Imputation de données manquantes

Résumé

Cette vignette présente les différents types de problèmes soulevés par la question très fréquente en pratique d'occurrences de données manquantes, que ce soit pour des données matricielles ou longitudinales. Les méthodes d'imputation de données manquantes sont décrites ; les plus rudimentaires : LOCF, imputation par la médiane, la moyenne..., de même que celles par modélisation ou apprentissage statistique : régression et régression locale, k-nn, régression PLS, SVD, Random Forest ou encore par imputation multiple. Ces méthodes sont illustrées et leur efficacité comparée sur trois jeux de données.

Retour au plan du cours

1 Introduction

Malgré la quantité croissante de données disponibles et l'émergence du *Big Data*, les problématiques de données manquantes restent très répandues dans les problèmes statistiques et nécessitent une approche particulière. Ignorer les données manquantes peut entraîner, outre une perte de précision, de forts biais dans les modèles d'analyse.

Les données sont constituées de p variables quantitatives ou qualitatives $(Y_1, \ldots, Y_p)$ observées sur un échantillon de n individus. Il existe des données manquantes représentées par la matrice M dite d'*indication des valeurs manquantes* [13] dont la forme dépend du type de données manquantes.

Définition des différents types de données manquantes et illustration de leurs répartitions possibles. Description des principales stratégies de gestion des données manquantes par suppression de données ou par complétion, sans souci d'exhaustivité.

2 Typologie des données manquantes

2.1 Types de données manquantes

Afin d'aborder correctement l'imputation des données manquantes il faut en distinguer les causes, surtout si elles ne sont pas le simple fruit du hasard. Une typologie a été développée par Little & Rubin (1987) [13], les répartissant en 3 catégories :

MCAR (*missing completely at random*). Une donnée est MCAR, c'est-à-dire manquante de façon complètement aléatoire si la probabilité d'absence est la même pour toutes les observations. Cette probabilité ne dépend donc que de paramètres extérieurs indépendants de cette variable. Par exemple : si chaque participant à un sondage décide de répondre à la question du revenu en lançant un dé et en refusant de répondre si la face 6 apparaît [1]. À noter que si la quantité de données MCAR n'est pas trop importante, ignorer les cas avec des données manquantes ne biaisera pas l'analyse. Une perte de précision dans les résultats est toutefois à prévoir.

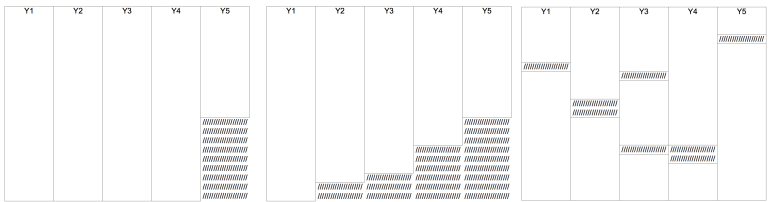
MAR (*Missing at random*). Le cas des données MCAR est peu courant. Il arrive lorsque les données ne manquent pas de façon complètement aléatoire ; si la probabilité d'absence est liée à une ou plusieurs autres variables observées, on parle de *missingness at random* (MAR). Il existe des méthodes statistiques appropriées qui permettrons d'éviter de biaiser l'analyse (voir 4)

MNAR (*Missing not at random*) La donnée est manquante de façon non aléatoire (MNAR) si la probabilité d'absence dépend de la variable en question. Un exemple répandu [1][9] est le cas où des personnes avec un revenu important refusent de le dévoiler. Les données MNAR induisent une perte de précision (inhérente à tout cas de données manquantes) mais aussi un biais qui nécessite le recours à une analyse de sensibilité.

2.2 Répartition des données manquantes

Soit $Y = (y_{ij}) \in \mathbb{R}^{n \times p}$ la matrice rectangulaire des données pour p variables $Y_1, \ldots, Y_p$ et n observations. Considérons $M = (m_{ij})$ la matrice d'indication des valeurs manquantes [13], qui va définir la répartition des données manquantes. On considèrera alors 3 types de répartition :

1. Les valeurs manquantes **univariées**. Pour une variable Y_k seulement, si une observation y_{ki} est manquante, alors il n'y aura plus d'observation de cette variable. Une illustration est donnée figure 1a.
2. Les valeurs manquantes sont dites **monotones** si Y_j manquante pour un individu i implique que toutes les variables suivantes $\{Y_k\}_{k>j}$ sont manquantes pour cet individu (figure 1b). L'indicateur de données manquantes M est alors un entier $M \in (1, 2, \ldots, p)$ pour chaque individu, indiquant le plus grand j pour lequel Y_j est observé.
3. Les valeurs manquantes sont **non monotones** (ou **arbitraires**), comme le représente la figure 1c Dans ce cas, on définit la matrice de valeurs manquantes par $M = (m_{ij})$ avec $m_{ij} = 1$ si y_{ij} est manquant et zéro sinon.



(a) (b) (c)

FIGURE 1 – Répartitions des données manquantes. (a) univariées, (b) monotones et (c) arbitraires/non monotones

Cette répartition est valable pour les données longitudinales (voir figure 2). La répartition monotone correspond alors à une censure à droite.

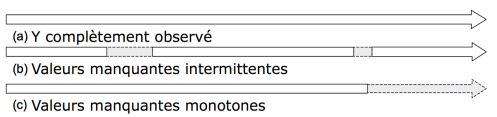


FIGURE 2 – Répartitions des données manquantes pour des variables longitudinales. (a) jeu complet, (b) arbitraires/non monotones et (c) monotones

2.3 Probabilité d'absence

La probabilité d'absence selon le type de données manquantes (MCAR, MAR, MNAR) peut être exprimé en fonction de M [13]. Les données sont divisées en deux selon la matrice M d'indication des données manquantes. On définit donc $Y_{obs} = Y\mathbb{1}_{\{M=0\}}$ les données observées et $Y_{mis} = Y\mathbb{1}_{\{M=1\}}$ les données manquantes telles que $Y = \{Y_{obs}, Y_{mis}\}$. Le mécanisme des données manquantes est caractérisé par la distribution conditionnelle de M sachant Y donnée par $p(M|Y)$.

- Dans le cas des données **MCAR**, l'absence de données ne dépend pas des valeurs de Y donc

$$p(M|Y) = p(M) \text{ pour tout } Y. \tag{1}$$

- Dans le cas **MAR** : soit Y_{obs} la partie observée du jeu de données et Y_{mis} les données manquantes. MAR signifie que l'absence de données dépend uniquement de Y_{obs} :

$$p(M|Y) = p(M|Y_{obs}) \text{ pour tout } Y_{mis}. \tag{2}$$

- Enfin, les données sont **MNAR** si la distribution de M dépend aussi de Y_{mis}.

Exemple pour un échantillon aléatoire univarié

Soit $Y = (y_1, \ldots, y_n)^\top$ où y_i est l'observation d'une variable aléatoire pour l'individu i, et $M = (M_1, \ldots, M_n)$ où $M_i = 0$ pour les données observées et $M_i = 1$ pour les données manquantes. On suppose également que la distribution conjointe est indépendante des individus. Alors

$$p(Y, M) = p(Y)p(M|Y) = \prod_{i=1}^{n} p(y_i) \prod_{i=1}^{n} p(M_i|y_i) \tag{3}$$

où $p(y_i)$ est la densité de y_i et $p(M_i|y_i)$ est la densité d'une loi de Bernoulli pour l'indicateur binaire M_i avec la probabilité $\mathbb{P}(M_i = 1|y_i)$ que y_i soit manquante.

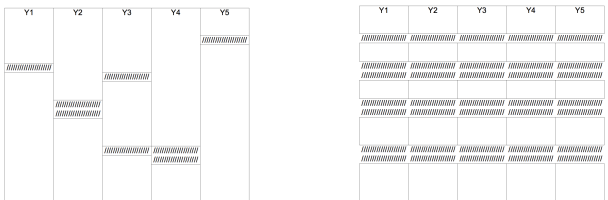
Si $\mathbb{P}(M_i = 1|y_i) = \alpha$ avec α une constante qui ne dépend pas de y_i alors c'est un cas MCAR (ou dans ce cas aussi MAR). Si $\mathbb{P}(M_i = 1|y_i)$ dépend de y_i, le mécanisme de données manquantes est MNAR.

3 Analyse sans complétion

3.1 Méthodes avec suppression de données

Dans certains cas, l'analyse est possible sans imputer les données manquantes. En général, on se reporte à deux méthodes classiques :

- **L'analyse des cas concrets**, qui consiste à ne considérer que les individus pour lesquels toutes les données sont disponibles, i.e. en supprimant les lignes comportant des valeurs manquantes. C'est ce qui est fait automatiquement avec R (`na.action=na.omit`). Cette méthode, on le voit bien figure 3, risque de supprimer trop de données et d'augmenter de beaucoup la perte de précision. De plus, si les données ne sont pas MCAR, retirer des observations va induire un biais dans l'analyse puisque le sous-échantillon des cas représentés par les données manquantes ne sont pas forcément représentatifs de l'échantillon initial.



(a) (b)

FIGURE 3 – Répartitions des données manquantes. (a) données d'origine, valeurs manquantes arbitraires, (b) observations restantes en analyse des cas complets

- **L'analyse des cas disponibles**. Afin d'éviter de supprimer trop de données, il est possible de faire de la suppression par paires (*pairwise deletion*) ou analyse des cas disponibles (*available-case analysis*). Différents aspects du problème sont alors étudiés avec différents sous-échantillons. Cependant, les différentes analyses ne seront pas nécessairement compatibles entre elles.
  L'analyse des cas disponibles correspond aussi au cas où une variable est supprimée du jeu de données à cause de sa trop grande quantité de valeurs manquantes.

3.2 Méthodes tolérant des données manquantes

Si la plupart des méthodes d'analyse suppriment automatiquement les données manquantes, certaines les tolèrent. C'est le cas par exemple des arbres CART qui considèrent des *surrogate splits* ou divisions de substitution : Au moment du split d'un nœud, plusieurs couples variables / seuil optimaux sont considérés et mémorisés. Au moment de l'utilisation, si la donnée est manquante pour une observation, ce n'est pas la meilleure division qui est utilisée mais celle juste après lui est substituée [7].

L'algorithme NIPALS [3] de la régression PLS permet de tolérer les données manquantes tout en les imputant (voir section 4.5).

Par ailleurs `XGBoost` propose également une stratégie assez complexe permettant de tolérer des données manquantes.

4 Méthodes d'imputation

Cette section donne un aperçu non exhaustif des méthodes de complétion les plus courantes. Un jeu de données est constitué de p variables quantitatives ou qualitatives $(Y_1, \ldots, Y_p)$ observées sur un échantillon de n individus ; M désigne la matrice d'indication des valeurs manquantes par $m_{ij} = \mathbb{1}_{\{y_{ij} \text{ manquante}\}}$

4.1 Complétion stationnaire

Il existe plusieurs complétions stationnaires possibles : valeur la plus fréquemment représentée (*Concept Most Common Attribute Value Fitting*, CMCF [14]) ou plus simplement dernière valeur connue (*Last observation carried forward*, LOCF) :

$$(y_{ij})_{mis} = y_{i^\star j^\star} = \{y_{i^\star j} | m_{i^\star j} = 0, j < j^\star\} \tag{4}$$

Cette méthode peut sembler trop naïve mais est souvent utilisée pour poser les bases d'une comparaison entre méthodes de complétion.

4.2 Complétion par une combinaison linéaire des observations

Une autre technique répandue consiste à remplacer toutes les valeurs manquantes par une combinaison linéaire des observations. On retiendra le cas d'imputation par la moyenne :

$$(y_{ij})_{mis} = y_{i^\star j^\star} = \bar{Y}_{j^\star} \tag{5}$$

ou par la médiane :

$$(y_{ij})_{mis} = y_{i^\star j^\star} = \tilde{Y}_{j^\star} \tag{6}$$

Mais ce cas se généralise à toute combinaison linéaire pondérée des observations.

Au lieu d'utiliser toutes les valeurs disponibles, il est possible de se restreindre à des méthodes qui sélectionnent les valeurs les plus influentes par agrégation locale ou régression voire en combinant différents aspects.

4.3 Méthode des plus proches voisins (KNN)

La complétion par k plus proches voisins (*k-nearest neighbors* ou KNN) consiste à exécuter l'algorithme suivant qui modélise et prévoit les données manquantes.

Algorithme des k plus proches voisins (k-nn)

---

1. Choix d'un entier k : $1 \geq k \geq n$.
2. Calculer les distances $d(Y_{i^\star}, Y_i)\,, \quad i = 1, \ldots, n$
3. Retenir les k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ pour lesquelles ces distances sont les plus petites.
4. Affecter aux valeurs manquantes la moyenne des valeurs des k voisins :

$$(y_{ij})_{mis} = y_{i^\star j^\star} = \frac{1}{k}\left(Y_{(i_1)} + \ldots + Y_{(i_k)}\right) \tag{7}$$

Comme pour la classification par KNN, la méthode des plus proches voisins nécessite le choix du paramètre k par optimisation d'un critère. De plus, la notion de distance entre les individus doit être choisie avec précaution. On considèrera usuellement la distance Euclidienne ou de Mahalanobis.

4.4 Régression locale

La régression locale (en anglais *LOcal regrESSion* : LOESS) [15] permet également d'imputer des données manquantes. Pour cela, un polynôme de degré faible est ajusté autour de la donnée manquante par moindres carrés pondérés, en donnant plus de poids aux valeurs proches de la donnée manquante.

Soit $Y_{i^\star}$ une observation à laquelle il manque q valeurs manquantes. On impute ces données manquantes par régression locale en suivant l'algorithme ci-après.

Algorithme LOESS

---

1. Obtention des k plus proches voisins $Y_{(i_1)}, \ldots, Y_{(i_k)}$
2. Création des matrices $A \in \mathbb{R}^{k\times(n-q)}$, $B \in \mathbb{R}^{k\times q}$ et $w \in \mathbb{R}^{(n-q)\times 1}$ de sorte que :
   - Les lignes de A correspondent aux voisins privés des valeurs aux indices des données manquantes de $Y_{i^\star}$
   - Les colonnes de B correspondent aux valeurs des voisins aux indices des données manquantes de $Y_{i^\star}$
   - Le vecteur w correspond aux $(n-q)$ valeurs observées de $Y_{i^\star}$: $w_j = (y_{i^\star j})_{obs}$
3. Résolution du problème des moindres carrés

$$\min_{x\in\mathbb{R}^k} \parallel A^\top x - w \parallel \tag{8}$$

   où $\parallel \cdot \parallel$ est la norme quadratique de $\mathbb{R}^k$.
4. Le vecteur u des données manquantes s'exprime alors par

$$u = B^\top x = B^\top (A^\top)^{-1} w \tag{9}$$

   avec $(A^\top)^{-1}$ la matrice pseudo-inverse de $A^\top$.

4.5 Algorithme NIPALS

L'algorithme NIPALS (*Nonlinear Iterative Partial Least Squares*) est une méthode itérative pour estimer la régression PLS. Cet algorithme peut être adapté pour l'imputation de données manquantes [3]. Soit $Y = (Y_1, \ldots, Y_p)$ tel que $\forall i \in 1, \ldots, p$, $E(Y_i) = 0$ (chaque colonne de la matrice est centrée). L'expansion de Y en termes de composantes principales et de facteurs principaux est donnée par

$$Y = \sum_{h=1}^{q} \xi_h u_h \tag{10}$$

où $q = dim L_2(Y)$ et $\{\xi_h\}_{h=1,\ldots,q}$ sont les composantes principales et $\{u_h\}_{h=1,\ldots,q}$ les vecteurs principaux de l'ACP de Y. Donc pour chaque variable Y_i on a

$$Y_i = \sum_{h=1}^{q} \xi_h u_h(i) \tag{11}$$

L'idée étant que pour chaque h, $u_h(i)$ représente la pente de la régression linéaire de Y_i sur la composante ξ_h. L'algorithme NIPALS va permettre d'obtenir $\{\hat{\xi}_h\}_{h=1,\ldots,q}$ et $\{\hat{u}_h\}_{h=1,\ldots,q}$ les approximations de $\{\xi_h\}_{h=1,\ldots,q}$ et $\{u_h\}_{h=1,\ldots,q}$.

Algorithme NIPALS

1. $Y^0 = Y$
2. **Pour** $h = 1, \ldots, q$ **faire**
   (a) $\xi_h = Y_1^{h-1}$
   (b) **Tant que** u_h n'a pas convergé **faire**
      i. **Pour** $i = 1, \ldots, p$ **faire**

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

      ii. Normaliser u_h
      iii. **Pour** $i = 1, \ldots, N$ **faire**

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

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

Les données manquantes sont approchées par

$$(\hat{y}_{ij})_{mis} = \sum_{h=1}^{q} \hat{\xi}_h(i) \hat{u}_h(j) \tag{12}$$

4.6 Par décomposition en valeurs singulières (SVD)

4.6.1 Cas où il y a suffisamment de données observées

S'il y a bien plus de données observées que de données manquantes, on sépare le jeu de données Y en deux groupes : d'un côté Y^c avec les observations complètes et de l'autre Y^m comprenant les individus pour lesquels certaines données manquent. On considère alors la décomposition en valeurs singulières (SVD) tronquée du jeu complet [6] :

$$\hat{Y_J^c} = U_J D_J V_J^\top \tag{13}$$

où D_J est la matrice diagonale comprenant les J premières valeurs singulières de Y^c. Les valeurs manquantes sont alors imputées par régression :

$$\min_{\beta \in \mathbb{R}^J} \sum_{i \text{ observées}} \left(Y_{i^\star} - \sum_{j=1}^{J} v_{lj} \beta_j \right)^2 \tag{14}$$

Soit $V_J^\star$ la version tronquée de V_J, c'est à dire pour laquelle les lignes correspondant aux données manquantes de la ligne $Y_{i^\star}$ sont supprimées. Une solution du problème 14 est alors

$$\hat{\beta} = (V_J^{\star\top} V_J^\star)^{-1} V_J^{\star\top} Y_{i^\star} \tag{15}$$

La prédiction des données manquantes est donc donnée par

$$Y_{i^\star} = V_J^{(\star)} \hat{\beta} \tag{16}$$

où $V_J^{(\star)}$ est le complément de $V_J^\star$ dans V_J.

Comme pour KNN, cette méthode nécessite le choix du paramètre J. On se ramènera alors à un problème de minimisation :

$$\min_{X \text{ de rang } J} \| Y^c - X \|_F \tag{17}$$

avec $\| \cdot \|_F$ la norme de Frobenius.

4.6.2 Cas où il y a trop de données manquantes

Si les données manquantes sont trop nombreuses, cela induira un biais important dans le calcul de la base de décomposition. De plus, il arrive qu'il y ait au moins une donnée manquante pour toutes les observations. Dans ce cas, il faut résoudre le problème suivant :

$$\min_{U_J, V_J, D_J} \| Y - m - U_J D_J V_J^\top \|_\star \tag{18}$$

où $\| \cdot \|_\star$ somme les carrés des éléments de la matrice, en ignorant les valeurs manquantes. m est le vecteur des moyennes des observations. La résolution de ce problème suit l'algorithme suivant :

Algorithme de Complétion par SVD

---

1. Créer une matrice Y^0 pour laquelle les valeurs manquantes sont complétées par la moyenne,
2. Calculer la SVD solution du problème (18) pour la matrice complétée Y^i. On crée ainsi Y^{i+1} en remplaçant les valeurs manquantes de Y par celles de la régression.
3. Itérer l'étape précédente jusqu'à ce que $\| Y^i - Y^{i+1} \| / \| Y^i \| < \epsilon$, seuil arbitraire (souvent à 10^{-6})

4.7 Utilisation des Forêts aléatoires

Stekhoven et Bühlmann (2011)[5] ont proposé une méthode de complétion basée sur les forêts aléatoires appelée `missForest`. Une librairie R éponyme lui est associée. Cette méthode nécessite une première imputation naïve, par défaut une complétion par la moyenne, afin d'obtenir un échantillon d'apprentissage complet. Puis une série de forêts aléatoires sont ajustées jusqu'à la première dégradation du modèle.

Pour formaliser cela, on décompose le jeu de données initial en quatre parties. Pour chaque variable Y^s, $s = 1, \ldots, S$ dont les valeurs manquantes sont indexées par $i^s_{mis} \subseteq \{1, \ldots, n\}$, on définit

1. y^s_{obs} les valeurs observées dans Y^s
2. y^s_{mis} les valeurs manquantes dans Y^s
3. $X^s = Y \setminus Y^s$ l'ensemble des régresseurs de Y^s parmi lesquels on considère
   (a) x^s_{obs} les régresseurs observés pour $i^s_{obs} = \{i, \ldots, n\} \setminus i^s_{mis}$
   (b) x^s_{mis} les régresseurs manquants pour i^s_{mis}

La méthode suit alors l'algorithme suivant :

Algorithme MissForest

---

1. Première complétion "naïve" des valeurs manquantes.
2. Soit k le vecteur des indices de colonnes de Y triées par quantité croissante de valeurs manquantes ;
3. **Tant que** γ n'est pas atteint **faire**
   (a) Y^{old}_{imp} = matrice précédemment imputée
   (b) **Pour** s dans k **faire**
      i. Ajuster $y^{(s)}_{obs} \sim x^{(s)}_{obs}$ par forêt aléatoire
      ii. Prédire $y^{(s)}_{mis}$ avec les régresseurs $x^{(s)}_{mis}$
      iii. Y^{new}_{imp} est la nouvelle matrice complétée par les valeurs prédites $y^{(s)}_{mis}$
   (c) mettre à jour le critère γ

Avec un critère d'arrêt γ atteint dès que la différence entre la matrice de données nouvellement imputé et la précédente augmente pour la première fois. La différence de l'ensemble des variables continues est définie comme

$$\Delta_N = \frac{\sum_{j \in N} \left(Y_{imp}^{new} - Y_{imp}^{old}\right)^2}{\sum_{j \in N} \left(Y_{imp}^{new}\right)^2} \tag{19}$$

En cas de variables qualitatives on définit la différence par

$$\Delta_F = \frac{\sum_{j \in F} \sum_{i=1}^{n} \mathbb{1}_{Y_{imp}^{new} \neq Y_{imp}^{old}}}{\#NA} \tag{20}$$

4.8 Inférence Bayésienne

Soit θ la réalisation d'une variable aléatoire et soit $p(\theta)$ sa distribution *a priori*. La distribution *a posteriori* est donc donnée par :

$$p(\theta|Y_{obs}) \propto p(\theta) f(Y_{obs}; \theta) \tag{21}$$

La méthode de *data augmentation* de Tanner et Wong (1987) [10] simule de manière itérative des échantillons aléatoires des valeurs manquantes et des paramètres du modèle, compte tenu des données observées à chaque itération, constituée d'une étape d'imputation (I) et d'une étape "postérieure" (P).

Soit $\theta^{(0)}$ un tirage initial obtenu à partir d'une approximation de la distribution a posteriori de θ. Pour une valeur $\theta^{(t)}$ de θ à un instant t

Imputation (I) : soit $Y_{mis}^{(t+1)}$ avec une densité $p(Y_{mis}|Y_{obs}, \theta^{(t)})$

Postérieure (P) : soit $\theta^{(t+1)}$ avec une densité $p(\theta|Y_{obs}, Y_{mis}^{(t)})$

Cette procédure itérative finira par obtenir un tirage de la distribution conjointe de $(Y_{mis}, \theta|Y_{obs})$ lorsque $t \to +\infty$

4.9 Imputation multiple

L'imputation multiple consiste, comme son nom l'indique, à imputer plusieurs fois les valeurs manquantes afin de combiner les résultats pour diminuer l'erreur (le bruit) due à l'imputation [4]. Cela permet également de définir une mesure de l'incertitude causée par la complétion.

Le maintien de la variabilité d'origine des données se fait en créant des valeurs imputées qui sont basées sur des variables corrélées avec les données manquantes et les causes d'absence. L'incertitude est prise en compte en créant des versions différentes de données manquantes et l'observation de la variabilité entre les ensembles de données imputées.

4.10 Amelia II

Amelia II est un programme d'imputation multiple développé par James Honaker et al. (2011) [8]. Le modèle s'appuie sur une hypothèse de normalité : $Y \sim \mathcal{N}_k(\mu, \Sigma)$, et nécessite donc parfois des transformations préalables des données.

Soit M la matrice d'indication des données manquantes et $\theta = (\mu, \Sigma)$ les paramètres du modèle. Une autre hypothèse est que les données sont **MAR** donc

$$p(M|Y) = p(M|Y_{obs}) \tag{22}$$

La vraisemblance $p(Y_{obs}|\theta)$ s'écrit alors

$$p(Y_{obs}, M|\theta) = p(M|Y_{obs}) p(Y_{obs}|\theta) \tag{23}$$

Donc

$$L(\theta|Y_{obs}) \propto p(Y_{obs}|\theta) \tag{24}$$

Or en utilisant la propriété itérative de l'espérance

$$p(Y_{obs}|\theta) = \int p(Y|\theta) dY_{mis} \tag{25}$$

On obtient donc la loi à posteriori

$$p(\theta|Y_{obs}) \propto p(Y_{obs}|\theta) = \int p(Y|\theta) dY_{mis} \tag{26}$$

L'algorithme EMB d'Amelia II combine l'algorithme EM classique (du maximum de vraisemblance) avec une approche bootstrap. Pour chaque tirage, les données sont estimées par *bootstrap* pour simuler l'incertitude puis l'algorithme EM est éxécuté pour trouver l'estimateur *a posteriori* $\hat{\theta}_{MAP}$ pour les données *bootstrap*. Les imputations sont alors créées par tirage de Y_{mis} selon sa distribution conditionnelle sur Y_{obs} et des tirages de θ.

5 Exemple

5.1 Fraudes sur la consommation en gaz

Les différentes méthodes de complétion ont été testées et comparées sur un exemple de détection de fraudes sur la consommation en gaz. Soit $Y \in \mathbb{R}^{N\times 12}$ tel que y_{ij} soit la consommation de gaz de l'individu i au mois j. La répartition des données manquantes est non monotone et on fait l'hypothèse de données MAR. Après une transformation en log afin d'approcher la normalité, la complétion a été effectuée. Les résultats ont été comparés avec un échantillon test de 10% des données, préalablement retiré du set.

Ce jeu de données réel comporte au moins une valeur manquante par individu, et au total 50.4% des données sont manquantes. Si on ne considère que la consommation mensuelle individuelle, sans variables exogènes, on obtient la répartition des erreurs de chaque méthode représentée Figure 4.

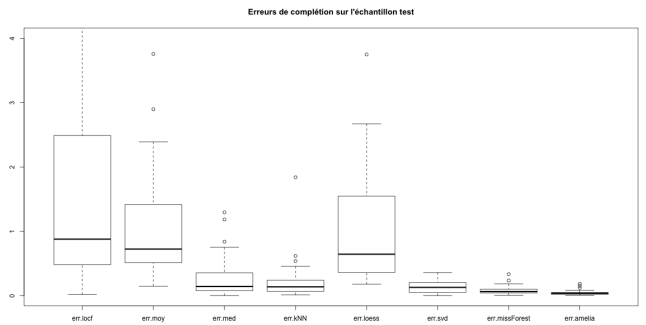


FIGURE 4 – Fraudes - Erreurs de complétion sur un échantillon test

5.2 Encours Boursiers Parisiens (EBP)

On s'intéresse aux cours des actifs boursiers sur la place de Paris de 2000 à 2009. On considère 252 cours d'entreprises ou indices régulièrement cotés sur cette période. En se limitant au cas MCAR, on crée artificiellement de plus en plus de données manquantes à imputer. Pour 10% de données manquantes, une comparaison des méthodes d'imputations est données Figure 5. Trois méthodes se détachent visiblement : SVD, missForest et AmeliaII.

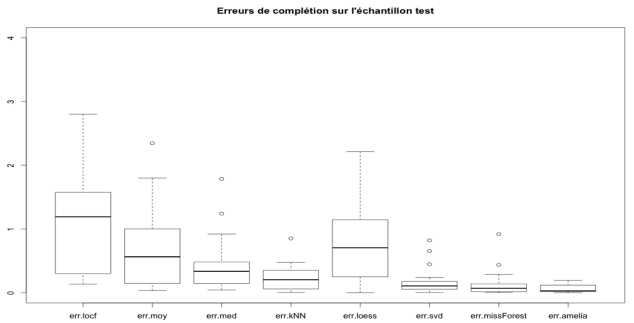


FIGURE 5 – EBP - Erreurs de complétion sur un échantillon test de 10%

La robustesse de ces méthodes a été testée en augmentant graduellement la quantité de données manquantes. Les résultats sont donnés Figure 6 pour AmeliaII et Figure 7 pour missForest.

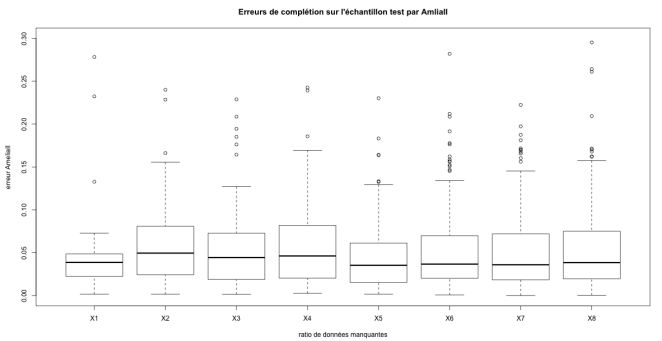


FIGURE 6 – EBP - Erreurs de complétion sur un échantillon test par AmeliaII quand la quantité de valeurs manquantes augmente

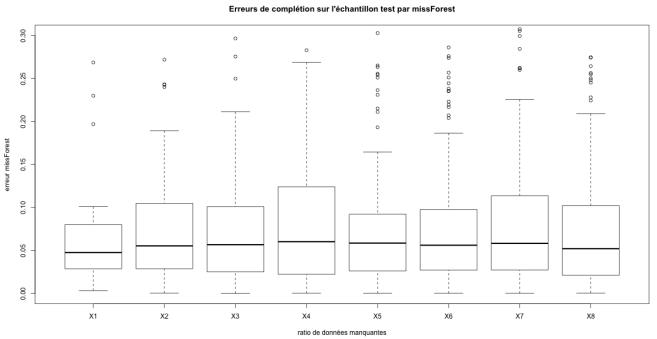


FIGURE 7 – EBP - Erreurs de complétion sur un échantillon test par missForest quand la quantité de valeurs manquantes augmente

5.3 Maladies Coronariennes (CHD)

La plupart des méthodes d'imputation de sont définies que pour des variables quantitatives. Mais certaines méthodes présentées ci-dessus permettent d'imputer des données qualitatives, voire hétérogènes. C'est le cas de LOCF, KNN et missFavorest qui ont donc été testées sur un jeu de données de référence sur les problèmes de complétion [11]. Les données ont été acquises par Detranao et al. (1989) [12] et mises à disposition par Bache et Lichman (2013)[2]. Elles se présentent sous la forme d'une matrice d'observations médicales $Y \in \mathbb{R}^{N\times 14}$ de 14 variables hétérogènes pour N patients. Le jeu de données contient donc des variables quantitatives (`age, pression, cholestérol, fréquence cardiaque maximale, oldpeak`) et qualitatives (`sexe, douleur, sucre, cardio, angine, pente du pic, nombre de vaisseaux cardiaques, thalassémie, absence/présence de maladie cardiaque`).

En se limitant toujours au cas MCAR, on crée artificiellement de plus en plus de données manquantes à imputer. L'adéquation de l'imputation est donnée par la moyenne de l'erreur en valeur absolue dans le cas des données quantitatives et par la distance de Hamming dans le cas des données qualitatives. Les résultats sont représentés Figure 8.

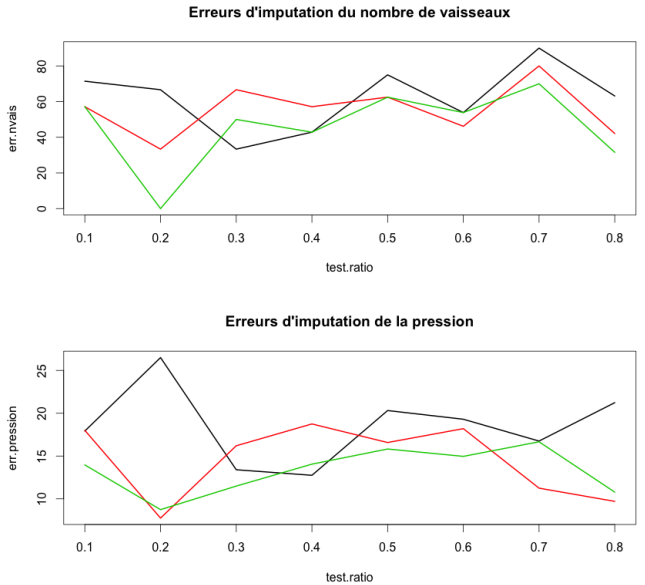


FIGURE 8 – CHD - Erreurs de complétion sur un échantillon test par LOCF (noir), KNN (rouge) et missForest (vert) quand la quantité de valeurs manquantes augmente, pour une variable qualitative (au dessus) et quantitative (en dessous)

Références

[1] Gelman A. et Hill J., *Data Analysis Using Regression and Multilevel/-Hierarchical Models*, chap. 25, p. 529–563, Cambridge University Press, 2007.

[2] K. Bache et M. Lichman, *UCI Machine Learning Repository*, 2013, http://archive.ics.uci.edu/ml.

[3] Preda C., Saporta G. et Hedi Ben Hadj Mbarek M., *The NIPALS algorithm for missing functional data*, Romanian Journal of Pure and Applied Mathematics **55** (2010), n° 4, 315–326.

[4] Rubin D.B., *Multiple Imputation for Nonresponse in Surveys*, Wiley, 1987.

[5] Stekhoven D.J. et Bühlmann P., *MissForest - nonparametric missing value imputation for mixed-type data*, Bioinformatics Advance Access (2011).

[6] Hastie et al, *Imputing Missing Data for Gene Expression Arrays*, Rap. tech., Division of Biostatistics, Stanford University, 1999.

[7] A. J. Feelders, *Handling Missing Data in Trees : Surrogate Splits or Statistical Imputation.*, PKDD, Lecture Notes in Computer Science, t. 1704, Springer, 1999, p. 329–334.

[8] Honaker J., King G. et Blackwell M., *Amelia II : A Program for Missing Data*, Journal of statistical software **45** (2011), n° 7.

[9] Glasson Cicignani M. et Berchtold A., *Imputation de Donnees Manquantes : Comparaison de Differentes Approches*, 42e Journees de Statistique, 2010.

[10] Tanner M.A. et Wong W.H., *The Calculation of Posterior Distributions by Data Augmentation*, Journal of the American Statistical Association **82** (1987), n° 398, 528–540.

[11] Setiawan N.A., Venkatachalam P.A. et Hani A.F.M., *A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set*, 4th Kuala Lumpur International Conference on Biomedical Engineering 2008 (University of Malaya Department of Biomedical Engineering Faculty of Engineering, réd.), t. 21, Springer Berlin Heidelberg, 2008, p. 266–269.

[12] Detrano R., Janosi A., Steinbrunn W., Pfisterer M., Schmid J., Sandhu S., Guppy K., Lee S. et Froelicher V., *International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease*, American Journal of Cardiology **64** (1989), 304–310.

[13] Little R.J.A. et Rubin D.B., *Statistical Analysis with Missing Data*, Wiley series in probability and statistics, 1987.

[14] Grzymala Busse J. W., Grzymala Busse W. J. et Goodwin L. K., *Coping With Missing Attribute Values Based on Closest Fit in Preterm Birth Data : A Rough Set Approach*, Computational Intelligence **17** (2001), 425–434.

[15] Cleveland W.S. et Devlin S.J., *Locally-Weighted Regression : An Approach to Regression Analysis by Local Fitting*, Journal of the American Statistical Association **83** (1988), n° 403, 596–610.