la base des approches fondées sur la maximisation de la vraisemblance qui sont décrites dans la section 3.

Enfin, pour utiliser au mieux les informations sur la répartition des données manquantes et leur mécanisme de génération, un autre type d'approche est décrit dans Tierney *et al.* (2015). Les auteurs proposent l'utilisation d'arbres de décision pour déterminer quelles sont les variables permettant d'expliquer la présence de manquants. Ces approches peuvent permettre d'utiliser l'information obtenue sur la présence de valeurs manquantes pour mettre en œuvre des stratégies plus efficaces d'analyse des données manquantes (pondération des cas complets, comme décrit dans la section 2.1, modèles à effets aléatoires ou modèles de mélange de profil, comme décrits dans la section 6, par exemple). Ils montrent également que cette approche est performante, y compris dans le cas MCAR, sur un cas pratique de données médicales.

 ISSN: 2102-6238

2. Méthodes fondées uniquement sur les données observées

Une première approche pour pouvoir utiliser et analyser des données contenant des valeurs manquantes consiste à utiliser uniquement les observations disponibles. Ces approches présentent l'avantage de ne pas avoir recours à la spécification un modèle d'imputation (c'est-à-dire de remplacement des données) dont la qualité conditionne fortement les résultats de l'analyse. En revanche, elles sont souvent relativement inefficaces, biaisées ou induisent une perte de puissance importante.

Nous présentons, dans cette section, les approches possibles fondées sur ce paradigme, en décrivant les avantages et limites de celles-ci.

2.1. Analyse des cas complets et pondération

Une des premières possibilités pour traiter un jeu de données présentant des données manquantes est l'*analyse des cas complets* [10]. Cette méthode est la plus simple et la plus courante et c'est la méthode souvent implémentée par défaut dans les logiciels. Elle consiste à ne considérer que les individus pour lesquels toutes les données sont disponibles et donc à supprimer tout individu ayant au moins une valeur manquante.

Comme déjà souligné dans la section 1.3.2, l'analyse des cas complets est principalement valable dans le cas où les données manquantes sont MCAR et, même dans ce cas-ci, elle peut conduire à la suppression d'un nombre important d'individus (et donc à une perte de puissance dans les problèmes d'inférence). Graham (2009) déconseille l'utilisation de cette méthode lorsque les individus présentant des valeurs manquantes représentent plus de 5% de la population. En outre, dans le cas de la régression linéaire de Y_1 sur les autres variables $Y^{-1} = (Y_2, ..., Y_p)$, Seaman et White (2011) montrent que l'analyse des cas complet produit des estimations non biaisées du modèle linéaire uniquement dans le cas où R est indépendante de Y_1 sachant Y^{-1} : $\mathbb{P}(R = 1|Y) = \mathbb{P}(R = 1|Y^{-1})$.

Une approche pour réduire les biais d'estimation dans l'analyse des cas complets consiste à repondérer les cas complets disponibles : c'est la *pondération par probabilité inverse* (IPW) [11] (voir Seaman et White (2011) pour une revue de ce type d'approches). Généralement, la pondération est choisie comme l'inverse de la probabilité d'un individu d'être observé complètement, $\frac{1}{\eta_i}$. Les probabilités $(\eta_i)_{i=1,...,n}$ étant inconnues, elles sont estimées par un modèle de régression dont la variable à prédire est la variable R. Des équivalences asymptotiques ont été montrées dans Robins et Wang (2000) et Reilly et Pepe (1997) entre IPW et l'imputation multiple (voir section 5.2), dans le cas où Y est MAR et où les modèles d'imputation (pour l'imputation multiple) et de génération des données manquantes (IPW) sont correctement spécifiés. En pratique, Seaman et White (2011) notent que les études empiriques donnent, en général, un avantage d'efficacité à l'imputation multiple mais soulignent aussi quelques avantages de IPW : sa simplicité conceptuelle et de mise en œuvre, sa meilleure efficacité lorsque la distribution de $Y_{1,\text{obs}}$ est très différente de celle de $Y_{1,\text{miss}}$ ou lorsque les cas non complets tendent à avoir des valeurs manquantes pour beaucoup de (et non pour quelques) variables. Enfin, Seaman et White (2011) soulignent que IPW peut produire des poids très instables, lorsque l'estimation de η_i est faible.

[10] *listwise deletion* en anglais.
[11] *inverse probability weighting* en anglais.

 ISSN: 2102-6238

Les auteurs proposent quelques solutions pour aborder ce problème, comme la stabilisation des poids et l'augmentation de IPW (AIPW). Enfin, au niveau de l'implémentation, le package **ipw** (van der Wal et Geskus, 2011) permet de déterminer les probabilités inverses à utiliser pour l'imputation.

Conclusion et recommandations :

— *Avantages* : faciles à mettre en œuvre ; ne requièrent pas de spécifier un modèle d'imputation correct ;

— *Désavantages* : principalement valables dans les cas MCAR (analyse des cas complets) et MAR (IPW) ; requièrent que le nombre de cas complets corresponde à une proportion importante des données de départ ; en pratique souvent moins efficaces que l'imputation multiple.

2.2. Analyse des cas disponibles

Afin d'éviter la diminution trop importante du nombre d'individus dans l'analyse statistique, une alternative à l'analyse des cas complets est l'*analyse des cas disponibles* [12] (Allison (2001) et Pigott (2001)). Cette approche consiste à estimer différents aspects du problème avec différents sous-échantillons en utilisant le maximum d'information disponible dans chacun des sous-problèmes. On inclut aussi dans l'analyse des cas disponibles, le cas où une variable entière est retirée du jeu de données parce que son taux de valeurs observées est trop faible ou inférieur à 1 (dans ce dernier cas, la méthode prend le nom d'*analyse des variables complètes*).

De manière plus précise, deux exemples typiques d'utilisation de cette approche sont présentés ci-dessous :

— si l'analyse statistique requiert l'estimation d'une matrice de covariance des variables Y_j, on peut estimer la covariance entre chaque paire de variables à partir de

$$\mathrm{Cov}(Y_j, Y_{j'}) = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} y_{ij'} r_{ij} r_{ij'} - \bar{\mathbf{y}}_j^{jj'} \bar{\mathbf{y}}_{j'}^{jj'}$$

où $n_{jj'} = \sum_{i=1}^{n} r_{ij} r_{ij'}$ est le nombre de cas disponibles pour Y_j et $Y_{j'}$ et $\bar{\mathbf{y}}_j^{jj'} = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} r_{ij} r_{ij'}$ est la moyenne empirique de Y_j sur ces cas disponibles. Parfois, pour utiliser l'information maximale disponible, la moyenne est estimée par $\bar{\mathbf{y}}_j^{jj'} = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} r_{ij}$, moyenne empirique sur les cas disponibles pour Y_j. Cet estimateur peut être utilisé, par exemple, dans le cas d'un modèle linéaire (avec Y la variable à expliquer ou les variables explicatives) dans lequel l'estimation des paramètres ne fait intervenir que des estimateurs des moments du premier et du second ordre (càd, de la moyenne et des variances/covariances) mais Allison (2001) indique qu'alors, en dehors du cas MCAR, les estimations sont biaisées, comme pour l'analyse des cas complets ;

— pour l'apprentissage d'arbres de classification ou de régression (Friedman (1977) et Breiman *et al.* (1984)), les données sont partitionnées récursivement de manière binaire en recherchant, pour dans chaque nœud t déjà construit, une variable Y_j et un seuil s_j^* qui

[12] *pairwise deletion ou available-case analysis* en anglais.

 ISSN: 2102-6238

maximisent un critère d'homogénéité des ensembles $\{i \in t : y_{ij} < s_j^*\}$ et $\{i \in t : y_{ij} \geq s_j^*\}$. Les données manquantes lors de l'apprentissage sont prises en compte en définissant, pour chaque variable, le seuil de partition optimal, s_j^*, à partir des observations non manquantes, $\{i \in t : r_{ij} \neq 0\}$, uniquement. Le critère d'homogénéité est également construit sur ces observations uniquement.

Une fois le meilleur ensemble (Y_j, s_j^*) défini par minimisation du critère d'homogénéité, les données sont ensuite partitionnées en deux sous-ensembles (non disjoints) $\{i \in t : y_{ij} < s_j^* \text{ et } r_{ij} = 1\} \cup \{i \in t : r_{ij} = 0\}$ et $\{i \in t : y_{ij} \geq s_j^* \text{ et } r_{ij} = 1\} \cup \{i \in t : r_{ij} = 0\}$, ce qui correspond à la propagation des observations manquantes dans les deux branches de l'arbre. Cette approche est appelée *partitionnement probabiliste* et une alternative à celle-ci est la définition de *variables de substitution* (voir section 2.4 pour une discussion et des éléments de comparaison).

Les approches d'analyse des cas disponibles posent en général des problèmes de deux types différents, qui viennent du fait que les différents composants des modèles (covariances ou bien partition dans un arbre) sont calculés sur des sous-échantillons différents :

— d'une part, cette approche peut favoriser (ou défavoriser) de manière artificielle certaines variables selon leur taux de valeurs manquantes dans l'analyse ou la prédiction. Par exemple, en présence de données manquantes MAR, Breiman *et al.* (1984) montrent que cette approche ne dégrade que peu les performances en apprentissage de la méthode sauf si les données sont manquantes de manière plus importantes pour les variables susceptibles d'être les plus pertinentes pour partitionner l'échantillon. Dans ce dernier cas, l'utilisation de la stratégie d'analyse des cas disponibles a des effets sur les performances en apprentissage : les erreurs en apprentissages sont majorés par rapport à d'autres approches comme l'utilisation de variables de substitution ;

— d'autre part, dans le cas du calcul d'une matrice de covariance ou de corrélation, l'analyse des cas disponibles produit une matrice avec des corrélations calculées sur des individus différents et/ou sur un nombre différent d'individus. Les résultats de cette méthode résultent d'une série d'analyses sur divers sous-échantillons qui peuvent être représentatifs de populations différentes. Ce problème complique les interprétations des corrélations et limite la généralisation à une population spécifique : comme les corrélations sont calculées sur des sous-échantillons de tailles différentes, les erreurs standards des estimateurs habituels sont difficiles à obtenir. Par exemple, la stratégie consistant à les calculer en utilisant la taille moyenne des échantillons sous-estime les erreurs standards (Little, 1992). Enfin, si les moyennes sont calculées sur les cas disponibles pour chacune des deux variables indépendamment, il est possible d'obtenir des corrélations incohérentes (non comprises entre -1 et 1), en particulier pour des variables fortement corrélées (van Buuren, 2012).

Le package **regtools** propose des implémentations de type « analyse des cas disponibles » de plusieurs méthodes statistiques en étendant, par exemple, les fonctions `lm` (régression linéaire), `prcomp` (ACP) et `loglin` (modèles log-linéaires).

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; ne requiert pas de spécifier un modèle d'imputation correct ; permet de prendre en compte plus d'individus par rapport à l'analyse des cas disponibles ;

 ISSN: 2102-6238

— *Désavantages* : principalement valable dans le cas MCAR ; favorise artificiellement certaines variables ; produit des statistiques sur des sous-populations différentes, difficilement comparables.

2.3. Ajustement par variable binaire

L'*ajustement par variable binaire* [13] s'utilise dans des modèles de régression lorsque l'analyse des cas complets n'est pas possible en raison d'un trop faible nombre de cas complets (Cohen *et al.*, 1985). Elle consiste à associer à chaque variable explicative incomplète, Y_j, la variable Y_j^* définie par :

$$Y_j^* = \begin{cases} Y_j & \text{si } Y_j \text{ est observée,} \\ A & \text{sinon.} \end{cases}$$

où $A \in \mathbb{R}$ est une constante arbitraire (souvent 0 ou la moyenne de Y_j, mais sa valeur n'est pas importante). Il suffit alors de remplacer chaque variable incomplète Y_j par le couple (Y_j^*, R_j).

Par rapport à l'analyse des cas complets, cette méthode permet d'améliorer la précision de certains estimateurs en utilisant l'intégralité des individus disponibles dans le jeu de données initial. Néanmoins, cette méthode produit des estimateurs qui sont biaisés dans tous les cas.

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; alternative à l'analyse des cas complets lorsque le nombre de cas complets est trop faible ;

— *Désavantages* : produit presque systématiquement des estimateurs biaisés dans le cadre de problèmes d'inférence ; pas recommandée en pratique.

2.4. Approche par substitution de variables

Dans le cas particulier d'un modèle de prédiction (régression ou classification supervisée) dans lequel Y sont les variables explicatives, on peut aussi obtenir des prédictions à partir d'observations incomplètes de Y en utilisant des approches par substitution de variables. Ces approches sont particulièrement utilisées dans le cas d'arbres de régression ou de classification (Breiman *et al.*, 1984), qui utilisent la notion de « partition de substitution » : la partition de substitution d'une partition du nœud t par la variable Y_j et le seuil s_j^* est définie comme la partition par la variable $Y_{j'}$ (pour un $j' \neq j$) et le seuil $s_{j'}^*$ qui minimise une mesure d'association entre les deux partitions sur les individus observés.

Breiman *et al.* (1984) montrent que l'utilisation des partitions de substitution pour la prédiction d'une observation avec des données manquantes donne des performances de qualité dès lors que les observations sont manquantes aléatoirement et que plusieurs des variables explicatives Y_j sont corrélées (ce qui induit des mesures d'association élevées entre une partition et sa ou ses partitions de substitution). Ding et Simonoff (2010) vont plus loin et proposent une étude exhaustive, théorique et empirique, des diverses méthodes classiques de prise en charge des valeurs manquantes : analyse des cas disponibles (section 2.1) et analyse des variables complètes (section 2.2), imputation par la moyenne (section 4.1), création d'une modalité particulière « *manquant* » utilisé comme une modalité supplémentaire (qui est à rapprocher de la méthode décrite

[13] *dummy variable adjustment* en anglais.

 ISSN: 2102-6238

en section 2.3), utilisation de partitions probabilistes (section 2.2) et, enfin, variables de substitution. Les résultats théoriques et empiriques montrent que la qualité de l'approche dépend de deux critères :

— si les données à prédire (et pas seulement les données d'apprentissage) contiennent elles-aussi des données manquantes et que la variable à prédire est liée au processus de génération des données manquantes (ce cas contient des situations MAR et MNAR) alors l'approche par création d'une modalité supplémentaire est la plus efficace en terme d'erreur de prédiction ;

— dans tous les autres cas, les approches par substitution de variables, utilisation des variables complètes et partitions probabilistes sont, de manière à peu près équivalentes, les meilleures, avec un désavantage pour l'approche par variables complètes dans les cas de taux de manquants faibles et un désavantage pour l'approche par imputation dans les cas de taux de manquants importants.

Conclusion et recommandations :

— *Avantages* : a montré son efficacité empirique dans le cadre des arbres de régression et de classification ;

— *Désavantages* : principalement valable dans le cas MAR et lorsque les covariables sont fortement corrélées ; gourmande en temps de calcul.

3. Inférence statistique en présence de valeurs manquantes

Lorsque l'objet de l'analyse statistique est l'inférence, les approches fondées sur la modélisation paramétrique de la distribution multivariée des données, $f(Y;\theta)$ permettent d'obtenir des estimations de θ sans avoir à imputer les données et en garantissant une estimation non biaisée de ce paramètre, à condition que l'hypothèse d'ignorabilité du mécanisme de génération des données manquantes soit vérifiée. Les premiers travaux de ce type ont été proposés par Schafer (1997) et se fondent sur des approches de maximisation de la vraisemblance dans le cadre d'un modèle gaussien. On les retrouve fréquemment résumés sous le nom générique de « modélisation jointe [14] », qui regroupe des approches fréquentistes et bayésiennes.

3.1. Approches fréquentistes

Lorsque la densité $f(Y;\theta)$ est spécifiée et dans le cas d'un mécanisme ignorable, l'équation (1) indique que la vraisemblance de θ pour les données observées est de la forme

$$\mathscr{L}(\theta|Y_{\text{obs}}) \propto \log \int f(Y;\theta) dY_{\text{miss}}.$$

Les estimateurs du maximum de vraisemblance offrent des estimations non biaisées de θ mais, à cause de l'intégration, le calcul direct de la vraisemblance précédente n'est possible que dans de très rares cas en présence de données incomplètes. Les approches fréquentistes pour l'estimation

[14] *Joint Modelling* en anglais

 ISSN: 2102-6238

de θ dans ce cadre-ci peuvent être regroupées en deux grands types de méthodes : la première utilise une approche EM (Dempster *et al.*, 1977) et la seconde se fonde une approche par maximum de vraisemblance à information incomplète [15], originellement proposée par Finkbeiner (1979).

— **Algorithme EM.** L'idée de l'utilisation de l'algorithme EM consiste à alterner deux étapes :

une étape E (Expectation) dans laquelle les statistiques suffisantes du modèle sont « complétées » en tenant compte des valeurs observées et de la valeur courante du paramètre, $\theta^{(t)}$. La forme de ces statistiques dépend du modèle considéré ;

une étape M (Maximization) dans laquelle la valeur du paramètre courant est mise à jour pour obtenir $\theta^{(t+1)}$ par maximisation de la vraisemblance complétée à l'étape E.

L'approche EM présente l'avantage d'être convergente (Dempster *et al.*, 1977). Toutefois, si dans le cas d'une distribution gaussienne, les formules explicites des étapes E et M sont données dans ((Little et Rubin, 2002) et Enders (2001)), la mise en œuvre de cette approche peut s'avérer plus complexe pour d'autres distributions, comme discuté par Meng et Rubin (1993). Enfin, Enders (2001) liste un certain nombre de désavantages à cette approche, en particulier, le fait qu'elle ne fournit pas d'estimation de la variabilité des estimations de θ : une étape supplémentaire (utilisant par exemple une approche par bootstrap ; voir (Graham, 2009) et section 5.3) est nécessaire pour obtenir des estimations des erreurs types.

— **FIML.** L'approche par maximum de vraisemblance à information incomplète, quant à elle, ne remplit pas les valeurs manquantes mais détermine une vraisemblance partielle pour chaque observation i. Celle-ci, notée $\mathscr{L}_i$, est obtenue par calcul de la vraisemblance ordinaire sur les variables observées pour i (les paramètres non estimables car fondés sur des variables manquantes pour i sont remplacés par 0). Dans le cas gaussien, si on note $\theta = (\mu, \Sigma)$ les paramètres (moyenne et variance) de la loi jointe, on obtient

$$\mathscr{L}_i = K_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{y}_i^* - \mu_i)^\top \Sigma_i^{-1}(\mathbf{y}_i^* - \mu_i)$$

où $\mathbf{y}_i^*$ est le vecteur des variables observées pour l'individu i, μ_i et Σ_i correspondent respectivement au vecteur moyenne et à la matrice de covariance restreints aux variables observées pour i. K_i est une constante qui dépend du nombre de valeurs observées pour i.

Ces n quantités sont alors sommées pour obtenir la fonction de log-vraisemblance sur l'ensemble de l'échantillon :

$$\widetilde{\mathscr{L}}(\theta|Y) = \sum_i^n \mathscr{L}_i.$$

θ est enfin obtenu comme le maximum de cette vraisemblance $\widetilde{\mathscr{L}}$. Outre l'estimation du paramètre de la loi jointe des données, cette approche permet d'obtenir des erreurs types sur le paramètre, ce qui est un avantage sur l'approche précédente. Comme noté par Enders (2001), elle peut aussi être plus simple à mettre en œuvre que l'approche EM car elle ne nécessite pas de dériver une étape E spécifique à chaque modèle.

[15] FIML : *Full Information Maximum Likelihood*, aussi connue sous les noms de *direct maximum likelihood* ou *raw maximum likelihood*.

 ISSN: 2102-6238

Notons que les deux approches décrites ci-dessus dépendent toutes les deux de l'hypothèse d'ignorabilité du mécanisme de génération des données manquantes. Elles sont donc restreintes au cas de données MAR et non applicables dans le cadre MNAR. Elles sont, en outre, fortement dépendantes de la véracité du modèle sous-jacent de génération des données, souvent supposé gaussien.

Enfin, ces approches sont fréquemment utilisées dans le cadre de l'imputation de données (voir section 4) : une fois θ estimé, l'imputation, c'est-à-dire, le remplacement de la valeur manquante par une valeur plausible, peut être réalisée en échantillonnant selon la loi $f(Y;\theta)$ pour compléter les valeurs manquantes. Notons toutefois que le cadre d'application de l'approche dépasse celui de l'imputation : l'approche par maximum de vraisemblance est initialement destinée à l'estimation du paramètre de la loi jointe de Y, θ, et peut donc être utilisée directement (sans avoir recours à l'imputation) si l'estimation de θ est la question d'intérêt pour le statisticien. Elle offre, en particulier, un cadre général pour l'inférence et rend possible l'utilisation de tests du rapport de vraisemblance.

Conclusion et recommandations :

— *Avantages* : bien adaptées au cadre de l'inférence statistique ; ne requièrent pas l'imputation de valeurs ; fournissent des estimations non biaisées dans le cadre d'un mécanisme ignorable ; peuvent être utilisées également pour l'imputation des valeurs manquantes ; fournissent des estimations des erreurs sur les paramètres estimés.

— *Désavantages* : seulement valables dans le cas MAR ; requièrent des hypothèses fortes sur la loi jointe des données ; gourmande en temps de calcul ; garanties asymptotiques qui requièrent des échantillons de grande taille.

3.2. Approches bayésiennes

Une autre approche pour estimer le paramètre θ de la loi jointe $f(Y;\theta)$ est le recours à une approche bayésienne dans laquelle une loi a priori est définie sur θ, $p(\theta)$. Cette loi *a priori* est utilisée pour déterminer la loi *a posteriori* du paramètre connaissant les données observées :

$$p(\theta|Y_{\text{obs}}) \propto f(Y_{\text{obs}}|\theta)p(\theta).$$

L'inférence bayésienne consiste à déterminer cette loi *a posteriori*.

En présence de valeurs manquantes, comme dans le cadre fréquentiste, l'hypothèse d'un mécanisme ignorable permet d'écrire

$$p(\theta|Y_{\text{obs}}) = \int p(\theta|Y)f(Y_{\text{miss}}|Y_{\text{obs}},\theta)dY_{\text{miss}}. \tag{2}$$

Tanner et Wong (1987) proposent un cadre général pour l'inférence bayésienne sous cette hypothèse, avec une approche par augmentation de données. Celle-ci consiste à itérer deux étapes :

une étape d'imputation (étape I) dans laquelle M tableaux de données complets sont générés selon la loi $f(Y_{\text{miss}}|Y_{\text{obs}},\theta)$ courante. Cette étape consiste à échantillonner θ M fois dans la distribution courante de $p(\theta|Y_{\text{obs}})$, $p_t(\theta|Y_{\text{obs}})$, et à utiliser les valeurs échantillonnées et la donnée de $f(Y|\theta)$ pour générer les données complètes $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}$. Tanner et Wong (1987) notent la similarité d'approche entre cette étape et l'imputation multiple (décrite en section 5.2), d'où le nom « étape d'imputation » qu'ils lui ont donné ;

 ISSN: 2102-6238

une étape postérieure (étape P) dans laquelle la valeur courante de la loi *a posteriori* est obtenue par

$$p_{t+1}(\theta|Y_{\text{obs}}) = \frac{1}{M}\sum_{m=1}^{M} p\left(\theta|Y = \mathbf{Y}^{(M)}\right)$$

L'étape P demande de pouvoir calculer analytiquement $p(\theta|Y)$. La mise en œuvre de cette approche peut donc être plus ou moins facile selon le choix de l'*a priori* effectué. Dans le cas gaussien de paramètre $\theta = (\mu, \Sigma)$, si on choisit pour *a priori* de θ l'*a priori* (non informatif) de Jeffrey (Gelman *et al.*, 2013), on sait que $\mathscr{L}(\Sigma|Y)$ est une loi de Wishart inverse à $n-1$ degré de liberté et de paramètre d'échelle S^{-1}, où S est la matrice de covariance empirique de θ, et $\mathscr{L}(\mu|\Sigma, Y) \sim \mathscr{N}(\overline{\mathbf{y}}, \frac{1}{n}\Sigma)$. Toutefois, la loi a posteriori n'a pas toujours une forme explicite simple à déterminer et on peut alors avoir recours à des algorithmes itératifs (comme l'algorithme de Gibbs) pour pouvoir échantillonner dans la loi *a posteriori*.

Tanner et Wong (1987) montrent que l'approche proposée, sous des conditions relativement peu restrictives, converge bien vers la vraie loi a posteriori $p(\theta|Y_{\text{obs}})$. En outre, l'estimation bayésienne s'adapte assez bien à tout type de répartition de données manquantes et, contrairement aux approches fréquentistes, elle est bien adaptée aux échantillons de petites tailles puisqu'elle ne repose pas sur des résultats asymptotiques. Par ailleurs, elle fournit directement une estimation de la variance associée à l'estimation des paramètres via la loi *a posteriori* et permet également, comme les approches fréquentistes, de pratiquer une imputation des données manquantes en utilisant un échantillonnage similaire à l'étape I décrite plus haut.

Conclusion et recommandations :

— *Avantages* : bien adaptée au cadre de l'inférence statistique ; ne requiert pas l'imputation de valeurs ; garanties théoriques de convergence ; peut être utilisée également pour l'imputation des valeurs manquantes ; adaptée aux échantillons de petite taille ;

— *Désavantages* : seulement valable dans le cas MAR ; gourmande en temps de calcul ; requiert des hypothèses fortes sur la loi jointe des données.

3.3. Packages R

Divers packages proposent des implémentations de ces approches :

— **Amelia** (Honaker *et al.*, 2011) propose diverses méthodes d'imputation EM et IP et des graphiques diagnostiques. Le package propose des versions fondées sur des approches bootstrap ou bayésienne pour estimer les incertitudes et gère les imputations multiples (voir section 5). Le package possède une interface graphique (AmeliaView) permettant aux personnes non familières avec R de l'utiliser ;

— **lavaan** (Rosseel, 2012) propose une approche par maximum de vraisemblance à information incomplète pour prendre en compte les données manquantes dans les modèles à équations structurelles ;

— **norm** (Schafer et Olsen, 1998) est un package proposant l'analyse de données multivariées suivant une distribution normale. La fonction `em.norm` donne les estimations des paramètres obtenues par approche EM. Pour obtenir une imputation des données manquantes, la fonction `imp.norm` peut être utilisée avec les paramètres estimés par la fonction précédente. Enfin, la fonction `da.norm` implémente l'approche bayésienne décrite

 ISSN: 2102-6238

ci-dessus. En particulier, Schafer et Olsen (1998) conseillent l'utilisation des résultats de l'algorithme EM pour initialiser l'approche bayésienne et mieux calibrer le nombre d'itérations nécessaires pour celle-ci. **cat** et **mix** (Schafer et Olsen, 1998) sont l'équivalent du package **norm** pour l'imputation de variables catégorielles et mixtes. **cat** estime les paramètres d'une distribution multinomiale pour les variables catégorielles.

4. Imputation simple

Une alternative aux approches qui se fondent sur les données observées uniquement est probablement l'approche la plus courante de traitement des données manquantes : l'imputation de celles-ci par une valeur unique utilisée pour « remplacer » la valeur non observée. On appelle cette approche *imputation simple*. Comme souligné par Schafer et Graham (2002), les approches par imputation présentent plusieurs avantages par rapport aux approches utilisant uniquement les données observées : d'une part, elles permettent de limiter la perte de puissance liée à la taille réduite de l'échantillon correspondant aux individus complètement observés. D'autre part, si les données observées contiennent suffisamment d'information pour permettre de prédire les valeurs non observées, l'inférence statistique conserve sa précision initiale. Enfin, une fois les données manquantes imputées, l'utilisateur obtient un tableau de données complet de n individus sur lequel n'importe quelle analyse statistique classique peut être pratiquée, sans nécessité d'avoir un traitement particulier personnalisé pour les valeurs non observées : ces approches ne sont donc pas restreintes au cadre de l'inférence statistique.

Selon que l'objectif de l'imputation est l'inférence statistique d'une quantité d'intérêt ou bien l'obtention d'un tableau complet permettant diverses analyses statistiques, l'impact des erreurs d'imputation est différent. Les différentes méthodes d'imputation s'intéressent donc à conserver au mieux certains aspects dans les variables observées (distribution univariée, corrélations entre variables, etc) en fonction de l'objectif de l'utilisateur. L'erreur commise par la méthode d'imputation est alors mesurée soit en terme d'erreur commise sur la valeur imputée elle-même (erreur d'imputation, voir section 5.1 sur les outils de diagnostic), soit sur le résultat de l'analyse.

Dans cette section, nous décrivons les méthodes les plus courantes d'imputation simple, que nous avons organisées en trois grandes familles (complétion stationnaire, imputation fondée sur des similarités entre individus, imputation fondée sur des méthodes de prédiction) auxquelles s'ajoutent les méthodes d'imputation adaptées à l'analyse factorielle des données. Dans toutes ces familles, des approches existent pour imputer des variables numériques ou catégorielles. Nous présentons les avantages et inconvénients de ces méthodes, qui sont toutes principalement adaptées au cadre MAR. En particulier, nous essayons de systématiquement mettre en avant le cadre approprié d'utilisation de celles-ci, qui est lié, à la fois, à l'usage que l'utilisateur souhaite avoir du tableau imputé, mais aussi, au type de répartition des données manquantes. Enfin, nous discutons, en conclusion de la section, d'une méthodologie appropriée pour l'analyse globale d'un tableau de données contenant des valeurs manquantes ainsi que d'ouvertures pour l'imputation dans le cadre de données ayant une structure particulière (séries temporelles, par exemple).

 ISSN: 2102-6238

4.1. Complétion stationnaire

L'imputation par *complétion stationnaire* (Schafer et Graham, 2002 et Kaiser, 2014) consiste à remplacer les valeurs manquantes de la variable Y_j par une valeur identique, m_j, pour tous les individus. Différents types de complétion stationnaire existent :

— pour une variable catégorielle prenant ses valeurs dans un ensemble fini $\{1,\dots,M\}$, le mode des valeurs $\{y_{ij} : r_{ij} \neq 0\}$ est utilisé pour l'imputation [16] : $m_j = \arg\max_{u=1,\dots,M} \text{Card}\{i : r_{ij} \neq 0 \text{ et } y_{ij} = u\}$;

— pour une variable numérique, la valeur moyenne ou médiane des $\{y_{ij} : r_{ij} \neq 0\}$ est utilisée pour l'imputation : $m_j = \frac{1}{\sum_{i=1}^n r_{ij}} \sum_{i=1}^n y_{ij} r_{ij}$. L'imputation par la moyenne est simple à mettre en œuvre mais ses propriétés sont limitées : elle distord la distribution de la variable d'intérêt même dans le cas MCAR. Par conséquent, certaines caractéristiques de la distribution sont biaisées, en particulier la variabilité qui est réduite ;

— pour une variable numérique, une combinaison convexe des valeurs $\{y_{ij} : r_{ij} \neq 0\}$ peut également être utilisée pour l'imputation : $m_j = \frac{1}{\sum_{i=1}^n r_{ij}} \sum_{i=1}^n w_i y_{ij} r_{ij}$ où w_i sont les poids de la combinaison linéaire tels que $\frac{\sum_{i=1}^n w_i r_{ij}}{\sum_{i=1}^n r_{ij}} = 1$. L'imputation par la moyenne est un cas particulier d'imputation par combinaison linéaire (dans lequel $w_i = 1$).

Schafer et Graham (2002) soulignent un des principaux problèmes de cette approche : dans le cas simple de l'estimation de la moyenne de la variable Y_j, contenant des valeurs manquantes, l'imputation par la moyenne diminue la taille attendue de l'intervalle de confiance d'une part en introduisant un biais qui diminue la valeur de l'écart type empirique de Y_j et d'autre part en sur-estimant, par n, le nombre de valeurs observées. Les auteurs montrent que pour 25% de valeurs manquantes, le taux d'erreur observé sur l'intervalle de confiance de la moyenne est près de trois fois ce qu'il devrait être. Enfin, outre une sous-estimation de la variabilité des variables, y compris dans le cas MCAR, cette approche modifie les corrélations entre variables. Pour limiter ces problèmes, des variantes de l'imputation stationnaire peuvent être mises en œuvre : en particulier, lorsque la population est naturellement stratifiée en sous-populations homogènes, l'imputation par complétion stationnaire peut être réalisée indépendamment dans chacune des sous-populations.

Enfin, un autre exemple de méthode d'imputation se rapprochant de la complétion stationnaire est celui de données longitudinales où la variable Y_j est mesurée pour les individus i à divers pas de temps $t = 1,\dots,T$. Dans ce cas, l'imputation d'une valeur manquante y_{ijt} peut être faite par la dernière valeur connue de cette variable pour cet individu, y_{ijt^*}, pour $t^* = \arg\max_{u=1,\dots,t-1}\{r_{iju} \neq 0\}$. Cette approche, souvent abrégée par LOCF [17] et aussi connue sous le nom de « analyse du point final [18] », fait l'hypothèse implicite qu'il n'y a pas eu de changement entre t^* et t. C'est une approche de gestion des données manquantes très largement pratiquée dans le cadre d'études cliniques longitudinales, plus particulièrement des études dites « en intention de traiter », dans lesquelles deux groupes de malades, un groupe traité et un groupe contrôle, sont suivis de manière longitudinale. Molnar *et al.* (2008) soulignent que, dans le cas où le taux de sortie de l'étude du

[16] *Concept Common Attribute Value Fitting* en anglais.
[17] *Last Observation Carried Forward* en anglais.
[18] *endpoint analysis* en anglais

 ISSN: 2102-6238

groupe traité est lié au traitement, cette approche biaise les conclusions en faveur du traitement, avec des conséquences potentiellement très importantes pour la prise en charge médicale des malades. Ces conclusions sont confirmées par l'étude par simulations de Unnebrink et Windeler (2001) qui montre une violation du degré de significativité et une perte de puissance importante dans les tests de comparaison entre les deux groupes dans ce cas-ci.

Les approches par complétion stationnaire sont disponibles, par exemple, dans les packages **simputation** (imputation par la médiane), **Hmisc** (imputation aléatoire, par la moyenne, par la médiane, par le mode...) et **ForImp** (imputation par la moyenne, par la médiane, par le mode). De manière plus générique, la fonction `impute` du package **Hmisc** permet d'utiliser une fonction arbitraire des valeurs observées pour une imputation par complétion stationnaire.

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; permet d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ;

— *Désavantages* : biaise (diminue) l'estimation des variabilités des variables ; modifie les corrélations entre variables ; sur-estime la taille de l'échantillon observé ; non recommandée en pratique, même dans les cas MCAR, sauf si le nombre de valeurs manquantes est très faible et que l'on ne sait pas mettre en œuvre une autre méthode décrite dans ce papier.

4.2. Méthodes fondées sur des similarités entre individus

Une autre approche pour l'imputation simple consiste à utiliser les valeurs observées des individus similaires à l'individu pour lequel une valeur est manquante. Ces méthodes sont liées à des imputations par k plus proches voisins (kNN) ou à des méthodes regroupées sous le nom générique d'approches « hot-deck » (les deux dénominations étant parfois confondues selon les publications).

4.2.1. Méthode des k plus proches voisins (kNN)

La méthode kNN est une méthode d'imputation multivariée fondée sur une notion de distance entre individus, $d(i,i')$, obtenue à partir de q covariables entièrement observées, X. Pour une valeur manquante y_{ij}, l'approche consiste, d'une part, à calculer l'ensemble des distances $d(i,i')$ pour les $i' \neq i$ tels que $r_{i'j} \neq 0$ et à retenir les k observations (pour un $k \in \mathbb{N}^*$), $y_{(1)j}, \ldots, y_{(k)j}$, correspondant aux k plus petites distances. Les k valeurs $(y_{(i)j})_{i=1,\ldots,k}$ des plus proches voisins sont alors agrégées pour imputer la valeur manquante y_{ij}. Généralement, si la variable Y_j est numérique, la valeur manquante est imputée par la moyenne (ou la médiane) des $(y_{(i)j})_{i=1,\ldots,k}$. L'approche se généralise facilement au cas où il n'y a pas de covariables complètement observées en calculant des distances, pour chaque individu, qui sont basées sur un sous-ensemble d'individus et/ou de variables complètement observées.

La méthode requiert le choix de deux hyper-paramètres : d, la distance choisie, et k, le nombre de voisins utilisés pour l'estimation. Des choix classiques pour d sont la distance euclidienne entre valeurs observées,

$$d(i,i') = \sum_{j'=1}^{q} (x_{ij'} - x_{i'j'})^2, \tag{3}$$

 ISSN: 2102-6238

ou la distance de Mahalanobis. Lorsque le jeu de données contient des variables catégorielles, Zhang (2012) propose l'utilisation d'une distance particulière prenant en compte l'existence de ces variables et la valeur imputée est alors le mode des $(y_{(i)j})_{i=1,\dots,k}$. Moeur et Stage (1995) proposent une approche alternative fondée sur l'analyse canonique des corrélations entre les covariables X et les cas complets de Y : les plus proches voisins sont alors définis dans l'espace factoriel de projection de X. L'idée sous-jacente est de sélectionner les plus proches voisins dans un espace de corrélation optimale avec les variables à imputer.

Pour le choix de k, Jönsson et Wohlin (2004) soulignent que les recommandations pour le choix de cette valeur varient selon les auteurs : par exemple, Chen et Shao (2000) et Huisman (2000) utilisent $k = 1$ ou 2, Baretta et Santaniello (2016) recommandent d'utiliser une valeur faible de k alors que Troyanskaya *et al.* (2001) recommandent une valeur de k comprise entre 10 et 20 pour des jeux de données de grande taille. Dans leurs expériences, Jönsson et Wohlin (2004) mettent en valeur une dépendance de k à la taille du jeu de données et suggèrent de choisir k égal à la racine carrée du nombre moyen de cas complets des variables utilisées pour l'imputation.

L'imputation par kNN est implémentée dans de nombreux packages R. Parmi ceux-ci, on peut citer :

— **DMwR** : ce package regroupe des fonctions utiles pour la fouille de données et est associé à l'ouvrage de Torgo (2010). La fonction `knnImputation` de ce package propose deux méthodes d'imputation des valeurs manquantes. La méthode par défaut est une moyenne pondérée, le poids de l'individu i' étant donné par $\exp(-d(i,i'))$ où d est la distance euclidienne entre l'individu imputé, i et i'. L'approche alternative consiste à remplacer chaque valeur manquante par la médiane des kNN (ou bien par le mode quand la variable à imputer est catégorielle) ;

— **impute** (Troyanskaya *et al.*, 2001) : ce package Bioconductor est destiné à l'imputation de données d'expressions de gènes (puces à ADN) et requiert donc un tableau de variables numériques. La méthode proposée dans ce package calcule des voisins dans l'espace des gènes et non dans l'espace des individus. Pour accélérer le calcul des distances euclidiennes entre gènes, le package utilise un pré-traitement par classification non supervisée et réduit le calcul des distances à un sous-groupe de gènes. L'imputation par la moyenne des kNN est finalement réalisée ;

— **VIM** (Kowarik et Templ, 2016) : ce package autorise l'imputation par kNN pour des données mixtes. Pour ce faire, les k voisins sont choisis en utilisant une variation de la distance de Gower (Gower, 1971). Cette distance peut s'appliquer à un ensemble de variables à la fois numériques, catégorielles et binaires. Elle est fondée sur une notion de *contribution* de la covariable X_j qui est définie par

$$\begin{array}{cc} X_j \text{ est numérique} & X_j \text{ est catégorielle} \\ S_{ii'j} = \frac{|x_{ij} - x_{i'j}|}{\max_l(x_{lj}) - \min_l(x_{lj})} & S_{ii'j} = \begin{cases} 1 & \text{si } x_{ij} = x_{i'j} \\ 0 & \text{sinon.} \end{cases} \end{array}$$

De cette notion, on peut déduire une distance entre individus i et i' comme suit :

$$d(i,i') = \frac{\sum_{j=1}^{p} S_{ii'j}}{n}.$$

 ISSN: 2102-6238

Les variables numériques sont finalement imputées par la médiane des valeurs des voisins tandis que les variables catégorielles sont imputées par le mode des valeurs des voisins ;

— **yaImpute** (Crookston et Finley, 2008) : ce package met à disposition une grande variété de méthodes d'imputation par *k*NN, dont l'approche d'imputation par analyse canonique des corrélations décrite plus haut et propose plusieurs outils diagnostiques pour l'évaluation et la comparaison des approches d'imputation.

4.2.2. Hot-deck

L'imputation hot-deck est une approche qui a été introduite en 1947 pour traiter les valeurs manquantes dans les réponses des sondages démographiques (*Current Population Survey*) par le bureau national américain des sondages (*US Census Bureau*). Andridge et Little (2010) font une revue des méthodes hot-deck et de leurs propriétés.

L'imputation hot-deck est fondée sur le concept de *donneur*, qui est proche du concept de plus proche voisin. De manière plus précise, pour un individu i ayant une valeur manquante y_{ij}, on définit un ensemble de donneurs $\mathscr{D}(i)$ qui sont des individus i' « similaires » à l'individu i et pour lesquels $r_{i'j} \neq 0$. Une des valeurs $y_{i'j}$ pour $i' \in \mathscr{D}(i)$ est alors imputée pour y_{ij}. Les variantes de la méthode hot-deck diffèrent à deux niveaux : dans la phase de définition de l'ensemble des donneurs et dans la phase d'imputation.

Généralement, l'ensemble des donneurs d'un individu i est défini par le biais d'une mesure de similarité ou de distance calculée sur des covariables complètement observées, X, mais d'autres approches sont parfois pratiquées. Les plus courantes sont les suivantes :

— **Hot-deck métrique ou plus proches voisins**
Dans cette variante, l'ensemble des donneurs est défini comme l'ensemble des *k*NN de l'individu i pour une distance donnée calculée sur un ensemble de covariables X, complètement observées. La distance euclidienne est généralement utilisée. Cette approche est similaire au cas de l'approche *k*NN (voir section 4.2.1) mais diffère dans la phase d'imputation (voir ci-dessous), sauf pour le cas $k = 1$.

— **Hot-deck métrique avec score d'affinité**
Une autre méthode pour calculer la similarité entre deux individus a été proposée par Cranmer et Gill (2012) : le score d'affinité. Le score d'affinité $s(i,i')$ mesure le degré de similarité qui existe entre l'individu receveur i et chaque donneur potentiel i', pour lequel les p variables du jeu de données ont été observées. Il a été établi, dans un premier temps, pour des données discrètes et se définit alors comme la proportion de valeurs communes entre i et i' parmi les variables observées pour le receveur i :

$$s(i,i') = \frac{\sharp\{j = 1,\ldots,p : r_{ij} = 1 \text{ et } y_{ij} = y_{i'j}\}}{\sum_{j=1}^{p} r_{ij}}.$$

Dans le cas de variables numériques continues, Cranmer et Gill (2012) proposent d'adapter le score d'affinité de la manière suivante :

$$s(i,i') = \frac{\sum_{j=1}^{p} r_{ij} \mathbf{1}_{\{|y_{ij} - y_{i'j}| < \sigma\}}}{\sum_{j=1}^{p} r_{ij}}$$

 ISSN: 2102-6238

où σ est un seuil à fixer (qui peut éventuellement être adapté en fonction de l'échelle de la variable). Dans les deux cas, l'ensemble des donneurs, $\mathscr{D}(i)$, se définit alors par $\mathscr{D}(i) = \left\{i' : \ s(i,i') = \max_{l \neq i} s(i,l)\right\}$.

— **Hot-deck hiérarchisé**
L'approche hot-deck hiérarchisé est similaire au cas d'imputation de données longitudinales décrit dans la section 4.1. Elle est utilisée lorsqu'il existe un ordre naturel entre les variables ($j = 1, \ldots, p$) et consiste à remplacer la valeur manquante y_{ij} par la valeur d'un individu qui a les mêmes valeurs pour les variables $Y_1, Y_2, ..., Y_{j-1}$. S'il n'en existe pas, elle est remplacée par la valeur d'un individu ayant les mêmes valeurs pour les variables $Y_1, Y_2, ..., Y_{j-2}$. Ce processus est itéré jusqu'à obtention d'au moins un individu correspondant à un critère de correspondance. Cette méthode est donc fondée sur une définition modifiée de l'ensemble des donneurs $\mathscr{D}(i)$ qui sont des individus identiques à l'individu i pour certaines variables et a une phase d'imputation spécifique bien définie.

Une fois l'ensemble des donneurs $\mathscr{D}(i)$ défini, l'imputation est pratiquée selon diverses méthodes :

— **Hot-deck aléatoire avec ou sans remise**
L'approche hot-deck aléatoire consiste à remplacer une valeur manquante y_{ij} par la valeur $y_{i'j}$ pour un i' choisi au hasard dans $\mathscr{D}(i)$. Cette approche peut être utilisée pour des variables numériques ou catégorielles mais nécessite que les individus du jeu de données aient un profil homogène pour que les valeurs imputées ne soient pas éloignées de la vraie valeur. Aussi, si la population s'avère trop hétérogène, il est préférable de constituer des classes d'imputation réputées plus homogènes. La méthode hot-deck aléatoire est alors appliquée à l'intérieur de ces sous-populations et on parle alors de « hot-deck par classes ». En pratique, les classes d'imputation sont souvent définies en stratifiant le jeu de données selon des covariables entièrement observées ou en appliquant des procédures usuelles de classification sur le jeu de données (Joenssen et Bankhofer, 2012).

— **Hot-deck séquentiel**
L'approche hot-deck séquentielle (Little et Rubin, 2002) est utilisée lorsqu'il existe un ordre naturel au sein des individus $i = 1, \ldots, n$. Si une valeur y_{ij} est manquante, elle est alors imputée par la valeur non manquante la plus récente parmi l'ensemble des donneurs $\mathscr{D}(i)$, y_{i^*j} avec $i^* = \arg\max_{i'=1,\ldots,i-1}\{y_{i'j} : \ r_{i'j} = 1\}$. En pratique, les variables sont ordonnées par le choix d'une variable (ou de plusieurs variables) de tri parmi les covariables X_j observées pour tous les individus. Celle-ci doit expliquer au mieux la variable à imputer (à partir des observations correspondant aux individus répondants) et, si besoin, les covariables de tri suivantes sont utilisées pour ordonner les ex-aequos. Comme l'estimateur obtenu dépend de l'ordre dans lequel les données sont ordonnées, il est nécessaire que la covariable de tri choisie ne soit pas fortement corrélée avec la probabilité de non-réponse. La conséquence du non respect de cette règle est l'imputation de la même valeur pour un grand nombre d'individus et donc la distorsion de la distribution de la variable imputée (Kalton et Kasprzyk, 1986), qui entraîne une distorsion de la distribution des données et diminue artificiellement la variance estimée. Une solution de type hot-deck hiérarchisé, comme décrite ci-dessus, permet de limiter ce type de problème.

L'imputation hot-deck est implémentée dans les packages R suivant :

 ISSN: 2102-6238

— **hot.deck** (Cranmer et Gill, 2012) : outre l'imputation simple par hot-deck métrique avec score d'affinité, ce package propose une imputation multiple (voir section 5.2) ;

— **HotDeckImputation** : ce package propose différentes méthodes d'imputation hot-deck : hot-deck séquentiel, hot-deck aléatoire, hot-deck métrique par *k*NN ainsi qu'une méthode appelée « hot-deck séquentiel CPS ». Cette dernière permet d'appliquer l'approche hot-deck séquentiel parmi les classes d'imputation ;

— **VIM** : outre les fonctionnalités d'analyse exploratoire des données manquantes, ce package propose également plusieurs approches d'imputation hot-deck (dont le hot-deck aléatoire et le hot-deck séquentiel) dans la fonction `hotdeck`. Le package **simputation** possède également une fonction `impute_hotdeck`, qui utilise les fonctions de VIM et permet divers types d'imputation hot-deck ;

4.2.3. Cold-deck

Cette approche est proche de la méthode hot-deck présentée dans la section précédente mais, dans ce cas-ci, les donneurs ne sont pas des individus du jeu de données initial. De manière plus précise, les mêmes variables Y ont été observées sur un second ensemble d'individus $i = n+1,\ldots,n+m$ et les donneurs sont définis au sein de cet ensemble. Par exemple, l'imputation de la valeur manquante y_{ij} pour un $i \leq n$, requiert la définition de l'ensemble des donneurs $\mathscr{D}(i) \subset \{n+1,\ldots,n+m\}$, par exemple par calcul des distances euclidiennes :

$$\forall i' = n+1,\ldots,n+m, \qquad d(i,i') = \sum_{j' \neq j} r_{ij}(y_{ij'} - y_{i'j'})^2.$$

Les cas typiques d'utilisation sont les cas où les donneurs proviennent d'enquêtes antérieures, de données historiques ou de l'expertise d'un spécialiste (Andridge et Little, 2010).

4.2.4. Conclusion et recommandations

L'avantage principal des méthodes basées sur des mesures de similarité est qu'elles ne requièrent pas d'hypothèses sur la distribution des données : elles peuvent être utilisées de manière souple avec des données de types variés et peuvent même s'adapter à des métriques d'intérêt spécifiques aux données étudiées (comme les distances basées sur la phylogénie entre espèces utilisées en biologie, par exemple ; Cranmer et Gill, 2012).

Les approches hot-deck préservent la distribution univariée des données dans le cadre MCAR (Enders, 2010, chap. 2) et des modifications de l'approche permettent d'obtenir des estimateurs sans biais de la moyenne dans le cadre MAR (Andridge et Little, 2010). Les valeurs imputées sont des valeurs observées donc réalistes et elles ne nécessitent pas d'hypothèses paramétriques fortes. Elles permettent, en outre, d'imputer à la fois des variables numériques et catégorielles. Toutefois, ces méthodes produisent des estimateurs biaisés de nombreux paramètres pour tout type de mécanisme de génération des données manquantes (y compris MCAR). En particulier, ces approches ne sont pas adaptées à l'estimation des mesures d'association entre les variables (Schafer et Graham, 2002), même si quelques solutions ont été proposées pour résoudre ce problème dans le cas de données manquantes monotones (Andridge et Little, 2010). Fay

 ISSN: 2102-6238

(1996) montre également, sur des simulations, que la variance de l'estimateur de la moyenne est sous-estimée lorsque calculée directement sur les données imputées par hot-deck. Ce dernier problème peut être limité par l'utilisation de méthodes de ré-échantillonnage ou par l'imputation multiple (voir section 5.2). Enfin, Andridge et Little (2010) soulignent que hot-deck est moins sensible à une mauvaise spécification des hypothèses qui sous-tendent l'imputation (imputation hiérarchique, par plus proches voisins, ...) que les méthodes paramétriques mais que cet avantage est principalement visible lorsque la taille de l'échantillon est suffisamment grande. L'imputation hot-deck est, en effet, très dépendante de la richesse de l'ensemble des donneurs potentiels et celle-ci se dégrade rapidement lorsque la taille de l'échantillon est faible.

Enfin, les approches kNN sont principalement étudiées et évaluées d'un point de vue empirique. En particulier, Baretta et Santaniello (2016) montrent, sous divers types de mécanismes de génération des données manquantes, que prendre $k > 1$ permet d'améliorer la qualité de l'imputation par rapport à $k = 1$ en terme d'erreur sur la valeur imputée et d'erreur quadratique moyenne sur l'estimation de diverses statistiques (coefficient de corrélation et de régression) à partir des données imputées mais l'augmentation de k tend à déformer, de manière croissante, la distribution univariée des variables imputées et, notamment, à modifier leurs variances.

Conclusion et recommandations :

— *Avantages* : faciles à mettre en œuvre ; permettent d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ; non paramétriques et peuvent prendre en compte divers types de distance ; préserve la distribution univariée des données (HD) ; sans biais dans le cas MCAR pour l'estimation de la moyenne (HD) ;

— *Désavantages* : déforme la distribution univariée des données (kNN) ; déforment les relations multivariées ; pas recommandée si n est faible (HD).

4.3. Approches par prédiction

Une approche alternative pour imputer des valeurs manquantes est d'avoir recours à des approches par prédiction. Pour imputer la valeur manquante y_{ij}, ces méthodes estiment un modèle de régression de Y_j sur les autres variables, $(Y_{j'})_{j' \neq j}$, pour lesquelles $y_{ij'}$ est observée ou sur les covariables complètement observées, X. La prédiction obtenue pour l'individu i est alors utilisée pour imputer y_{ij}.

Parmi ces méthodes, on peut citer la régression locale (ou LOESS, Cleveland et Devlin, 1988), fréquemment utilisée. Elle consiste à construire un polynôme de faible degré, ajusté autour de la donnée manquante, par kNN. De manière plus précise, si seule la valeur y_{ij} est manquante pour l'individu i, les kNN de i sont sélectionnés parmi l'ensemble des individus pour lesquels toutes les variables sont observées. Si ces observations sont notées $(1), \ldots, (k)$, le problème de régression linéaire par moindres carrés est estimé :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p-1}} \sum_{i'=1}^{k} \left(\beta^\top \mathbf{y}_{(i')}^{-j} - y_{i'j} \right)^2$$

où $\mathbf{y}_{(i')}^{-j}$ est le vecteur des observations des $p-1$ variables autres que Y_j pour l'individu (i'). La valeur y_{ij} est alors imputée par

$$\hat{\beta}^T \mathbf{y}_i^{-j}.$$

 ISSN: 2102-6238

Cette approche se généralise de manière évidente au cas où plusieurs variables sont à imputer pour un même individu i ou bien en remplaçant la sélection des kNN par un poids décroissant en la distance entre l'individu pour lequel la valeur est manquante, i, et les autres individus, utilisés pour estimer β.

Au-delà de la méthode LOESS décrite ci-dessous, de nombreuses autres approches de régression ou de classification, paramétriques et non paramétriques, sont utilisées de la même manière pour imputer des valeurs manquantes pour des variables numériques ou catégorielles. Parmi celles-ci, on peut citer les plus courantes comme

— la régression linéaire ou sa version robuste utilisant des M estimateurs (Hubert et Ronchetti, 2009) ;

— les régressions linéaires pénalisées de types Lasso (Tibshirani, 1996), ridge (Hoerl et Kennard, 1970), elasticnet (Zou et Hastie, 2005), régression pas à pas (Hocking, 1976) ;

— les méthodes de régression non paramétriques comme les arbres de régression CART (Breiman *et al.*, 1984) ou les forêts aléatoires (Breiman, 2001). Stekhoven et Bühlmann (2012) proposent également une approche d'imputation par prédiction qui est itérative et fondée sur les forêts aléatoires. Celle-ci est implémentée dans le package **missForest**.

En outre, le package **simputation** est un package permettant d'effectuer de l'imputation par prédiction de manière très générique et avec une syntaxe simplifiée. Certaines méthodes de régression y sont pré-implémentées (régression linéaire, régression linéaire robuste, CART, forêts aléatoires, ...) et la fonction `impute_proxy` permet de mettre en œuvre une méthode d'imputation définie après estimation d'une fonction de prédiction arbitraire. Ainsi, par exemple, l'imputation par LOESS peut être réalisée en combinant cette fonction avec un modèle obtenu par le package **locfit**. Le package **VIM** propose également des méthodes d'imputation fondées sur la régression linéaire ou la régression linéaire généralisée (fonction `regressionImp`). L'imputation par régression est aussi utilisée dans le contexte d'études génétiques : dans celles-ci, des marques de mutation (appelées SNP) sont collectées à divers endroits du génome d'individus d'intérêt et ce type de données contient généralement un grand nombre de valeurs manquantes. Dans ce cadre-ci, le package **snpStats** (Bioconductor) propose une imputation qui combine une régression pas à pas pour sélectionner un ensemble de marqueurs permettant de bien expliquer un marqueur d'intérêt et un modèle de régression généralisé utilisant cet ensemble de marqueurs pour la prédiction.

Il existe plusieurs types d'amélioration des méthodes par prédiction :

— l'approche par *régression stochastique* se propose d'injecter un bruit aléatoire lors de l'étape de prédiction (Little et Rubin, 2002). Ceci a pour objectif de limiter la sous-estimation de la variabilité et la sur-corrélation des variables imputées. Cette méthode (prédiction par régression ridge puis injection de bruit) est implémentée dans la fonction `mice.impute.norm` du package **mice** (van Buuren et Groothuis-Oudshoorn, 2011) (fonction `mice`) ;

— l'approche par *spécification de lois conditionnelles (FCS)* [19] (van Buuren, 2007) spécifie, de manière paramétrique et pour toute variable Y_j ayant des valeurs manquantes, la densité conditionnelle des lois $f(Y_j|Y_{-j}, R; \theta_j)$, avec Y_{-j} l'ensemble des variables différentes de

[19] *Fully Conditional Specification*, en anglais.

 ISSN: 2102-6238

Y_j et θ_j le paramètre permettant de spécifier la loi conditionnelle. Après une initialisation de l'imputation (par exemple, une imputation par la moyenne), et pour chaque variable j, traitée par ordre croissant du nombre de valeurs manquantes, deux étapes sont itérées :

— $\theta_j^{(t)}$ est tirée aléatoirement selon la loi $p(\theta_j | Y_j = \mathbf{y}_{1,\text{obs}}, Y^{-j} = \mathbf{Y}^{-j,(t-1)})$;

— $\mathbf{y}_1^{(t)}$ est tirée aléatoirement selon la loi $f(Y_{\text{miss}} | Y_j = \mathbf{y}_{1,\text{obs}}, Y^{-j} = \mathbf{Y}^{-j,(t-1)}; \theta_1^{(t)})$.

L'approche est donc relativement similaire aux approches bayésiennes décrites dans la section 3.2 mais permet de créer des modèles de spécification des données plus flexibles, qui prend en compte les spécificités de chaque variable (contraintes de positivité, dépendances conditionnelles entre variables, ...) de manière plus naturelle. Comme les méthodes de la section 3.2, elle est fréquemment utilisée pour l'imputation multiple (voir section 5.2).

Les approches d'imputation par régression sont très largement utilisées pour produire un jeu de données complet avant analyse. Elles sont relativement flexibles, s'adaptant aux *a priori* sur les données, par l'utilisation de modèles de prédiction paramétriques ou non paramétriques. Leur performance est donc fortement dépendante de deux aspects : le premier est la capacité à pouvoir estimer des valeurs réalistes pour les valeurs manquantes à partir des valeurs observées sur les autres variables. Elles requièrent donc une dépendance entre les variables utilisées pour l'imputation et celles qui sont imputées. Elles ne couvrent donc pas non plus, *a priori*, le cas MNAR. Le deuxième aspect est la nécessité de bien spécifier la méthode de régression (ou le modèle de régression dans un cadre paramétrique) permettant d'imputer les variables : les approches classiques d'évaluation des méthodes de prédiction (validation croisée, ...) peuvent donc être utiles pour évaluer la fiabilité de l'approche choisie. Par ailleurs, il faut noter que l'approche est difficilement praticable lorsque certaines variables ont un fort ratio de manquants (les modèles de régression, dont la précision dépend directement du nombre de valeurs observées pour la variable à imputer, sont alors difficilement estimables) ou lorsque les valeurs manquantes entre les diverses variables sont fréquemment liées aux mêmes individus (il est alors difficile d'avoir suffisamment de variables observées pour estimer un modèle de régression) : elles sont donc mieux adaptées aux répartitions de données manquantes sans structure. Enfin, les garanties théoriques pour ces méthodes concernent principalement l'erreur commise sur la valeur imputée (par rapport à la valeur réelle non observées, et pas l'inférence statistique qui pourraient être pratiquées sur le tableau de données imputées) et découlent directement des garanties théoriques connues pour les diverses méthodes de régression utilisées.

Conclusion et recommandations :

— *Avantages* : permettent d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ; flexibles (large choix d'approches de régression) ;

— *Désavantages* : principalement valables dans le cas MAR ; requièrent une bonne spécification de la méthode de régression ; requièrent une bonne prédictibilité des variables ayant des valeurs manquantes par les autres variables ; cadre théorique lié à l'erreur quadratique sur la valeur imputée (et non aux résultats de l'analyse statistique pratiquée).

4.4. Approches factorielles pour l'analyse exploratoire

Il est important de souligner qu'un grand nombre de travaux étudiant le traitement des données manquantes se placent dans un cadre inférentiel (c'est le cas, par exemple, de l'ouvrage de

 ISSN: 2102-6238

référence de Little et Rubin, 2002). Ceux-ci peuvent ne pas être bien adaptés à un cadre exploratoire comme l'analyse de données, dans lequel des critères géométriques sont privilégiés par rapport aux hypothèses de nature probabilistes. Parmi les analyses exploratoires, l'Analyse en Composantes Principales (ACP) tient une place importante et son extension en présence de valeurs manquantes a été largement étudiée (Josse *et al.*, 2009 et Ilin et Raiko, 2010). De nombreux problèmes sont soulignés pour la pratique de l'ACP en présence de manquants : difficulté pour le centrage et la réduction des variables, non unicité de la solution de minimisation de la fonction de coût classique en ACP, extension non triviale de la notion de base de l'ACP, ...

Dans l'étude de l'ACP en présence de valeurs manquantes, deux objectifs complémentaires sont visés : celui de la réalisation d'une ACP en présence de valeurs manquantes et celui de l'utilisation de l'ACP pour imputer des valeurs manquantes. Dans le cadre d'études de simulations où des données manquantes sont produites de manière artificielle pour évaluer la qualité des algorithmes (sur-imputation ; voir section 5.1.1), ces deux objectifs sont évalués par des métriques de performance différentes (Josse *et al.*, 2009) : coefficient RV (Escoufier, 1973) entre les coordonnées des individus sur les données complètes par rapport aux coordonnées produites par les approches d'ACP adaptées, d'une part, et erreur de reconstitution entre valeurs initiales et valeurs imputées, d'autre part.

De nombreuses variantes des methodes de prises en compte des valeurs manquantes dans l'ACP ont été proposées dont les principales sont :

— **Nonlinear Iterative Partial Least Squares (NIPALS)** (Wold, 1966). Le principe de cette méthode est aussi à la base de la régression PLS (*Partial Least Squares* ; Tenenhaus, 1998). Il permet de réaliser une ACP avec données manquantes sans supprimer les individus i pour lesquelles une valeur y_{ij} est manquante et sans imputer les valeurs manquantes. En ce sens, la méthode se rapproche des méthodes fondées sur l'analyse des cas disponibles, décrites dans la section 2.2, mais elle peut, en outre, être utilisée comme base pour l'imputation des valeurs manquantes.

De manière plus précise, si on suppose les variables $(Y_1, ..., Y_p)$ centrées, l'algorithme NIPALS utilise la formule de décomposition de l'ACP suivante :

$$\mathbf{Y} \simeq \sum_{h=1}^{d} \mathbf{t}_h \rho_h^\top$$

où $d \leq p$ est la dimension de projection permettant d'obtenir une « bonne » reconstitution des données et $\{\mathbf{t}_h\}_{h=1,...,d} \subset \mathbb{R}^n$ et $\{\rho_h\}_{h=1,...,d} \subset \mathbb{R}^p$ sont, respectivement, les composantes principales et les vecteurs directeurs des axes principaux de l'ACP. Ceci implique que les observations de la variable Y_j peuvent s'écrire comme une régression linéaire sur les composantes $(\mathbf{t}_h)_h$: $Y_j = \sum_{h=1}^{d} \rho_{hj} \mathbf{t}_h$ (et respectivement pour l'individu i qui peut être écrit comme une régression sur les axes principaux).

L'algorithme NIPALS utilise cette remarque et estime, de manière itérative et jusqu'à convergence, les $(\rho_h)_h$ et les $(\mathbf{t}_h)_h$ par régressions successives sur les valeurs observées, en initialisant les composantes principales, par exemple, à une colonne de $\mathbf{Y}$. Contrairement à l'approche standard de l'ACP où les axes sont déterminés simultanément par décomposition spectrale, l'approche NIPALS calcule les axes successivement en utilisant une étape de déflation.

 ISSN: 2102-6238

Une fois les $(\mathbf{t}_h)_{h=1,\dots,d}$ et les $(\rho_h)_{h=1,\dots,d}$ estimés, il est possible de proposer une estimation des valeurs manquantes en utilisant la formule de reconstitution des individus :

$$\widehat{y}_{ij} = \sum_{h=1}^{d} t_{hi}\rho_{hj}. \tag{4}$$

En pratique, l'approche NIPALS fournit des solutions raisonnables lorsque le taux de manquant est faible mais elle souffre de plusieurs désavantages. Le premier est que lorsqu'une proportion importante de valeurs sont manquantes, la procédure itérative de NIPALS propage les erreurs d'axe en axe et sa convergence n'est pas garantie. Par ailleurs, si l'ACP est pratiquée sur les données centrées et réduites, NIPALS ne peut réaliser une mise à jour de l'écart-type des variables (à cause de la déflation) et produit donc un résultat qui ne correspond pas à une ACP réduite. Enfin, les axes obtenus ne sont pas nécessairement orthogonaux et le critère classique de minimisation de l'erreur de reconstitution de l'ACP,

$$\sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{h=1}^{d} t_{hi}\rho_h \right\|^2, \tag{5}$$

n'est pas minimisé par la procédure séquentielle.

— **ACP itérative** (Kiers, 1997). L'ACP itérative est une approche itérative qui vise à minimiser l'erreur de reconstitution de l'ACP (équation (5)). L'initialisation de la méthode attribue une valeur arbitraire aux données manquantes (souvent la moyenne de la variable considérée). Une ACP est ensuite effectuée sur ce jeu de données rendu complet et les données initialement manquantes sont alors mises à jour via la formule de reconstitution de l'équation (4). Les deux étapes d'estimation de l'ACP et d'imputation sont répétées jusqu'à convergence, (Kiers, 1997) montrant que la procédure converge nécessairement, éventuellement vers un minimum local.

En raison de l'alternance des étapes d'estimation et d'imputation, similaires aux étapes *Expectation* et *Maximization* des algorithmes EM, l'ACP itérative est souvent appelée ACP-EM. En effet, l'ACP peut être vue comme un modèle statistique dans lequel les données ont une structure dans un espace à faible dimension (d) et sont corrompues par un bruit (Candès *et al.*, 2013). Cette formulation se ré-écrit sous la forme d'un modèle à effet fixe (Caussinus, 1986)

$$y_{ij} = \sum_{h=1}^{d} t_{hi}\rho_{hj} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ (i.i.d.)}, \tag{6}$$

que Josse *et al.* (2009) utilisent pour montrer que l'ACP itérative peut effectivement être vue exactement comme un algorithme EM et bénéficie donc des propriétés et des caractéristiques de ces approches.

Toutefois, l'approche souffre d'un problème de sur-ajustement aux données, particulièrement dans les cas de grande dimension ($p > n$) (Josse *et al.*, 2009). Aussi, pour pallier le problème du sur-ajustement, la version régularisée de l'ACP itérative lui est préférée. La régularisation peut être effectuée en choisissant une dimension réduite, $d \ll p$, pour la reconstitution ou bien en ajoutant un terme de pénalité en norme ℓ_2 (*ridge*) lors de l'étape

 ISSN: 2102-6238

d'imputation. Verbanck *et al.* (2015) montrent que l'ACP régularisée *ridge* peut être vue comme une extension de l'équation (6) au modèle mixte

$$\mathbf{y}_i = \mathbf{R}\mathbf{t}_i + \boldsymbol{\varepsilon}_i, \tag{7}$$

où $\mathbf{R}$ est une matrice de dimension $p \times d$, $\mathbf{t}_i \sim \mathcal{N}(0, \mathbb{I}_d)$ et $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2)$ (i.i.d.). Ce modèle, connu sous le nom d'« ACP probabiliste », est proposé initialement dans (Tipping et Bishop, 1999).

— **ACP bayésienne** (Ilin et Raiko, 2010 et Verbanck *et al.*, 2015). Diverses approches bayésiennes sont proposées dans la littérature pour l'ACP, fondées sur le modèle à effets fixes de l'équation (6) ou le modèle d'ACP probabiliste de l'équation (7). En particulier, Verbanck *et al.* (2015) montrent que l'ACP probabiliste peut être vue comme un traitement bayésien des effets fixes du modèle de l'équation (6) ou bien comme un traitement bayésien direct des effets fixes avec le modèle

$$\mathbf{y}_i = \widetilde{\mathbf{y}}_i + \boldsymbol{\varepsilon}_i, \qquad \widetilde{\mathbf{y}}_i \sim \mathcal{N}(0, \tau_d)$$

où la matrice $\widetilde{\mathbf{Y}} = \begin{pmatrix} \widetilde{\mathbf{y}}_1 \\ \vdots \\ \widetilde{\mathbf{y}}_n \end{pmatrix}$ est de dimension $n \times d$. (Ilin et Raiko, 2010) proposent d'autres *a priori* bayésiens et font le lien entre diverses variantes de l'ACP probabiliste. Ils proposent également des versions rapides de l'estimation, utilisant des approches en ligne ou des approximations variationnelles, qui montrent des résultats encourageants sur les données de la compétition Netflix (2007) (qui consiste à compléter un tableau de notes de $p = 17\,770$ films évalués par $n = 480\,189$ spectateurs et contenant plus de 98% des données manquantes).

Enfin, comme beaucoup de méthodes d'analyse factorielle s'apparentent à l'ACP, il est possible d'étendre l'imputation par ACP à celles-ci. Ainsi, une méthode d'imputation fondée sur l'Analyse des Correspondances Multiples (ACM), proposée par Audigier *et al.* (2016a), permet de gérer l'imputation de variables catégorielles et une méthode fondée sur l'Analyse Factorielle Multiple (AFM), proposée par Josse *et al.* (2012), permet de prendre en compte la structuration d'un jeu de données en blocs de variables. De même, une méthode fondée sur l'Analyse Factorielle des Données Mixtes (AFDM) de Audigier *et al.* (2016b) permet d'imputer des données mixtes (catégorielles et numériques). Les approches d'ACP en présence de valeurs manquantes ont également été étendues au cadre de l'imputation multiple (voir section 5.2) par Josse et Husson (2012) pour l'ACP itérative et Audigier *et al.* (2015) pour l'ACP bayésienne.

Les méthodes factorielles qui prennent en compte les valeurs manquantes sont implémentées dans plusieurs packages R dont les principaux sont :

— **ade4** (Chessel *et al.*, 2004) qui permet l'analyse exploratoire de données écologiques et environnementales et propose une implémentation de NIPALS ;

— **missMDA** (Josse *et al.*, 2012) qui propose des implémentations de plusieurs méthodes d'analyse factorielle en présence de valeurs manquantes ;

 ISSN: 2102-6238

— **mixOmics** (Lê Cao *et al.*, 2009) qui propose des méthodes d'analyses multivariées pour l'exploration et l'intégration de données biologiques (en particulier les données 'omiques) et impute les valeurs manquantes avec l'approche NIPALS ;

— **pcaMethods** (Stacklies *et al.*, 2007) qui est un package Bioconductor[20] qui propose de nombreuses méthodes d'ACP en présence de valeurs manquantes (dont NIPALS, les méthodes d'ACP probabiliste et d'ACP bayésienne) ainsi que des outils pour la validation croisée et la visualisation des résultats.

Conclusion et recommandations :

— *Avantages* : bien adaptées à l'analyse exploratoire ; garanties théoriques fondées sur les modèles à effets fixes ou mixtes ; variantes adaptées à la grande dimension et au grand volume ;

— *Désavantages* : cadre théorique restreint aux modèles de génération des données fondés sur les modèles à effets fixes ou mixtes décrits plus haut : mêmes limitations que celles décrites dans la section 3.

4.5. Conclusions sur l'imputation simple

Dans cette section, nous avons présenté les principales méthodes d'imputation simple, en les catégorisant en trois grandes familles : complétion stationnaire, imputation fondée sur des similarités entre individus et méthodes de prédiction. Dans le cadre particulier des analyses factorielles, nous avons aussi présenté les approches développées spécifiquement pour ces cas-ci.

La complétion stationnaire est probablement l'approche la plus simple et la plus rapide. Pour ces raisons, elle peut apparaître comme très attractive. Cependant, même pour des taux de manquants relativement faibles, cette approche n'est pas recommandée car elle ignore les relations de corrélation entre variables et entre individus, elle sous-estime fortement la variabilité des variables imputées et en déforme leurs distributions.

Les méthodes qui utilisent une information de ressemblance entre individus (comme les approches hot-deck) sont particulièrement bien appropriées dans le cas de données discrètes (catégorielles ou numériques discrètes). D'une manière générale, toutefois, si elles préservent la distribution univariée des données, elles tendent à fortement déformer les corrélations entre variables. Dans le cas où le jeu de données contient des individus avec un grand nombre de valeurs manquantes, des individus entiers peuvent être utilisés pour imputer toutes les valeurs manquantes comme le suggèrent Voillet *et al.* (2016). Dans ce cas, elles permettent de mieux conserver les relations de corrélation entre variables et sont donc bien adaptées au cas où des analyses factorielles ou une inférence de réseaux sont réalisées après l'imputation comme dans Imbert *et al.* (2018). Toutefois, elles nécessitent de pouvoir obtenir une mesure de ressemblance ou une distance entre individus, ce qui peut être réalisé par l'utilisation de covariables complètement observée. Le choix de la distance et la nécessité d'avoir des données permettant de la calculer sont donc également deux limitations de la méthode.

Les approches d'imputation qui utilisent des méthodes de régression ou une modélisation jointe (comme les approches paramétriques multivariées de la section 3 ou les approches factorielles) sont généralement mieux adaptées pour la modélisation de la loi jointe des variables.

[20] https://www.bioconductor.org

 ISSN: 2102-6238

Elles sont plus difficiles à mettre en œuvre, en général, que les approches précédentes, nécessitent la définition correcte d'un modèle de loi jointe des données ou d'une méthode de régression dont la qualité de l'analyse dépend fortement. Dans le cas d'approches paramétriques, il est parfois possible d'obtenir une estimation de la variabilité du paramètre de la loi (voir section 5.3) et elles fournissent donc, par ce biais, une information sur l'incertitude liée à l'imputation.

Néanmoins, au sein d'un même jeu de données, il peut s'avérer utile d'utiliser une combinaison d'approches pour s'adapter au mieux aux spécificités de chaque variable ou chaque individu contenant des valeurs manquantes. La démarche standard consiste à commencer par une analyse exploratoire des valeurs manquantes puis, selon la distribution de celles-ci par variable et par individu, et les corrélations connues entre variables, à supprimer les variables et individus ayant un fort taux de manquants (s'ils sont peu nombreux) puis à combiner diverses méthodes d'imputation (par prédiction, par hot-deck, etc) selon la variable ou l'individu à imputer. Le package **simputation** permet de gérer facilement ce type d'approches en proposant une collection de méthodes standard pour l'analyse exploratoire des données manquantes et leur imputation. Enfin, il est recommandé de chercher à estimer l'incidence de l'imputation sur les analyses pratiquées *a posteriori*, par exemple en estimant l'incertitude liée à l'imputation (voir section 5). Des conseils pratiques détaillés sont fournis sur le site décrivant les grandes lignes directrices en matière de qualité dans le traitement des enquêtes de l'organisme public « Statistique Canada »[21] ainsi que par Fellegi et Holt (1976).

Enfin, l'imputation doit parfois être adaptée aux particularités du jeu de données. Par exemple, une approche pour l'imputation de variables ordinales est proposée dans Ferrari *et al.* (2011). Celle-ci alterne une ACP non linéaire et une imputation par *k*NN et est implémentée dans le package **ForImp**. Également, l'imputation de séries chronologiques peut être pratiquée en tenant compte de la tendance observée au cours du temps avec des approches par interpolation, par ajustement d'une courbe de lissage ou par estimation d'un modèle de régression longitudinale (ARIMA, par exemple, voir Kohn et Ansley, 1986). Les méthodes les plus courantes d'imputation de séries temporelles sont implémentées dans le package **imputeTS** (Moritz et Bartz-Beielstein, 2017) qui, à ce jour, est l'unique package d'imputation de données uniquement dédié aux séries temporelles. D'autres packages dont **zoo** (Zeileis et Grothendieck, 2005) et **forecast** incluent aussi des méthodes d'imputation pour les séries temporelles qui sont relativement sophistiquées. Également, les packages **spacetime** (Pebesma, 2012), **timeSeries** et **xts** incluent des approches plus basiques pour l'imputation de séries temporelles. Une comparaison des diverses méthodes d'imputation de séries temporelles est effectuée dans Moritz *et al.* (2015) qui montrent que les méthodes d'imputation les plus efficaces pour ce type de données sont fondées sur une prise en compte de la saisonnalité de la série temporelle.

5. Variabilité et fiabilité de l'imputation

Dans les méthodes d'imputation simple, il est fréquent qu'une valeur manquante soit remplacée par sa valeur imputée et qu'elle joue, dans la suite de l'analyse, le même rôle que les valeurs observées. Le risque est fort de biaiser ces analyses *a posteriori*, sans contrôle de l'incertitude liée à l'imputation. Par exemple, dans le cas de l'estimation d'un paramètre à partir des données, la

[21] https://www.statcan.gc.ca/pub/12-539-x/2009001/imputation-fra.htm

 ISSN: 2102-6238

variance du paramètre est souvent sous-estimée même si le modèle d'imputation est correctement spécifié (voir section 3).

On peut distinguer diverses approches pour aborder cette problématique : la première consiste à utiliser des outils diagnostiques destinés à évaluer la fiabilité de l'imputation. Cette question est discutée dans la section 5.1 et cherche à identifier des erreurs dans l'estimation de la valeur imputée par rapport à la valeur qui aurait dû être observée.

La seconde se concentre sur l'estimation de la variabilité liée au processus d'imputation. D'une part, elle fournit un diagnostic sur la fiabilité ou le domaine de validité des conclusions de l'analyse et, d'autre part, elle améliore la qualité de l'analyse elle-même (par des méthodes d'agrégation par exemple). Dans ce cadre, une approche fréquemment utilisée est l'*imputation multiple* que nous décrivons dans la section 5.2. La section 5.3 décrit les alternatives à cette approche dans le cadre particulier de l'algorithme EM et la section 5.4 conclut la section par une courte discussion sur ces diverses approches.

5.1. Outils de diagnostic

Les valeurs imputées étant des valeurs estimées, il est important de vérifier si elles sont plausibles. Pour cela, il est possible d'utiliser des outils de diagnostic. Cela consiste généralement à comparer les valeurs imputées aux valeurs observées soit à l'aide de graphiques, soit à l'aide de statistiques élémentaires.

5.1.1. Sur-imputation

La première approche pour évaluer la qualité d'une méthode d'imputation est de procéder par sur-imputation [22] en supprimant des données observées et en comparant les valeurs imputées aux valeurs réelles avant suppression, notamment par calcul de l'erreur quadratique moyenne (MSE) ou de sa racine carrée (RMSE), comme proposé dans les packages **Amelia** et **missMDA**. Cette approche est relativement intéressante pour évaluer la qualité d'une méthode donnée.

Une approche alternative consiste à utiliser uniquement valeurs observées et leur distribution pour évaluer la pertinence des valeurs imputées.

5.1.2. Outils généraux de diagnostic

De manière plus avancée et systématique, Abayomi *et al.* (2008) et Stuart *et al.* (2009) proposent trois types de diagnostic pour des données multivariées. La première approche consiste à représenter, de manière graphique, les données elles-mêmes (au travers, par exemple, de nuages de points) en différenciant valeurs observées et valeurs imputées. Ces graphiques permettent de repérer facilement des valeurs atypiques dans l'imputation, signe par exemple, d'un problème potentiel dans le choix de la méthode d'imputation.

La second type de diagnostic consiste à comparer, pour chaque variable, les densités entre valeurs imputées et celles observées en utilisant un test de Kolmogorov-Smirnov et en réalisant des graphiques diagnostiques (histogramme, courbe de densité, ...). Ceux-ci ont pour but de

[22] *Overimputation* en anglais.

 ISSN: 2102-6238

permettre, pour chaque variable, la comparaison visuelle entre les distributions des valeurs observées et les distributions des valeurs imputées. Les différences entre les valeurs imputées et observées ne sont pas forcément dues à un problème d'imputation. Il est possible qu'un sous-groupe de la population ait plus de données manquantes pour certaines variables. Ainsi, les graphiques diagnostiques permettent de mettre en évidence ces variables pour mieux les étudier.

Le dernier type de diagnostic utilise le fait que les imputations sont générées par des modèles ajustés sur les données observées. Il est donc possible de vérifier la qualité de l'ajustement de ces modèles en comparant la valeur prédite, pour un individu et une variable donnés, à la valeur observée ou bien en utilisant les outils diagnostiques spécifiques d'un modèle donné (graphique des résidus, QQ plot pour un modèle linéaire, par exemple). Ce type de diagnostic se rapproche de la sur-imputation dans la comparaison entre valeur observée et valeur prédite.

Enfin, de manière similaire, et au-delà du cas MAR, Simon et Simonoff (1986) étudient le cas de la régression linéaire multiple avec une covariable ayant des valeurs manquantes et proposent des formules explicites pour la dépendance entre le paramètre à estimer ou le coefficient de corrélation de la régression linéaire et les valeurs manquantes. Sous l'hypothèse d'une dépendance linéaire entre la variable contenant des manquants et les autres covariables, les auteurs proposent des graphiques permettant d'étudier l'effet potentiel des valeurs manquantes sur la régression qui peuvent être utilisées comme diagnostics pour évaluer la pertinence de l'imputation dans ce cadre-ci.

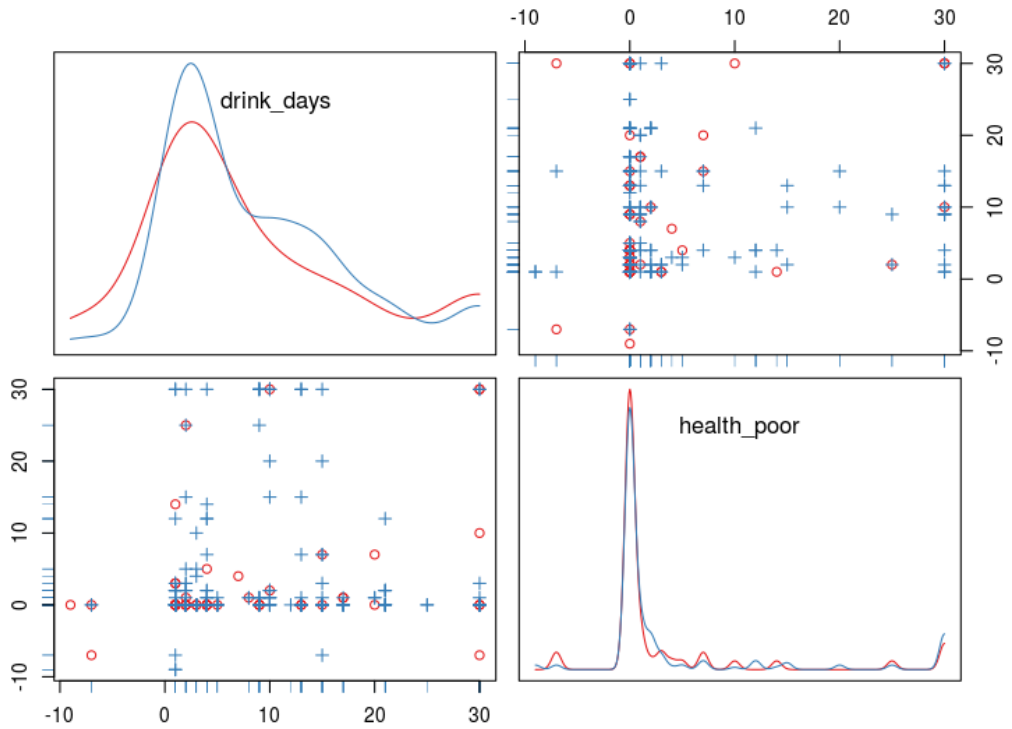

FIGURE 5. *Graphiques des distributions univariées (densités) des variables « drink_days » (en haut à gauche) et « health_poor » (en bas à droite) pour les valeurs manquantes (en rouge) ou observées (en bleu). Nuage de points des deux variables (en haut à droite et en bas à gauche).*

 ISSN: 2102-6238

Les packages **mi** et **VIM** proposent différents graphiques diagnostiques. Pour comparer les distributions, le package **VIM** fournit divers graphiques uni et bi-variés représentant de manière séparée ou simultanée les valeurs observées et les valeurs imputées. Par exemple, pour les variables « drink_days » (nombre de jours, au cours du dernier mois, où la personne a bu au moins un verre d'alcool) et « health_poor » (nombre de jours, au cours du dernier mois, où la personne n'a pu pratiquer une activité « habituelle » à cause de problèmes de santé), la figure 5 montre les distributions univariées des valeurs imputées et observées pour les deux variables et un nuage de points sur lequel les points correspondant à au moins une valeur imputée sont mis en valeur par une couleur distincte. Les densités des valeurs imputées et observées sont similaires et aucune répartition spécifique des points correspondant à des valeurs imputées n'est repérable sur le nuage de points, ce qui est un indicateur positif de la fiabilité de l'imputation. Le package **mi** utilise l'approche d'imputation FCS décrite dans la section 4.3 et fournit un graphique contenant distribution du tableau de données imputées et observées (par un histogramme) et graphiques comparant valeurs prédites et résidus aux valeurs observées (voir figure 6 pour la variable « weight_lbs »).

FIGURE 6. *Exemple de graphiques diagnostiques fourni par le package* **mi** *(pour la variable « weight_lbs ») : histogramme des valeurs observées et imputées, valeurs imputées (prédites) et résidus en fonction des valeurs observées.*

5.1.3. Erreur d'imputation et décomposition dans le cas des k-plus proches voisins

Comme indiqué dans la section 5.1.1, l'estimation de l'erreur d'imputation est souvent limitée à la comparaison entre valeurs observées et valeurs imputées. Dans Stage et Crookston (2007), les auteurs vont au-delà et proposent de décomposer l'erreur d'imputation en :

— *erreur de mesure*, qui est l'erreur commise entre les valeurs observées, y_{ij} et la « vraie » valeur de Y_j pour l'individu i, y^*_{ij} (qui reste inconnue en raison d'erreurs liées aux appareils de mesure ou bien de différences expérimentales incontrôlées entre les mesures par exemple). Contrairement au cadre habituel (qui suppose cette erreur nulle), le cadre de l'article de Stage et Crookston (2007) est celui d'erreurs de mesure non nulles mais qui ne présentent pas de biais et qui sont indépendantes de covariables complètement observées, X ;

— et *erreur pure* (qui peut être vue comme une erreur du modèle d'imputation) qui est spécifiée dans le cadre d'une approche d'imputation dans laquelle la variable avec des valeurs manquantes Y_j est imputée à partir d'un modèle faisant uniquement intervenir des covariables complètement observées X. Dans ce cadre-ci, l'erreur pure s'écrit :

$$y^*_{ij} - g_j(\mathbf{x}_i)$$

 ISSN: 2102-6238

où g_j est la fonction de prédiction permettant l'imputation de la valeur de Y_j. C'est cette erreur qui est d'intérêt pour diagnostiquer la méthode d'imputation choisie.

Dans le cadre de l'imputation par la méthode kNN et lorsque $k = 1$, ils montrent que l'on peut estimer l'erreur d'imputation pour la variable à imputer Y_j, à partir de la différence d'erreur quadratique moyenne (MSD) :

$$\mathrm{MSD}_j = \frac{\sum_{i=1}^{n} r_{ij} \left(y_{ij} - y_{\mathcal{N}_1(i),j}\right)^2}{\sum_{i=1}^{n} r_{ij}}$$

où $\mathcal{N}_1(i)$ est le plus proche voisin de i, parmi les individus pour lesquels Y_j est observée, au sens de la distance sur X comme définie dans l'équation (3). Enfin, ils proposent d'estimer l'*erreur standard d'imputation* (SEI) par

$$\mathrm{SEI}_j^2 = \mathrm{MSD}_j - \frac{1}{2}\mathrm{MMSD}(0)_j$$

où $\mathrm{MMSD}(0)_j$ est la valeur de MSD obtenue pour une petite fraction des paires d'individus ayant les plus petites distances entre eux, non pas au sens de l'équation (3) mais au sens de la distance de Mahalanobis (ces paires étant utilisées pour estimer l'erreur de mesure).

Cette proposition est généralisée aux cas où $k > 1$ en utilisant la valeur moyenne des kNN. Ces erreurs diagnostiques sont proposées dans le package R **yaImpute** (Crookston et Finley, 2008).

5.2. Imputation multiple

Pour tenter de mesurer l'impact de l'imputation et pour quantifier l'erreur commise lors de celle-ci, l'approche la plus répandue consiste à répéter l'imputation plusieurs fois en introduisant de l'aléa. Ces approches sont connues sous le nom d'imputation multiple.

5.2.1. Principe de l'imputation multiple

L'imputation multiple (Rubin, 1987, Rubin, 2012 et Schafer, 1999) consiste à proposer, pour chaque valeur manquante, non pas une mais plusieurs valeurs plausibles pour l'imputation. Cette méthode permet de mesurer la variabilité, sur le résultat final, du processus d'imputation.

L'imputation multiple se déroule en trois phases, représentées sur la figure 7 :

Phase d'imputation Le tableau de données initiales est dupliqué M fois et un modèle d'imputation est appliqué sur chaque nouveau tableau de données. Une part d'aléa est introduite, soit au niveau de la duplication du tableau initial (qui n'est pas reproduit à l'identique), soit au niveau de l'imputation elle-même, ce qui permet l'obtention de M tableaux différents de données complètes ;

Phase d'analyses statistiques L'analyse statistique retenue (régression, ACP, inférence de réseau, ...) pour analyser le tableau de données est mise en œuvre sur chacun des $m = 1, \ldots, M$ tableaux de données imputées pour obtenir M estimations ;

 ISSN: 2102-6238

Phase d'analyse combinée Les M résultats obtenus sont combinés selon les règles définies par Rubin (1987) pour obtenir une seule estimation finale ou pour estimer la variabilité des résultats par une analyse statistique cible pratiquée sur les données complétées.
Les procédures d'imputation qui incorporent une variabilité appropriée à travers les M jeux de données imputées dans le modèle sont dites « adéquates [23] » au sens de Rubin (1987) ou Little et Rubin (2002) : cela signifie que ces méthodes d'imputation reflètent correctement la variabilité de la méthode fondée sur les données imputées, en prenant en compte, à la fois, la variabilité intra-imputation (correspondant à la variabilité due à la méthode elle-même et au bruit dans les données) et la variabilité inter-imputation (attribuable à la présence de données manquantes).

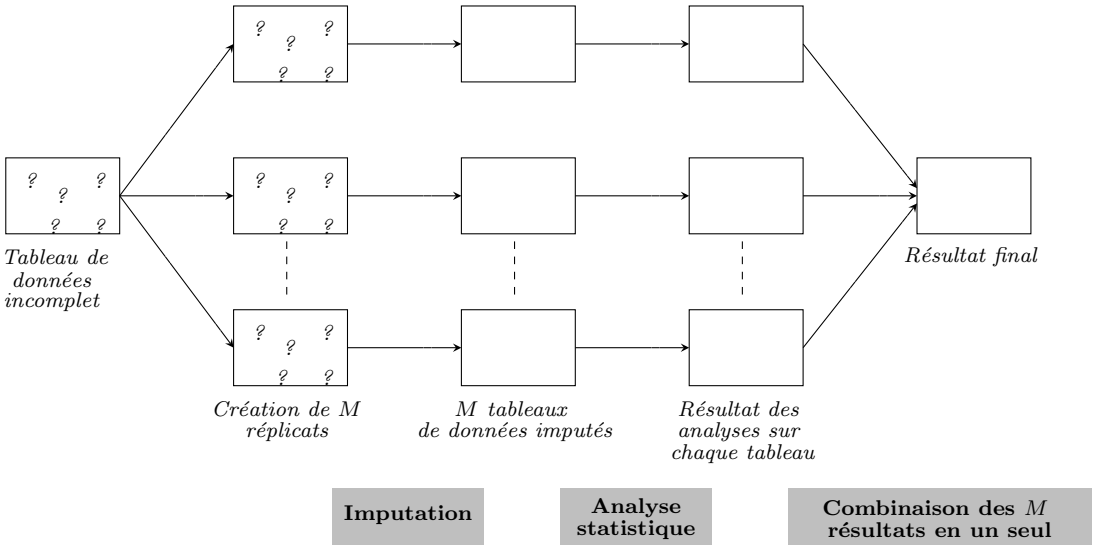

FIGURE 7. *Schéma de l'imputation multiple*

Ces diverses étapes et les approches principales pour leurs mises en œuvre sont décrites dans les sections suivantes.

5.2.2. Phase d'imputation

Plusieurs approches permettent d'obtenir des tableaux de données imputées différents, fondées soit sur des perturbations de l'échantillon initial, soit sur l'introduction d'un processus aléatoire dans l'imputation elle-même.

Approche par ré-échantillonnage Dans les approches par ré-échantillonnage, l'aléa est introduit au moment de la duplication du tableau de données initiales en M copies. Au lieu de

[23] *proper* en anglais.

 ISSN: 2102-6238

dupliquer le tableau initial, un sous-échantillonnage ou un ré-échantillonnage sont pratiqués pour obtenir M copies « perturbées » du tableau de données initiales. En pratique, les approches bootstrap ou bien Jackknife (avec $M = n$) sont les plus utilisées. L'analyse statistique conduit donc, par exemple, à l'estimation d'un paramètre θ par M valeurs $\hat{\theta}^{(m)}$ (pour $m = 1, \ldots, M$) qui sont les estimations obtenues par la méthode statistique cible à partir d'un tableau obtenu par ré-échantillonnage ou sous-échantillonnage puis imputation. L'approche par bootstrap est, par exemple, utilisée par Josse *et al.* (2011) et Audigier *et al.* (2016a) pour estimer la variabilité de la position d'un individu avec des valeurs manquantes dans l'ACP ou l'ACM. Une approche d'imputation multiple Jackknife pour estimer la variabilité d'un estimateur dans le cadre de l'imputation hot-deck est décrite par Burns (1990) : Rao et Shao (1992) montrent toutefois que celle-ci peut entraîner des biais importants et proposent une alternative fondée sur un estimateur Jacknife corrigé qui n'utilise qu'une imputation simple.

Approche de type « hot-deck » Les approches de type « hot-deck » (voir section 4.2.2) conduisent à la création, pour chaque valeur manquante, d'un ensemble de « donneurs » correspondant à un ensemble de valeurs plausibles pour la valeur manquante considérée. En effectuant un tirage aléatoire dans ce ensemble de donneurs pour chaque valeur manquante, M tableaux de données imputées différents sont obtenus (Cranmer et Gill, 2012).

Approche bayésienne Dans les approches bayésiennes (section 3.2), la phase d'imputation finale est fondée sur un échantillonnage selon la loi $f(Y_{\text{miss}}|Y_{\text{obs}}, \theta^{(T)})$ où T est le nombre d'itérations de l'algorithme et $\theta^{(T)}$ l'estimation courante du paramètre qui régit la loi jointe, θ. Il est donc possible d'utiliser cette approche pour générer M tableaux de données imputées différents. Cette approche est utilisée dans van Buuren et Groothuis-Oudshoorn (2011) et Su *et al.* (2011) pour une imputation fondée sur une méthode FCS (voir section 4.3) et par Audigier *et al.* (2015) pour une imputation multiple par ACP bayésienne.

5.2.3. *Combiner les résultats : cas de l'estimation d'une quantité numérique α et estimation de la variance de l'estimation*

Lorsque le but de l'analyse statistique est l'estimation d'une quantité numérique α, l'approche la plus fréquente pour combiner les résultats des M analyses statistiques après imputation est le simple calcul de l'estimateur moyen $\bar{\alpha}$ (Little et Rubin, 2002) :

$$\bar{\alpha} = \frac{1}{M} \sum_{m=1}^{M} \hat{\alpha}^{(m)}.$$

Dans le cas d'une approche par imputation multiple fondée sur le Jackknife, l'approche standard consiste à imputer le jeu de données entier avec une approche quelconque puis à obtenir $M = n$ estimateurs $\hat{\alpha}^{(m)}$ à partir des échantillons imputés correspondants aux individus $\{1, \ldots, n\} \setminus \{m\}$. L'estimation de α est alors réalisée de manière standard pour les approches Jackknife, en calculant la moyenne des pseudo-valeurs

$$\hat{\alpha} = \hat{\alpha}^{(0)} + (n-1)(\hat{\alpha}^{(0)} - \bar{\alpha}), \tag{8}$$

 ISSN: 2102-6238

où $\hat{\alpha}^{(0)}$ est l'estimateur de α obtenu à partir de l'échantillon entier après imputation (Rubin, 1987).

La variance de l'estimateur $\bar{\alpha}$ est, quant à elle, obtenue par

$$\mathrm{Var}(\bar{\alpha}) = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\mathrm{Var}(\hat{\alpha}^{(m)})}_{\text{variance intra-imputation:}W} + \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\alpha}^{(m)}-\bar{\alpha})^2}_{\text{variance inter-imputation:}B}$$

où B s'obtient directement à partir des m estimateurs $\hat{\alpha}^m$ et W dépend de la méthode employée pour obtenir cet estimateur (classiquement, par exemple, lorsque $\hat{\alpha}^{(m)}$ est une moyenne empirique, W s'obtient à partir des M variances empiriques des observations des M tableaux de données imputées). L'approximation de la variance peut être améliorée en multipliant B par $(1+\frac{1}{M})$ afin de prendre en compte le fait que les estimations de α ne sont que des approximations obtenues pour un nombre fini de tableaux, M : une variabilité supplémentaire, correspondant à l'erreur de simulation, peut être ajoutée et la variance totale de $\bar{\alpha}$ est alors estimée par

$$W + \frac{M+1}{M}B.$$

Dans le cas où l'imputation multiple est réalisée avec une approche bootstrap ou Jackknife, on peut aussi obtenir une estimation de la variance de l'estimateur sans avoir besoin d'un estimateur de $\mathrm{Var}(\hat{\alpha}^{(m)})$, en utilisant les échantillons dit « *out-of-bag* » (non sélectionnés dans l'échantillon bootstrap courant, pour l'approche bootstrap) ou bien par

$$\frac{1}{n(n-1)}\sum_{m=1}^{n}(\tilde{\alpha}^{(m)}-\hat{\alpha})^2,$$

avec $\tilde{\alpha}^{(m)} = n\hat{\alpha}^{(0)} - (n-1)\hat{\alpha}^{(m)}$ et les autres notations comme dans l'équation (8), pour l'approche Jackknife.

5.2.4. Autres approches pour la combinaison

Les approches décrites dans la section précédente ne permettent la combinaison des résultats que dans le cadre de l'estimation d'une quantité numérique. Lorsque les analyses statistiques pratiquées sur les M tableaux de données imputées produisent des résultats sous une forme plus complexe, d'autres approches peuvent être mises en œuvre soit pour visualiser la variabilité due à l'imputation, soit pour combiner les résultats.

Josse *et al.* (2012) proposent l'utilisation de l'imputation multiple en ACP pour obtenir des ellipses de confiance (sous hypothèse de distribution gaussienne) autour de la projection des individus dans l'ACP. Pour cela, une projection de référence est obtenue par ACP itérative et les résultats d'imputations multiples sont utilisées pour représenter les individus imputés comme individus supplémentaires, permettant ainsi l'estimation des contours des ellipses de confiance.

Lorsque le but de l'imputation multiple n'est pas seulement l'estimation de la variabilité de l'imputation mais aussi la définition d'un résultat « combiné » obtenu à partir de plusieurs imputations, diverses stratégies alternatives au calcul de la moyenne sont proposées : dans le cadre

 ISSN: 2102-6238

d'analyses factorielles, Voillet *et al.* (2016) proposent d'utiliser la méthode STATIS (Lavit *et al.*, 1994) pour combiner les différentes configurations obtenues lors d'une AFM (Multiple Factor Analysis, Escofier et Pagès (1994)) réalisée par imputation multiple : cette approche recherche une projection consensuelle, c'est-à-dire une projection la plus corrélée aux M projections obtenues à partir des données imputées. Enfin, Imbert *et al.* (2018) proposent une approche fondée sur l'analyse de la fréquence de prédiction d'une arête dans le cas où l'analyse statistique est une inférence de réseau : cette approche permet de ne conserver que les arêtes dont la prédiction est peu affectée par la valeur imputée et, ainsi, de diminuer le taux de faux positifs dans l'inférence.

5.2.5. Packages R

Divers packages proposent des implémentations pour effectuer des imputations multiples avec des approches différentes pour la partie imputation :

- **Amelia** propose une méthode d'imputation multiple fondée sur une approche par modélisation jointe gaussienne (estimée par EM ou par approche bayésienne), combinée à une imputation multiple par bootstrap (dans le cadre EM) ou bayésienne ;
- **hot.deck** propose une version multiple de l'imputation hot-deck fondée sur le score d'affinité proposé par Cranmer et Gill (2012) ;
- **jomo** et **pan** sont deux packages qui proposent de nombreux modèles d'imputation par modélisation jointe (approches bayésiennes) dans un cadre d'imputation multiple dit « multiniveaux », c'est-à-dire lorsque les individus sont stratifiés en classes ;
- **mi** propose des méthodes d'imputation multiple avec une approche dite par « équations chaînées », qui est une approche bayésienne fondée sur la méthode FCS (voir section 4.3). Le package contient un grand nombre de modèles pour variables numériques ou catégorielles, des approches par injection de bruit pour limiter les problèmes dûs aux colinéarités entre variables et propose également divers outils de diagnostic pour évaluer la fiabilité du modèle choisi ;
- **mice** est un des packages les plus utilisés pour l'imputation multiple. L'introduction de l'aléa dans l'imputation est réalisée via l'approche par équations chaînées (comme **mi**). Le package permet de traiter des variables de types variés (catégorielles ou numériques) et contient plusieurs outils diagnostiques ;
- **missMDA** propose des méthodes pour l'imputation multiple en analyse factorielle, soit par modélisation bayésienne, soit par approche bootstrap. L'imputation multiple est utilisée ici pour visualiser la variabilité de la projection sur les axes de l'ACP ou de l'AFM obtenus par imputation simple (ACP itérative) ou pour générer des valeurs multiples d'imputation par ACP (section 4.4) ;
- **mitools** permet de combiner des résultats d'imputations multiples de manière générique en aggrégeant n'importe quel résultat obtenu en combinant plusieurs imputations obtenues par ailleurs ;
- **MixedDataImpute** et **NPBayesInput** sont deux packages proposant des approches de modélisation jointe (approches bayésiennes) pour l'imputation, respetivement, de variables catégorielles et mixtes.

 ISSN: 2102-6238

5.3. *Estimation de l'incertitude dans les modèles EM*

L'approche précédente est fréquemment utilisée pour estimer l'erreur quadratique moyenne du paramètre θ dans les modèles d'imputation EM. Cependant, dans ce cas particulier, une alternative, moins coûteuse en temps de calcul, est proposée dans Meng et Rubin (1991) sous le nom de SEM [24].

Le principe de la méthode consiste à exprimer l'erreur quadratique moyenne de θ en fonction de deux quantités facilement estimable : l'erreur quadratique moyenne de θ sur les données observées et le taux de convergence de l'algorithme EM (qui est la différentielle de la fonction d'évolution de l'estimation du paramètre au cours de l'algorithme EM). Cette remarque permet d'obtenir directement l'erreur quadratique moyenne de θ au cours de l'algorithme EM.

5.4. *Discussion*

L'évaluation de l'incertitude liée à l'imputation est une phase importante pour évaluer la fiabilité des résultats d'une étude. Cette incertitude a diverses composantes, comme le soulignent Stage et Crookston (2007) : l'erreur standard du paramètre estimé ou de la valeur imputée est liée, d'une part, à l'incertitude existant sur les données observées et, d'autre part, à la part d'incertitude provenant de l'imputation elle-même. Dans la plupart des cas, ces deux composantes sont confondues et l'erreur globale est estimée.

Dans les approches EM, l'imputation est prise en charge par une hypothèse paramétrique nécessitant l'estimation d'un paramètre θ. L'incertitude liée à l'imputation est donc directement liée à la valeur de ce paramètre et à son erreur standard. Toutefois, cette dernière n'est obtenue directement que dans la méthode FIML et les autres approches ML requièrent l'insertion d'une étape supplémentaire dans la méthode (SEM ou bien approches par ré-échantillonnage) pour fournir une estimation de l'erreur standard sur l'estimation de θ. Toutefois, ces approches nécessitent d'avoir une taille d'échantillon assez élevée : dans le cas contraire, il est fréquent d'avoir recours à une approche bayésienne.

Enfin, la principale limite des approches EM est qu'elles nécessitent des hypothèses paramétriques et l'adaptation de l'approche pour chaque cadre d'hypothèses. Aussi, l'imputation multiple constitue-t-elle un cadre plus simple pour l'estimation de l'incertitude liée à l'imputation. Dans le cadre standard de l'estimation d'une quantité numérique, la combinaison des différents résultats se fait de manière naturelle par un simple calcul de moyenne même s'il peut être plus compliqué de trouver des règles de combinaison des résultats satisfaisants les propriétés préconisées dans Little et Rubin (2002) pour des analyses plus complexes. Toutefois, dans le cadre de l'inférence statistique, la supériorité, en terme de puissance statistique, de l'approche EM (en particulier FIML) sur l'imputation multiple est fréquemment soulignée (Collins *et al.*, 2007, Schafer et Graham, 2002, Graham *et al.*, 2007 et Dong et Peng, 2013).

6. Prendre en compte les données manquantes informatives (MNAR)

La plupart des approches présentées dans cette revue et implémentées dans les packages R sont fondées sur l'hypothèse implicite que les données sont manquantes de type MAR. En pratique,

[24] *Supplemental EM*, en anglais

 ISSN: 2102-6238

cette hypothèse est souvent abusive, particulièrement dans le cas de sondages portant sur des questions sensibles ou d'études cliniques longitudinales (dans lesquelles des patients peuvent sortir de l'étude pour des raisons liées aux variables d'intérêts mesurées : cette question est donc liée à la thématique des données censurées).

Lorsque les données sont manquantes de type MNAR, la loi de Y_{miss} n'est pas indépendante de la loi de R. Dans ce cas, les approches habituelles de traitement des données manquantes (qui consistent à estimer la loi multivariée $f(Y;\theta)$ à partir des données observées puis à utiliser cette loi pour l'inférence ou l'imputation) produisent des estimateurs ou des valeurs imputées biaisés.

Dans ce cas, l'estimation de la distribution jointe des données et de la probabilité d'absence, $f(Y,R;\theta,\psi)$ (ou $f(X,Y,R;\theta,\psi)$ si des covariables complètement observées sont disponibles), est la clé pour aborder cette question. Une approche courante consiste à proposer une factorisation réaliste de cette loi jointe qui soit estimable à partir des observations (Little, 1995). On distingue, en particulier, deux approches principales : les modèles de sélection [25] (Heckman (1976) et Diggle et Kenward (1994), section 6.1) et les modèles par mélange de profils [26] (Rubin, 1977) (section 6.2). Une troisième approche consiste à estimer les dépendances entre Y et R au moyen de variables latentes aléatoires : ce sont les modèles à paramètres partagés [27] (Little (1995) et Hogan et Laird (1997), section 6.3).

6.1. Modèles de sélection

Dans l'approche par modèle de sélection, la factorisation suivante de la loi jointe est utilisée :

$$f(Y,R;\theta,\psi) = f(Y|\theta)f(R|Y;\psi).$$

Cette factorisation est intuitive car elle modélise directement la distribution d'intérêt en utilisant la probabilité d'absence d'une donnée conditionnellement aux variables d'intérêt Y.

Un exemple typique est le modèle de Heckman (1979), dans lequel les valeurs d'une variable Y_j sont expliquées par

$$Y_j = X^\top\theta + \varepsilon, \tag{9}$$

où les erreurs ε sont indépendantes de X et suivent une loi gaussienne centrée de variance σ^2. La probabilité d'absence d'une valeur, R, conditionnellement à (X,Y_j) est, dans une première étape, estimée à l'aide (par exemple) d'un modèle PROBIT puis l'espérance conditionnelle $\mathbb{E}(Y|R=1)$, obtenue à partir de cette estimation, est utilisée comme variable explicative supplémentaire dans le modèle de régression de l'équation (9).

Des variantes de cette approche existent qui rentrent dans le cadre du modèle de sélection : par exemple, la méthode décrite dans l'article de Diggle et Kenward (1994) est une extension du modèle de Heckman au cas multivarié et Robins *et al.* (1995) et Rotnitzky *et al.* (1998) proposent des versions semi-paramétriques de ces approches pour la distribution des données complètes $f(Y;\theta)$ et les appliquent pour l'analyse des résultats d'un sondage sur le SIDA.

Une limite de ces approches est qu'elles sont souvent fondées sur des hypothèses paramétriques assez fortes, en particulier sur la spécification du modèle permettant d'obtenir $f(R|Y;\psi)$.

[25] *Selection model* en anglais.
[26] *Pattern mixture model* en anglais.
[27] *Shared-parameter model* en anglais.

 ISSN: 2102-6238

6.2. *Modèles de mélange de profils*

Comme les modèles de sélection, les modèles de mélange de profils utilisent une factorisation de la loi jointe $f(Y,R;\theta,\psi)$ pour estimer celle-ci. Dans ce cas-ci, la factorisation utilisée est

$$f(Y,R;\theta,\psi) = f(Y|R;\theta)f(R;\psi).$$

De manière concrète, la distribution conditionnelle décrit des profils distincts d'individus partageant le même profil de valeurs manquantes. Des sous-groupes d'individus, contenant les mêmes variables manquantes et observées, sont donc créés dans une première étape et dans chaque sous-groupe, la distribution, $f(Y|R;\theta)$, est estimée.

Les modèles de mélange de profils sont, par construction, sous-identifiés car, par définition des profils, certaines variables de $f(Y|R;\theta)$ sont toujours manquantes. Little (1993) propose, pour résoudre ce problème, d'utiliser des restrictions identificatrices, c'est-à-dire des contraintes sur les paramètres inestimables de $f(Y|R;\theta)$ pour les profils incomplets. Différentes restrictions sont proposées, comme par exemple :

— valeurs manquantes des cas complets (CCMV [28]) (Little, 1993) : le paramètre θ de $f(Y|R;\theta)$ est estimé pour le profil des cas complets et supposé identique pour tous les autres profils ;

— valeurs manquantes des cas disponibles (ACMV [29]) (Molenberghs *et al.*, 1998) : cette approche étend le cas précédent en estimant tous les paramètres estimables de θ directement dans chacun des profils et fixe les autres paramètres non estimables en utilisant un ordonnancement naturel (par exemple, dans le cas de données longitudinales) sur les différents profils.

Dans le cadre d'applications à l'analyse de données de qualité de vie chez des patientes atteintes du cancer du sein (qui sont des données censurées), Thijs *et al.* (2002) proposent une alternative aux restrictions identificatrices via des simplifications de modèle qui consistent à diminuer le nombre de paramètres à estimer. Ce principe est illustré par la description d'une stratégie d'estimation hiérarchique des lois dans les profils $f(Y|R;\theta)$ qui s'appuie sur la structuration longitudinale des variables.

6.3. *Modèles à paramètres partagés*

Dans les modèles à paramètres partagés, des variables aléatoires additionnelles, B, non observées, sont introduites pour modéliser la dépendance entre Y et R, qui sont alors supposées indépendantes sachant B. Dans ce cas, on a alors

$$f(Y,R|B;\theta,\psi) = f(Y|B;\theta)f(R|B;\psi)$$

et, par conséquent,

$$f(Y,R;\theta,\psi) = \int f(Y|B=b;\theta)f(R|B=b;\psi)f(b)db.$$

[28] *Complete Case Missing Value*, en anglais.
[29] *Available Case Missing Value*, en anglais.

 ISSN: 2102-6238

La stratégie standard consiste à faire une hypothèse paramétrique sur la distribution des effets aléatoires B. Un des premiers modèles à effets partagés a été proposé par Wu et Carroll (1988) qui ont introduit cette approche dans le cadre de données longitudinales gaussiennes. $f(Y|B=b,\theta)$ est modélisé comme un modèle linéaire avec effet aléatoire qui est combiné à $f(R|B=b,\psi)$, modèle PROBIT ou logistique à effet aléatoire.

Little (1995) explique que les modèles à paramètres partagés peuvent être considérés comme des modèles de sélection à coefficients aléatoires [30] via la factorisation suivante :

$$f(Y,R,B;\theta,\psi) = f(Y|B;\theta)f(R|Y,B;\psi)f(B)$$

et comme des modèles de mélange de profils à coefficients aléatoires [31], via la factorisation suivante :

$$f(Y,R,B;\theta,\psi) = f(Y|R,B;\theta)f(R|B;\psi)f(B).$$

Des extensions de cette approche sont proposées dans Follmann et Wu (1995) qui développent un modèle pour des réponses binaires dans le cadre d'une étude longitudinale et dans Albert et Follmann (2000) qui étendent l'approche initiale à l'analyse de données de comptage longitudinales. Gad et Darwish (2013) proposent également l'extension de l'algorithme EM stochastique pour estimer les paramètres du modèle à paramètres partagés. Ils y ajoutent une étape supplémentaire pour obtenir une erreur standard sur cette estimation.

6.4. Limites de ces approches

Le modèle de sélection est fondé sur des hypothèses paramétriques sur $f(R|Y;\psi)$. Cette particularité le rend sensible à une mauvaise spécification de cette loi. Bien que ne reposant pas sur des hypothèses explicites de paramétrage d'une distribution, les modèles de mélanges de profils sont aussi très sensibles aux hypothèses de restriction, qui ne sont pas vérifiables. Par ailleurs, un compromis est à effectuer pour déterminer un nombre de profils de données manquantes adéquat : en effet, un grand nombre de profils améliore la précision du modèle mais en augmentant le nombre de paramètres à estimer et donc en détériorant la qualité de l'estimation de chacun de ces paramètres. Enfin, dans cette approche, la loi marginale de Y n'est pas disponible directement (les paramètres de cette loi sont estimés conditionnellement à un profil donné). Estimer cette loi nécessite donc une marginalisation par rapport aux profils de données manquantes :

$$f(Y;\theta) = \sum_R f(Y|R;\theta_R)f(R;\psi).$$

Ces deux types d'approches sont plus adaptés au cas où la non réponse est directement liée aux variables observées (comme dans l'exemple d'un questionnaire portant sur des réponses sensibles). Par contre, lorsque l'absence d'une donnée est attribuable à un processus sous-jacent, par exemple la progression d'une maladie, il est préférable d'utiliser un modèle à paramètres partagés qui pourra prendre en compte ce processus à l'aide des effets aléatoires B. C'est le cas, par exemple, en présence de données censurées (Little, 1995).

[30] *Random-coefficient selection model*, en anglais.
[31] *Random-coefficient pattern-mixture model*, en anglais.

 ISSN: 2102-6238

6.5. Analyse de sensibilité

Les approches décrites précédemment sont fondées sur des hypothèses invérifiables sur le lien entre le processus de données manquantes et le processus d'intérêt. Verbeke *et al.* (2001) et Thijs *et al.* (2002) proposent une approche par analyse de sensibilité fondée sur une perturbation des données en direction de l'hypothèse MNAR pour vérifier la pertinence du modèle MAR. L'idée principale est de comparer les résultats obtenus sous ces deux hypothèses pour analyser la sensibilité des résultats à l'hypothèse MNAR.

Il existe différentes manières d'effectuer une analyse de sensibilité en présence de données manquantes. Une analyse de sensibilité relativement simple consiste à étudier les résultats de différents jeux de données imputés issus de modèles d'imputation différents. Ce principe est proposé dans le package **mice** qui met en place un certain nombre de scénarios plausibles et permet d'examiner les conséquences de chacun d'entre eux sur l'inférence finale. Dans le cas où l'hypothèse MAR semble violée, les auteurs proposent de multiplier les imputations par un facteur ou de leur ajouter une valeur fixe, les deux approches étant des formes basiques de modèles à mélange de profils.

Certaines méthodes utilisées pour imputer les données MNAR peuvent également être employées pour effectuer une analyse de sensibilité. Verbeke *et al.* (2001) proposent ainsi d'utiliser les modèles à mélange de profils pour l'analyse de sensibilité. Thijs *et al.* (2002) utilisent cette approche en comparant les résultats obtenus avec chacune des restrictions identificatrices possibles : cet ensemble de conclusions fournit ainsi un aperçu de la sensibilité aux hypothèses émises. Ce type d'approches peut donc s'avérer une première étape très utile pour détecter des évidences en faveur de l'hypothèse MNAR et trouver la stratégie qui semble la plus adéquate à leur prise en compte.

Enfin, notons que, si quelques approches et modèles permettent d'identifier et de prendre en compte les valeurs manquantes MNAR, une limite forte de celles-ci est l'absence d'implémentations dans les outils habituels de traitement des données manquantes. À notre connaissance, par exemple, aucun package R ne propose d'implémentation des modèles décrits plus hauts ni des approches d'analyse de sensibilité qui permettent de les évaluer.

7. Conclusion

Les données manquantes sont un problème fréquemment rencontré dans les analyses statistiques, quel que soit le domaine d'étude. La méthode la plus adéquate pour en tenir compte dépend de paramètres multiples comme la typologie des valeurs manquantes, le type de mécanisme qui a conduit à leur génération, leur distribution dans le jeu de données ainsi que les attentes de l'utilisateur en terme d'analyses statistiques. On peut toutefois dégager des recommandations générales en plusieurs étapes :

— la première étape consiste à décrire les données manquantes afin d'émettre des hypothèses sur le mécanisme des données manquantes. Ces hypothèses doivent guider le choix de la stratégie à utiliser pour les traiter, conduire à supprimer des données (individus ou variables) ou bien à compléter simplement certaines valeurs manquantes dont on a identifié l'origine (Fellegi et Holt, 1976) ;

 ISSN: 2102-6238

— lorsque le but de l'analyse statistique est l'inférence et que les données manquantes sont supposées MAR, les approches EM et bayésienne fournissent des estimations non biaisées pour lesquelles il est possible d'obtenir une bonne estimation des erreurs standards.
Dans d'autres cas d'analyses statistiques, les approches d'imputation multiple, qui permettent d'estimer la variabilité liée à l'imputation tout en fournissant un ou des tableaux de données complets, sont recommandées. Selon les hypothèses sur la distribution multivariée des données et selon le type d'analyse à effectuer *a posteriori*, ces imputations multiples pourront être basées sur des approches hot-deck, des approches par prédiction, des approches factorielles ou des approches bayésiennes. Confronter et comparer différents types d'imputation, notamment par analyse de sensibilité, peut permettre d'identifier les limites liées à chaque approche sur un cas d'application donné.
En revanche, si les données sont MNAR, ce qui est particulièrement fréquent dans le cas des études longitudinales, l'imputation doit alors être fondée sur des modèles spécifiques à ce type de données ;

— la dernière étape consiste à essayer d'obtenir une évaluation de la qualité de l'imputation ou de l'estimation statistique, soit en utilisant des outils ou des caractéristiques numériques diagnostiques, soit en procédant par analyse de sensibilité. En particulier, les hypothèses MAR/MNAR étant impossibles à vérifier par définition, il semble judicieux de systématiquement effectuer une analyse de sensibilité des résultats d'imputations sous hypothèses MAR/MNAR en cas de doute (lorsque la distribution des valeurs manquantes n'est pas homogène, par exemple). Toutefois, ces approches ne sont pas, à notre connaissance, implémentées dans les packages R actuellement disponibles.

Cette revue fournit un panorama des grandes familles de méthodes pouvant prendre en compte les données manquantes lors d'analyses statistiques. Nous nous sommes attachées à décrire des solutions logicielles disponibles pour utiliser ces méthodes, en listant les divers packages R dans lesquels elles sont implémentées. Des tableaux récapitulant les différentes méthodes et les packages R associés sont fournis après cette conclusion, organisés de la même manière que les sections de cet article (analyse descriptive, utilisation des données observées, inférence, imputation simple, variabilité liée à l'imputation). La liste des packages ne prétend pas à l'exhaustivité mais propose un panorama réaliste des packages utilisables pour mettre en œuvre une approche donnée.

 ISSN: 2102-6238

TABLE 1. *Packages permettant l'analyse descriptive des données manquantes*

Méthodes	**Packages R**	**Cadre d'application**
Identification de motifs de données manquantes	**mi** (Su *et al.*, 2011)	tableaux de données mixtes
Description des données manquantes	**naniar** ; **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016)	tableaux de données mixtes
Test MAR/MCAR	**BaylorEdPsych** ; **missMech** (Jamshidian *et al.*, 2014)	numériques et catégorielles

TABLE 2. *Récapitulatif des méthodes fondées uniquement sur les données observées*

Méthodes	**Packages R**	**Cadre d'application**
Analyse des cas complets	option disponible dans de nombreuses fonctions : `na.action=na.omit`	numériques et catégorielles
Analyses des cas disponibles	**regtools** ; option disponible dans certaines fonctions (par exemple, `method="pairwise"` dans la fonction `cor`)	numériques et catégorielles
Pondération par probabilité inverse (IPW)	**ipw** (van der Wal et Geskus, 2011)	numériques et catégorielles

TABLE 3. *Packages implémentant les approches paramétriques d'inférence statistique (EM ou bayésiennes)*

Méthodes	**Packages R**	**Cadre d'application**
FIML	**lavaan** (Rosseel, 2012)	modèle à équations structurelles
Approche EM avec un modèle multivarié normal	**norm** (Schafer et Olsen, 1998)	données multivariées gaussiennes
Approche EM avec un modèle log-linéaire	**cat** (Schafer et Olsen, 1998)	données multivariées catégorielles
Équivalent du package **norm** pour des données mixtes	**mix** (Schafer et Olsen, 1998)	données multivariées mixtes
EM avec approche bayésienne ou bootstrap	**Amelia** (Honaker *et al.*, 2011)	variables numériques

 ISSN: 2102-6238

TABLE 4. *Packages contenant des approches d'imputation simple*

Méthodes	**Packages R**	**Cadre d'application**
Moyenne, médiane	**ForImp** ; **Hmisc** ; **simputation**	variables numériques
Mode	**ForImp** ; **Hmisc**	variables catégorielles
LOCF	**zoo**	données longitudinales
k-plus proches voisins	**DMwR** (Torgo, 2010) ; **impute** (Troyanskaya *et al.*, 2001) ; **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016) ; **yaImpute** (Crookston et Finley, 2008)	variables numériques et/ou atégorielles, selon la distance choisie
Hot-deck	**hot.deck** (Cranmer et Gill, 2012) ; **HotDeckImputation** ; **simputation** ; **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016)	tableaux de données mixtes
Régression	**simputation** ; **snpStats** (Bioconductor) ; **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016)	variables numériques pour **simputation** et **VIM** ; données SNP pour **snpStats**
Régression LOESS	**locfit**	variables numériques
Régression stochastique	**mice** ($m = 1$) (van Buuren et Groothuis-Oudshoorn, 2011)	variables numériques
Arbres et forêts aléatoires	**missForest** (Stekhoven et Bühlmann, 2012)	tableaux de données mixtes
NIPALS	**ade4** (Chessel *et al.*, 2004) ; **pcaMethods** (Bioconductor, Stacklies *et al.*, 2007) ; **mixOmics** (Lê Cao *et al.*, 2009)	variables numériques
Analyses factorielles	**missMDA** (Josse *et al.*, 2012)	variables catégorielles et/ou numériques, selon la méthode choisie
Procédure d'imputation « en avant »	**ForImp** (Ferrari *et al.*, 2011)	variables ordinales
Interpolation, ajustement d'une courbe de lissage, estimation de régression longitudinales	**forecast** ; **imputeTS** (Moritz et Bartz-Beielstein, 2017) ; **spacetime** (Pebesma, 2012) ; **timeSeries** ; **xts** ; **zoo** (Zeileis et Grothendieck, 2005)	séries temporelles

 ISSN: 2102-6238

TABLE 5. *Packages incluant des approches d'évaluation de la variabilité due en présence de données manquantes ou due à l'imputation*

Méthodes	**Packages R**	**Cadre d'application**
Outils de diagnostic		
Calcul d'erreurs	**Amelia** (Honaker *et al.*, 2011); **missMDA** (Josse *et al.*, 2012); **yaImpute** (Crookston et Finley, 2008)	tableaux de données mixtes
Graphiques	**mi** (Su *et al.*, 2011); **VIM** (Templ *et al.*, 2012 et Kowarik et Templ, 2016)	tableaux de données mixtes
Imputation multiple		
Équations chaînées	**mi** (Su *et al.*, 2011); **mice** (van Buuren et Groothuis-Oudshoorn, 2011)	tableaux de données mixtes
Hot-deck	**hot.deck** (Cranmer et Gill, 2012)	tableaux de données mixtes
Analyses factorielles (MIPCA, MIMCA)	**missMDA** (Josse *et al.*, 2012)	tableaux de données mixtes
Approche de modélisation jointe (EM et bayésienne)	**Amelia** (Honaker *et al.*, 2011)	variables numériques
Approche de modélisation jointe (bayésienne)	**MixedDataImpute**; **NPBayesInput**	variables catégorielles et mixtes, respetivement
Approches de modélisation jointe (bayésiennes) multi-niveaux	**jomo**; **pan**	tableaux de données mixtes
Combinaison générique	**mitools**	tableaux de données mixtes stratifiés en classes

 ISSN: 2102-6238

Remerciements

Nous souhaitons remercier les deux rapporteurs anonymes pour leurs nombreuses remarques et suggestions qui ont permis de substentiellement améliorer la présentation de cette revue. Nous remercions également Vincent Audigier pour nous avoir pointé plusieurs packages et méthodes qui manquaient dans la version initiale de cette revue.

Références

ABAYOMI, K., GELMAN, A. et LEVY, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 57(3):273–291.

ALBERT, P. et FOLLMANN, D. (2000). Modeling repeated count data subject to informative dropout. *Biometrics*, 56(3):667–677.

ALLISON, P. (2001). *Missing Data*. Quantitative Applications in the Social Sciences. Sage Publications, Thousand Oaks, CA, USA.

ANDRIDGE, R. et LITTLE, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

AUDIGIER, V., HUSSON, F. et JOSSE, J. (2015). Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156.

AUDIGIER, V., HUSSON, F. et JOSSE, J. (2016a). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2):1–18.

AUDIGIER, V., HUSSON, F. et JOSSE, J. (2016b). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26.

BARALDI, A. et ENDERS, C. (2010). An introduction to modern missing data analysis. *Journal of School Psychology*, 48(1):5–37.

BARETTA, L. et SANTANIELLO, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16(Supp. 3):74.

BREIMAN, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

BREIMAN, L., FRIEDMAN, J., OLSEN, R. et STONE, C. (1984). *Classification and Regression Trees*. Chapman and Hall, Boca Raton, Florida, USA.

BURNS, R. (1990). Multiple and replicate item imputation in a complex sample survey. *In* of the CENSUS, B., éditeur : *Proceedings of the 6th Annual Research Conference*, pages 655–665, Washington DC, USA.

CANDÈS, E., SING-LONG, C. et TRZASKO, J. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657.

CARPENTER, J. et KENWARD, M. (2013). *Multiple Imputation and its Application*. Wiley.

CAUSSINUS, H. (1986). Models and uses of principal component analysis (with discussion). *In* de LEEUW, J., HEISER, W., MEULMAN, J. et CRITCHLEY, F., éditeurs : *Multidimensional Data Analysis. Proceedings of a Workshop, Pembroke College, Cambridge University, England*, pages 149–178, Leiden, The Netherlands. DSWO Press.

CHEN, J. et SHAO, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16(2):113–131.

CHESSEL, D., DUFOUR, A. et THIOULOUSE, J. (2004). The ade4 package – I: one-table methods. *R News*, 4(1):5–10.

CLEVELAND, W. et DEVLIN, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

COHEN, J., COHEN, P., WEST, S. et AIKEN, L. (1985). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2nd édition.

COLLINS, L. M., SCHAFER, J. L. et CHI-MING, K. (2007). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351.

COOK, D. et SWAYNE, D. (2007). *Interactive and Dynamic Graphics for Data Analysis*. Use R! Springer-Verlag, New York, NY, USA.

CRANMER, S. et GILL, J. (2012). We have to be discrete about this: a non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43:425–449.

CROOKSTON, N. et FINLEY, A. (2008). yaImpute: an R package for kNN imputation. *Journal of Statistical Software*, 23:10.

 ISSN: 2102-6238

DEMPSTER, A., LAIRD, N. et RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38.

DIGGLE, P. et KENWARD, M. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 43(1):49–93.

DING, Y. et SIMONOFF, J. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11:131–170.

DONG, Y. et PENG, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2:222.

ENDERS, C. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1):128–141.

ENDERS, C. (2010). *Applied Missing Data Analysis*. Guilford Press.

ESCOFIER, B. et PAGÈS, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18(1):121–140.

ESCOUFIER, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

FAY, R. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91(434):490–498.

FELLEGI, I. et HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35.

FERRARI, P. A., ANNONI, P., BARBIERO, A. et MANZI, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics & Data Analysis*, 55(7): 2410–2420.

FINKBEINER, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, 44(4):409–420.

FOLLMANN, D. et WU, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, 51(1):151–168.

FRIEDMAN, J. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, C-26(4):404–408.

GAD, A. et DARWISH, N. (2013). A shared parameter model for longitudinal data with missing values. *American Journal of Applied Mathematics and Statistics*, 1(2):30–35.

GELMAN, A., CARLIN, J., STERN, H. et RUBIN, D. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, USA, 3rd edition édition.

GELMAN, A. et HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierachical Models*. Cambridge University Press, New York, NY, USA.

GOWER, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–874.

GRAHAM, J. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60:549–576.

GRAHAM, J. W., OLCHOWSKI, A. E. et GILREATH, T. E. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.

HECKMAN, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

HOCKING, R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49.

HOERL, A. et KENNARD, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

HOGAN, J. et LAIRD, N. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medecine*, 16(1-3):239–257.

HONAKER, J., KING, G. et BLACKWELL, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software*, 45(7).

HUBERT, P. et RONCHETTI, E. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, USA.

HUISMAN, M. (2000). Imputation of missing item responses: some simple techniques. *Quality & Quantity*, 34(4): 331–351.

ILIN, A. et RAIKO, T. (2010). Practical approaches to Principal Component Analysis in the presence of missing values. *Journal of Machine Learning Research*, 11:1957–2000.

 ISSN: 2102-6238

IMBERT, A., VALSESIA, A., LE GALL, C., ARMENISE, C., LEFEBVRE, G., GOURRAUD, P., VIGUERIE, N. et VILLA-VIALANEIX, N. (2018). Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics*, 34(10):1726–1732.

JAMSHIDIAN, M. et JALAL, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

JAMSHIDIAN, M., JALAL, S. et JANSEN, C. (2014). MissMech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *Journal of Statistical Software*, 56(6):1–31.

JOENSSEN, D. et BANKHOFER, U. (2012). Donor limited hot deck imputation: effect on parameter estimation. *Journal of Theoretical and Applied Computer Science*, 6(3):58–70.

JÖNSSON, P. et WOHLIN, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *In Proceedings of the 10th International Symposium on Software Metrics*, pages 1530–1435, Chicago, IL, USA. IEEE.

JOSSE, J., CHAVENT, M., LIQUET, B. et HUSSON, F. (2012). Handling missing values with regularized iterative multiple correspondance analysis. *Journal of Classification*, 29(1):91–116.

JOSSE, J. et HUSSON, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99.

JOSSE, J., HUSSON, F. et PAGÈS, J. (2009). Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la Société Française de Statistique*, 150(2):28–51.

JOSSE, J., PAGÈS, J. et HUSSON, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3):231–246.

KAISER, J. (2014). Dealing with missing values in data. *Journal of Systems Integration*, 5(1):42–51.

KALTON, G. et KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12(1):1–16.

KIERS, H. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266.

KOHN, R. et ANSLEY, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, 81(395):751–761.

KOWARIK, A. et TEMPL, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.

LAVIT, C., ESCOUFIER, Y., SABATIER, R. et TRAISSAC, P. (1994). The ACT (STATIS method). *Computational Statistics and Data Analysis*, 18(1):97–119.

LÊ CAO, K., GONZÁLEZ, I. et DÉJEAN, S. (2009). *****Omics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855–2856.

LITTLE, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

LITTLE, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

LITTLE, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

LITTLE, R. et RUBIN, D. (2002). *Statistical Analysis with Missing Data*. Wiley.

LITTLE, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

MENG, S. et RUBIN, D. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278.

MENG, X. et RUBIN, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

MOEUR, M. et STAGE, A. (1995). Most similar neighbor: an improved sampling inference procedure for natural resources planning. *Forest Science*, 42(1):337–359.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. et DIGGLE, P. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161.

MOLNAR, F., HUTTON, B. et FERGUSSON, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research? *Canadian Medical Association Journal*, 179(8):751–753.

MORITZ, S. et BARTZ-BEIELSTEIN, T. (2017). imputeTS: time series missing value imputation in R. *The R Journal*, 9(1):207–218.

MORITZ, S., SARDÁ, A., BARTZ-BEIELSTEIN, T., ZAEFFERER, M. et STORK, J. (2015). Comparison of different methods for univariate time series imputation in R. Prepint arXiv 1510.03924.

PEBESMA, E. (2012). spacetime: spatio-temporal data in R. *Journal of Statistical Software*, 51(7):1–30.

 ISSN: 2102-6238

PIGOTT, T. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.
RAO, J. et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4):811–822.
REILLY, M. et PEPE, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medecine*, 16(1-3):5–19.
ROBINS, J., ROTNITZKY, A. et ZHAO, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.
ROBINS, J. et WANG, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1):113–124.
ROSSEEL, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2).
ROTNITZKY, A., ROBINS, J. et SCHARFSTEIN, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.
RUBIN, D. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
RUBIN, D. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):538–543.
RUBIN, D. (1987). *Multlipe Imputation for Nonresponse in Surveys*. Wiley.
RUBIN, D. (2012). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.
SCHAFER, J. (1997). *Analysis of Incomplete Multivariate Data*. CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, Boca Raton, FL, USA.
SCHAFER, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.
SCHAFER, J. et GRAHAM, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177.
SCHAFER, J. et OLSEN, M. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571.
SEAMAN, S. et WHITE, I. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.
SIMON, G. et SIMONOFF, J. (1986). Diagnostic plots for missing data in least squares regression. *Journal of the American Statistical Association*, 81(394):501–509.
STACKLIES, W., REDESTIG, H., SCHOLZ, M., WALTHER, D. et SELBIG, J. (2007). pcaMethods – a bioconductor package providing PCA methods for incomplete data. *Bioconductor*, 23(9):1164–1167.
STAGE, A. et CROOKSTON, N. (2007). Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. *Forest Science*, 53(1):62–72.
STEKHOVEN, D. et BÜHLMANN, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
STUART, E., AZUR, M., FRANGAKIS, C. et LEAF, P. (2009). Multiple imputation with large data sets: a case study of the children's mental health initiative. *American Journal of Epidemiology*, 169(9):1133–1139.
SU, Y., GELMAN, A., HILL, J. et YAJIMA, M. (2011). Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical Software*, 45:2.
TANNER, M. et WONG, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.
TEMPL, M., ALFONS, A. et FILZMOSER, P. (2012). Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1):29–47.
TENENHAUS, M. (1998). *La Régression PLS : Théorie et Pratique*. TECHNIP.
THIJS, H., MOLENBERGHS, G., MICHIELS, B., VERBEKE, G. et CURRAN, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288.
TIERNEY, N., HARDEN, F., HARDEN, M. et MENGERSEN, K. (2015). Using decision trees to understand structure in missing data. *BMJ Open*, 5(6):e007450.
TIPPING, M. et BISHOP, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Association, Series B (Statistical Methodology)*, 61:611–622.
TORGO, L. (2010). *Data Mining with R: Learning with Case Studies*. CRC Data Mining and Knowledge Discovery Series. Chapman and Hall, Boca Raton, Florida, USA.
TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. et

 ISSN: 2102-6238

ALTMAN, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

UNNEBRINK, K. et WINDELER, J. (2001). Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in Medecine*, 20(24):3931–3946.

van BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242.

van BUUREN, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Leiden, The Netherlands.

van BUUREN, S. et GROOTHUIS-OUDSHOORN, K. (2011). MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:3.

van der WAL, W. M. et GESKUS, R. B. (2011). ipw: an R package for inverse probability weighting. *Journal of Statistical Software*, 43(13).

VERBANCK, M., JOSSE, J. et HUSSON, F. (2015). Regularised PCA to denoise and visualise data. *Statistics and Computing*, 25(2):471–486.

VERBEKE, G., MOLENBERGHS, G., THIJS, H., LESAFFRE, E. et KENWARD, M. (2001). Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*, 57(1):7–14.

VOILLET, V., BESSE, P., LIAUBET, L., SAN CRISTOBAL, M. et GONZÁLES, I. (2016). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, 17(402). Forthcoming.

WOLD, H. (1966). Estimation of principal components and related models by iterative least squares. *In* KRISHNAIAH, éditeur : *Multivariate Analysis*, pages 1391–1420. Academic Press, New York, USA.

WU, M. et CARROLL, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188.

ZEILEIS, A. et GROTHENDIECK, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

ZHANG, S. (2012). Nearest neighbor selection for iterative kNN imputation. *Journal of Systems and Software*, 85(11):2541–2552.

ZOU, H. et HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, 67(2):301–320.

 ISSN: 2102-6238