4. LES DONNEES MANQUANTES SONT INEVITABLES[13]

4.1 Que faire avec des données manquantes ?

Dans certaines expériences, et dans certaines enquêtes il n'est pas possible de mesurer toutes les caractéristiques de certaines unités de base ; on dit que l'on a affaire à des **données manquantes**. Certaines techniques statistiques permettent de traiter de tels tableaux de données, et quelquefois de les estimer. Ces techniques, dites **techniques d'imputation** [14], consistent à attribuer une réponse en s'appuyant sur les autres valeurs du corpus de données ou avec des informations extérieures.

Dans tout traitement statistique le rôle à faire jouer aux données manquantes apparaît toujours. Par donnée manquante, il faut entendre des données qui sont manquantes pour quelques variables (mais pas pour toutes) et pour une partie des observations (mais pas pour toutes). Si une variable manque pour toutes les observations c'est une *variable non observée* ou une *variable latente*. Pourquoi les données manquantes représentent-elles un problème ? Tout simplement parce que la majorité des méthodes statistiques courantes et les logiciels qui leur sont associés supposent que toutes les variables d'un modèle spécifié sont mesurées pour toutes les observations. Il s'ensuit que dans pratiquement tous les logiciels une telle observation est supprimée[15] ; cette situation est illustrée sur la figure 5.

Observations\Variables	x^1	x^2	x^3	x^4	x^5
x_1	NA				NA
x_2					
x_3					NA
x_4	NA			NA	
x_5					
x_6					NA

*FIGURE 5. Schéma illustrant un tableau 6*5 avec données manquantes notées NA. Le tiers des observations seulement est complet ; seules 20 % des cellules sont manquantes*

Dans un tel cas, on est conduit à ne conserver que les cas complets soit deux observations sur les six de départ, une fraction faible du corpus initial. C'est la raison pour laquelle, depuis bien longtemps, tous les chercheurs ont essayé de « boucher les trous ». Mais la manière de récupérer ce qui n'a pas été mesuré ou observé est rarement décrite, la pratique revêtant un caractère un peu honteux[16]. C'est un problème d'éthique scientifique assez voisin de celui qui devrait se poser quand un institut de sondage n'indique pas la manière dont un sondage est redressé pour être représentatif d'une population. Et en réalité, jusqu'à la dernière décennie, le remplacement des données était

[13] En anglais : *missing data*.
[14] Le terme *imputation* (en français ou en anglais) a remplacé celui plus statistique d'*estimation* ; c'est une vision actuelle « comptable » et simplement calculatoire qui ne nous paraît pas la meilleure.
[15] En anglais : cette façon de faire s'appelle *listwise deletion* ou *complete case analysis*.
[16] Allison (2009) parle de *dirty little secret*.

une opération qui relevait plus du bricolage que d'une approche réellement scientifique. En fait, aucune des méthodes proposées n'avait de solides bases mathématiques.

4.2 Origine des données manquantes

Avant d'examiner quelques méthodes, il faut d'abord regarder quelles sont les suppositions possibles sur l'origine des données manquantes. Aucune méthode pour estimer des données manquantes ne peut fonctionner sans ces suppositions qui sont assez contraignantes et souvent … non vérifiables ! Très grossièrement, une donnée est manquante au hasard si le mécanisme qui aboutit à l'absence est indépendant de la valeur (non observée). Mais Little & Rubin (2002) ont donné une formulation plus détaillée avec des suppositions différentes pour le mécanisme de production de la donnée manquante.

La plus forte est que les données sont **manquantes entièrement au hasard** (***MCAR***)[17]. Supposons qu'il n'y ait qu'une seule variable avec données manquantes, nous la notons Z ; supposons que nous ayons un autre ensemble de variables représentées par une matrice $\mathbf{X}$ toujours observée. Soit un indicateur R_Z qui vaut 1 si Z est manquante et 0 si Z est observée, la supposition *MCAR* peut alors s'écrire :

$$P[R_Z = 1|\mathbf{X}, Z] = P[R_Z = 1].$$

Cette expression signifie que la probabilité qu'une donnée soit manquante ne dépend ni des valeurs observées $\mathbf{X}$ ni des possibles valeurs manquantes Z. Quelles sont les variables que doit contenir $\mathbf{X}$ de telle sorte que la supposition *MCAR* soit satisfaite ? La réponse est : seulement les variables dont on doit estimer les paramètres du modèle. On peut tester la supposition *MCAR* d'une façon assez simple en comparant les moyennes des variables de $\mathbf{X}$ selon les valeurs 0 ou 1 de R_Z. Une façon plus naturelle (pour un statisticien) est de faire une régression logistique de R_Z sur $\mathbf{X}$[18]. S'il y a des coefficients significatifs, ceci indiquera que la supposition *MCAR* n'est pas satisfaite. Mais il n'est pas possible de tester si l'absence de Z dépend de Z lui-même (conditionnellement à $\mathbf{X}$) car, pour le faire, il faudrait connaître les valeurs manquantes !

Exemple : Du sang est prélevé chez plusieurs patients. Lors du transport des tubes de sang de la salle de prélèvement à la salle d'analyse, des tubes tombent et se cassent. Ces échantillons de sang ne pourront pas être analysés. Ces données seront *MCAR*.

Une supposition plus faible (mais néanmoins forte) est celle de données **manquantes au hasard** (***MAR***) [19] ; on peut l'écrire :

$$P[R_Z = 1|\mathbf{X}, Z] = P[R_Z = 1|\mathbf{X}]$$

Cette équation signifie que l'absence de valeurs de Z peut dépendre de $\mathbf{X}$. mais pas de Z lui-même. Dans un dispositif expérimental, cette absence peut dépendre du groupe (témoin ou traité pour simplifier) mais pas de la valeur dans le groupe. Dans les deux cadres des deux suppositions, l'ensemble des variables de $\mathbf{X}$ dépend du modèle à estimer. Dans le cas d'une étude sur les revenus, l'introduction de variables comme l'âge, le sexe,

[17] En anglais : *missing completely at random* (*MCAR*).
[18] La régression logistique sera présentée au chapitre 12.
[19] En anglais : *missing at random* (*MAR*).

le niveau d'éducation ou le type d'activité peut rendre la supposition *MAR* plus raisonnable. On dit que le processus qui génère les données manquantes est **ignorable**, il peut être ignoré. En fait, il faut que les paramètres qui engendrent l'absence de données soient indépendants de ceux du modèle à estimer ; alors on peut estimer de manière valable les paramètres du modèle.

Exemple : On recueille la pression artérielle de plusieurs patients. La mesure de la pression artérielle étant suivie surtout chez les personnes âgées, il y aura plus de données manquantes chez les personnes jeunes. Ainsi les données manquantes de la variable « pression artérielle » vont dépendre de la variable « âge ».

Si nous ne pouvons pas supposer que la supposition *MAR* soit acceptable on dit que les données ne sont **pas manquantes au hasard *(MNAR)***[20]. Le mécanisme de leur génération ne peut plus être ignoré (on dit qu'il est **non ignorable**) ; mais les situations qui y aboutissent peuvent être très différentes, les mécanismes de description doivent être bien adaptés. De surcroît, il n'y a pas d'information dans le corpus qui puisse permettre de choisir un modèle approprié et aucune statistique qui indique la qualité de l'ajustement aux données. Pire, les résultats sont extrêmement sensibles au choix du modèle (voir chapitre 15 *Nonignorable Missing-Data Models* dans Little & Rubin, 2002). On comprend donc que la supposition du caractère ignorable soit la règle dans la majorité des logiciels.

Exemple : On recueille les revenus de plusieurs personnes. Les personnes ayant un revenu très élevé auront tendance à ne pas le donner. Ainsi, plus le revenu est haut et plus il y a un risque d'avoir une donnée manquante.

Il est important de noter qu'en pratique, il est bien souvent impossible de savoir si on se trouve en situation *MCAR*, *MAR* ou *MNAR*. Comme nous l'avons évoqué plus haut, on peut uniquement tester H_0 : « *MCAR* » contre H_1 : « non *MCAR* ». Il n'est par contre pas possible de déterminer si le processus de génération des données manquantes est *MAR* ou *MNAR* puisque les variables qui permettraient de déterminer l'appartenance à l'une de ces deux catégories n'ont pas pu être mesurées.

5. QUELLE METHODE POUR QUEL MECANISME ?

5.1 Méthodes conventionnelles : méthodes d'imputation simples

5.1.1. Principe d'estimation : la question des données manquantes est apparue très tôt dans le traitement des données statistiques ; la première approche est due à Yates (1933), la première automatisation de leur analyse est due à Healy & Westmacott (1956). L'idée de base est simple : (1) des valeurs approximatives sont substituées aux valeurs manquantes, (2) l'analyse des données est faite, (3) des valeurs prédites sont obtenues pour les données manquantes, (4) ces prédictions sont substituées aux valeurs manquantes, (5) une nouvelle analyse des données est faite. Ces étapes sont reprises jusqu'à ce que les valeurs manquantes ne soient plus modifiées ou, de façon équivalente, que la somme des carrés des résidus ne diminue plus. Cette manière de procéder est un

[20] En anglais : *missing not at random* (*MNAR.*)

exemple d'**algorithme *EM*** *(pour espérance-maximisation)*[21] maintenant largement utilisé. Très schématiquement,on a :

- Procédure itérative en 2 étapes :
  - Calcul de l'espérance : identification de la distribution des données manquantes en fonction des données observées et des variables explicatives,
  - Etape de maximisation : remplace les données manquantes par les valeurs estimées,
- Itération jusqu'à l'obtention d'une stabilisation des valeurs estimées.

On peut considérer qu'une méthode d'estimation doit satisfaire aux exigences suivantes :

1. Minimiser le biais,
2. Maximiser l'information disponible, éviter donc de supprimer au maximum des observations tout en essayant d'avoir des paramètres estimés de la plus grande précision possible,
3. Obtenir de bonnes estimations, c'est-à-dire des estimations précises des variances, des intervalles de confiances et des probabilités qui y sont attachées.

Il serait, de plus, idéal de ne pas faire de suppositions inutiles sur le mécanisme d'obtention. Seules les méthodes du maximum de vraisemblance et d'imputation multiple peuvent satisfaire « en partie » à ces exigences puisqu'elles nécessitent tout de même de supposer que les données manquantes sont ignorables (*MCAR* ou *MAR*).

5.1.2 Analyse des cas complets ou suppression des observations incomplètes : cette méthode est très souvent utilisée. Dans le cas de données *MCAR*, cela ne pose pas de problème majeur puisque le sous-échantillon de données complètes est un simple échantillon aléatoire de l'échantillon observé ; donc il ne peut pas introduire un biais ; mais la précision peut être dégradée puisque les résultats sont fondés sur moins de données que prévu. Par contre, dans le cas de données *MAR*, la suppression peut introduire un biais. Par exemple, nous voulons estimer le revenu moyen d'une population. Dans l'échantillon 80 % des femmes indiquent leur revenu, pour seulement 65 % des hommes : c'est une violation de type *MCAR*, pour chaque sexe l'absence ne dépend que du revenu. Si, comme on le sait, les hommes ont un revenu supérieur à celui des femmes, le revenu moyen sera sous-estimé. On peut cependant utiliser la méthode des cas complets sous l'hypothèse *MAR*, à condition d'ajuster le modèle sur les variables responsables de la non réponse de la variable d'intérêt.

Néanmoins, si le nombre de données manquantes est important, il vaut mieux utiliser d'autres méthodes d'analyse ou ne rien faire du tout si la qualité des données est jugée trop mauvaise.

5.1.3 Estimation par relation entre variables : on peut utiliser la relation entre les variables existantes pour estimer les données manquantes, par exemple en faisant des régressions de chaque variable sur les autres et en prenant le meilleur modèle.

[21] En anglais : *Expectation-Maximization.*

Quelquefois, cette procédure ne fonctionne pas car la matrice des corrélations entre les variables, dont les éléments sont calculés sur des échantillons de tailles différentes, n'est pas obligatoirement inversible. Estimer une valeur manquante par une moyenne peut entrer dans cette catégorie.

5.1.4 Utilisation de la ressemblance entre observations : si nous regardons les observations (les lignes du tableau de données), nous pouvons mesurer leur ressemblance par une distance calculée sur les variables présentes. On peut choisir les dix observations les plus voisines et estimer la valeur manquante par la médiane de ces dix observations ; si les variables sont discrètes on peut prendre le mode. On peut aussi prendre une distance pondérée ; si d est l'une des distances, pour estimer la donnée manquante on la pondérera avec un poids $w(d) = e^{-d}$; c'est ce qui est fait dans la library **DMwR** (Torgo, 2010). Bien que fortement décriées de nos jours, ces méthodes sont encore très souvent utilisées.

5.1.5 Exemple (calcium dans le sol et dans un légume, `CalciumRencher86a`) : prenons les données de Kramer & Jensen (1969) déjà utilisées par Rencher (1995) ; il s'agit d'un échantillon de $n = 10$ observations (lieu de culture) et $p = 3$ variables de teneur en calcium dans la même unité (milliéquivalents pour 100 g) :

$\boldsymbol{y}^1$: calcium disponible dans le sol

$\boldsymbol{y}^2$: calcium échangeable dans le sol

$\boldsymbol{y}^3$: calcium du navet vert poussé sur ce sol

Supposons que les deux valeurs entre parenthèses (3.5 pour $\boldsymbol{y}^2$ et 2.70 pour$\boldsymbol{y}^3$) soient manquantes (notée NA dans le fichier de données, cf. Tab. 7)

TABLEAU 7 - **`CalciumRencher86a`** :teneur en calcium (sol et navet).

Lieu	$\boldsymbol{y}^1$	$\boldsymbol{y}^2$	$\boldsymbol{y}^3$
1	35	~~(3.5)~~	2.80
2	35	4.9	~~(2.70)~~
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	35	4.6	2.88
8	35	10.9	2.90
9	35	8.0	3.28
10	30	1.6	3.20

Nous pouvons faire (sur les huit observations 3 à 10) la régression de $\boldsymbol{y}^2$ sur les deux variables pour obtenir l'estimation : $\hat{\boldsymbol{y}}^2 = -39.9433 + 0.1542\boldsymbol{y}^1 + 13.8011\boldsymbol{y}^3$; appliquée aux valeurs 35 et 2.80 des deux régresseurs, elle donne $\hat{y}^2 = 4.097$ à comparer à la valeur observée de 3.5. De même la régression (sur les mêmes huit observations) de $\boldsymbol{y}^3$ sur les deux autres variables donne : $\hat{\boldsymbol{y}}^3 = 2.814291 - 0.001377\boldsymbol{y}^1 + 0.049941\boldsymbol{y}^2$; appliquée aux valeurs 35 et 4.9 des deux régresseurs, elle donne $\hat{y}^3 = 3.011$ à comparer à la valeur observée de 2.70.

Si maintenant nous remplaçons les valeurs manquantes par les valeurs que nous venons d'estimer, nous avons un échantillon complet de dix observations. Nous

recommençons la procédure. Avec la nouvelle estimation $\hat{\boldsymbol{y}}^2 = -40.5136 + 0.1388\boldsymbol{y}^1 + 14.0554\boldsymbol{y}^3$, nous obtenons une nouvelle estimation pour la valeur manquante $\hat{y}^2 = 3.698$. Avec l'autre variable $\hat{\boldsymbol{y}}^3 = 2.82886 - 0.00301\boldsymbol{y}^1 + 0.05191\boldsymbol{y}^2$, nous obtenons l'autre estimation pour la valeur manquante $\hat{y}^3 = 2.978$. On peut poursuivre jusqu'à ce que les estimations ne changent plus. Si on compare ces valeurs estimées aux moyennes qui auraient pu être employées ($\bar{y}^2 = 7.589$, $\bar{y}^3 = 3.132$), elles sont à première vue meilleures, plus proches des valeurs réellement observées.

5.2 Estimation par Maximum de vraisemblance

La méthode du maximum de vraisemblance s'est avérée excellente pour la manipulation des données manquantes (Allison, 2001). La majorité des logiciels supposent l' « ignorabilité » donc le processus originel des données manquantes est *MAR*. Alors la formalisation est assez simple : il faut une fonction de vraisemblance qui exprime la probabilité des données comme une fonction des paramètres inconnus. Soit deux variables aléatoires discrètes X et Z, leur fonction de probabilité conjointe est $p(x, z|\theta)$ où θ est un vecteur de paramètres ; ce qui exprime que $p(x, z|\theta)$ donne la probabilité que $X = x \; et \; Z = z$.

Sans donnée manquante, la fonction de vraisemblance s'écrit :

$$L(\theta) = \prod_{i=1}^{n} p(x_i, z_i|\theta)$$

C'est cette fonction que nous rendons maximale pour estimer la valeur de θ.

Supposons que les données soient *MAR* pour les r premières observations de Z, et *MAR* pour X pour les s suivantes. Soit :

$$g(x|\theta) = \sum_{z} p(x, z|\theta)$$

la distribution marginale de $\boldsymbol{X}$ (en faisant la somme sur Z) et soit :

$$h(z|\theta) = \sum_{x} p(x, z|\theta)$$

la distribution marginale de Z (en faisant la somme sur $\boldsymbol{X}$).

La fonction de vraisemblance est alors :

$$L(\theta) = \prod_{i=1}^{r} g(x_i|\theta) \prod_{i=r+1}^{r+s} h(z_i|\theta) \prod_{i=r+s+1}^{n} p(x_i, z_i|\theta)$$

Ce qui exprime que cette fonction est factorisée en parties correspondant aux différentes structures présentées par les données manquantes. Si les variables sont continues et non discrètes, les sommations sont remplacées par des intégrales. Pour modéliser le maximum de vraisemblance pour des données manquantes, il faut donc un modèle pour la distribution conjointe des variables et une méthode numérique pour rendre maximum cette vraisemblance. Avec des données discrètes, on peut choisir un

modèle multinomial ou un modèle log-linéaire. Avec des données multiNormales, la vraisemblance peut être maximisée soit directement, soit par un algorithme *EM*. (Dempster et *al.*, 1977). Dans ce cas, les paramètres à estimer sont les moyennes, variances et covariances ; la description détaillée est fournie par Allison (2001)[22].

Cette méthode, qui est à relier à la méthode de pondération inverse dans le cas de données manquantes de type *MAR*, permet de tenir compte des données manquantes en donnant un poids plus important aux données qui ont une forte probabilité d'être manquantes. Soit Z une variable partiellement observée, X une variable totalement observée et M définie par :

$$M = \begin{cases} 1 \; si \; Z \; est \; observée \\ 0 \; sinon \end{cases}$$

Le poids utilisé pour une observation est en fait l'inverse de sa probabilité d'être observée, que l'on peut prédire par une régression logistique de M sur X.

En pratique, la méthode de la pondération inverse se résume en 3 étapes :

1. Faire une régression logistique de M sur X,
2. Pour chaque observation i, obtenir la valeur prédite $\hat{p}_i$ de $p_i = P[M_i = 1/X_i]$,
3. Réaliser l'analyse souhaitée en attribuant à chaque observation le poids $1/\hat{p}_i$.

5.3 Un exemple (`Eaux2010`)

Pour illustrer les procédures conventionnelles, nous prenons le corpus complet de 113 eaux minérales. On peut résumer et visualiser les données manquantes grâce à la fonction **`aggr`** de la library **`VIM`**. On peut compter le nombre de cellules avec données manquantes (**`countNA(Eaux2)`**, soit 39), le nombre d'échantillons incomplets (26 dans **`Eaux2[!complete.cases(Eaux2),]`**) et extraire un fichier de données complètes (87 observations dans **`EauxComplete`**) (Tab.8A et Fig.5).

TABLEAU 8A - **`Eaux2010`** : fichier des 113 eaux minérales.

```
> library(DMwR) ; library(VIM)
> Eaux2010<-read.table("Eaux2010.txt",h=T, row.names=7)
> dim(Eaux2010)
     [1] 113   8
> Eaux2<-Eaux2010[,1:6]
> dim(Eaux2)
     [1] 113   6
> countNA(Eaux2)
     [1] 39
# Fig.6, avec un paramétrage des couleurs.
> aggr(Eaux2, col=c("yellow","grey"))
> Eaux2.agr= aggr(Eaux2)
```

```
> summary(Eaux2.agr)
     Missings per variable:
      Variable Count
          HCO3     6
           SO4    10
            Cl    12
            Ca     2
            Mg     3
            Na     6
         Sigle     0
     Missings in combinations of variables:
      Combinations Count    Percent
     0:0:0:0:0:0:0    87 76.9911504
     0:0:0:0:0:1:0     2  1.7699115
     0:0:0:0:1:0:0     1  0.8849558
     0:0:1:0:0:0:0     6  5.3097345
     0:0:1:0:0:1:0     1  0.8849558
     0:0:1:0:1:0:0     1  0.8849558
     0:1:0:0:0:0:0     5  4.4247788
     0:1:1:0:0:0:0     2  1.7699115
     0:1:1:0:0:1:0     1  0.8849558
     0:1:1:1:0:0:0     1  0.8849558
     1:0:0:0:0:0:0     2  1.7699115
     1:0:0:0:0:1:0     2  1.7699115
     1:0:0:1:0:0:0     1  0.8849558
     1:1:0:0:1:0:0     1  0.8849558
> Eaux2[!complete.cases(Eaux2),]
          HCO3    SO4     Cl  Ca     Mg     Na
Kam     NA    5.0  4.50   5   9.0  10.00
Cub    354   16.0 22.00 112   3.0     NA
Ama    312  372.0    NA 176  46.0  28.00
Hep    403 1479.0    NA 555 110.0     NA
Par    223     NA 19.00  65   5.0  20.00
Wat    135   24.0    NA  35  11.0   3.00
Dus    329  110.0    NA  80  40.0  18.00
Ker    487  144.0    NA 144  34.0  32.00
Ame    429     NA 18.00 157   8.0  14.00
Ann     29    8.0    NA  12    NA   1.00
Gio    113     NA 86.00  29  16.0  45.00
Lar     31    1.4  0.40   1   0.3     NA
Lei     59   14.0    NA  21   2.0   2.00
Lev     74     NA 73.00  15  14.0  39.00
Pra    476     NA  9.00 140  12.0   4.00
Vam     NA    5.0  0.26  NA   2.7   1.35
Mal     NA     NA 65.00  35    NA  15.00
Car     36   26.0  7.00   3    NA  15.00
Hig    133    7.0    NA  32   8.0  45.00
Ays     NA    0.0  7.00  27   5.0     NA
Oru     NA   14.0  9.00  26   6.0     NA
Sal     NA    6.0  7.00  36   9.0   7.00
Bad   1700     NA    NA 272 102.0 180.00
Kek   1150     NA    NA 246  56.0  36.00
San   1360     NA    NA  NA  60.0 290.00
Sve   1220     NA    NA 228  38.0     NA
> nrow(Eaux2[!complete.cases(Eaux2),])
     [1] 26
> EauxComplete<-na.omit(Eaux2)
> dim(EauxComplete)
     [1] 87  6
```

[22] Dans le logiciel SAS la procédure s'appelle `PROC MI`. Avec R, nous utilisons **`DMwD`** (Torgo, 2010) ; une autre library est aussi fort intéressante **`Amelia`** (Honaker et *al.*, 2010).

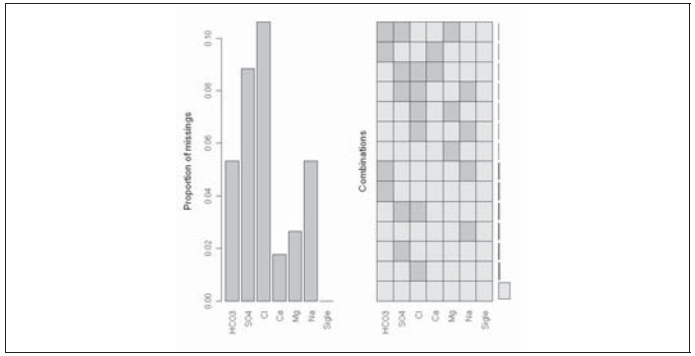

FIGURE 6. **Eaux2010** *: visualisation des données manquantes.*

On peut calculer la matrice des coefficients de corrélation et la comparer aux valeurs du *scatterplot* de la figure 6 du chapitre 2. (Tab.8B).

TABLEAU 8B - **Eaux2010** : coefficients de corrélation entre les 6 variables et présentation schématique.

```
> round(cor(EauxComplete[,1:6]),3)

      HCO3   SO4    Cl    Ca    Mg    Na
HCO3 1.000 0.077 0.642 0.280 0.365 0.872
SO4  0.077 1.000 0.078 0.815 0.497 0.051
Cl   0.642 0.078 1.000 0.140 0.378 0.746
Ca   0.280 0.815 0.140 1.000 0.591 0.083
Mg   0.365 0.497 0.378 0.591 1.000 0.207
Na   0.872 0.051 0.746 0.083 0.207 1.000
```

```
> symnum(cor(EauxComplete[,1:6]))
     H S Cl Ca M N
HCO3 1
SO4    1
Cl   ,   1
Ca     +    1
Mg   . . .  .  1
Na   +   ,       1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8
'+' 0.9 '*' 0.95 'B' 1
```

On peut alors estimer les données manquantes de deux manières différentes et comparer les résultats (Tab.8C et Tab.8D).

TABLEAU 8C - **Eaux2010** : estimation des données manquantes.

```
# Estimation des valeurs NA par distance pondérée des 10 obs les + proches
> Eaux2[,1:6]<-knnImputation(Eaux2[,1:6],k=10)
> EauxReconsMoy<-Eaux2 ; dim(EauxReconsMoy)
     [1] 113   7
    ...
# Estimation des valeurs NA par la médiane des k =10 voisins
> Eaux2<-Eaux2010[,1:6]
> Eaux2[,1:6]<-knnImputation(Eaux2[,1:6],k=10,meth="median")
> EauxReconsMed<-Eaux2 ; dim(EauxReconsMed)
     [1] 113   6
# Collage des 3 matrices base(avec NA) et les deux reconstituées
> EauxReconsVerif<-cbind(Eaux2010[,1:6],EauxReconsMoy[,1:6], EauxReconsMed[,1:6])
> nrow(EauxReconsVerif[!complete.cases(EauxReconsVerif),])
     [1] 26
# Impression des observations avec les 2 estimations de valeurs manquantes
> round(EauxReconsVerif[!complete.cases(EauxReconsVerif),],1)
```


Le tableau 8D nous permet de voir que les grandes différences entre les deux types d'estimation sont surtout nettes sur les quatre dernières. Plus généralement, une estimation de données manquantes fournit des valeurs qu'il faut examiner avec grande attention. Quand on observe deux valeurs très différentes d'une même quantité, il faut obligatoirement s'interroger sur la qualité de la reconstitution et essayer d'en comprendre la raison. Ici elle apparaît moins bonne pour les valeurs élevées HCO_3.

TABLEAU 8D - **Eaux2010** : résultats (extrait) de l'estimation des données manquantes.

	Données initiales					
Sig	HCO_3	SO_4	Cl	Ca	Mg	Na
Kam	**NA**	5.0	4.5	5	9.0	10.0
Cub	354	16.0	22.0	112	3.0	**NA**
Ama	312	372.0	NA	176	46.0	28.0
Hep	403	1479.0	NA	555	110.0	NA
...						
Bad	1700	**NA**	NA	272	102.0	180.0
Kek	1150	**NA**	NA	246	56.0	36.0
San	1360	**NA**	NA	**NA**	60.0	290.0
Sve	1220	**NA**	NA	228	38.0	**NA**
	Estimation par moyennes pondérées					
Sig	HCO_3	SO_4	Cl	Ca	Mg	Na
Kam	70.0	5.0	4.5	5.0	9.0	10.0
Cub	354.0	16.0	22.0	112.0	3.0	**12.4**
Ama	312.0	372.0	35.6	176.0	46.0	28.0
Hep	403.0	1479.0	11.0	555.0	110.0	**11.5**
...						
Bad	1700.0	283.7	108.0	272.0	102.0	180.0
Kek	1150.0	216.7	34.9	246.0	56.0	36.0
San	1360.0	139.4	61.8	**143.1**	60.0	290.0
Sve	1220.0	157.2	65.1	228.0	38.0	**145.2**
	Estimation par médianes					
Sig	HCO_3	SO_4	Cl	Ca	Mg	Na
Kam	**73.0**	5.0	4.5	5.0	9.0	10.0
Cub	354.0	16.0	22.0	112.0	3.0	**12.5**
Ama	312.0	372.0	**15.0**	176.0	46.0	28.0
Hep	403.0	1479.0	**15.0**	555.0	110.0	**11.5**
...						
Bad	1700.0	**261.5**	**19.0**	272.0	102.0	180.0
Kek	1150.0	**221.0**	**16.5**	246.0	56.0	36.0
San	1360.0	**42.0**	**27.0**	**107.5**	60.0	290.0
Sve	1220.0	**142.5**	**19.0**	228.0	38.0	**43.5**

On peut contrôler ces résultats en comparant les moyennes des observations complètes (les 87 observations) et les deux tableaux estimés (Tab.8E) à l'aide de tests de *Student* résumés dans le tableau 8F.

TABLEAU 8E - **Eaux2010** : comparaison de l'échantillon complet aux deux estimations (n=26, moyenne pondérée ou médiane).

```
> for (i in 1:6  )
+    {
+ a1<-t.test(EauxReconsVerif[complete.cases(EauxReconsVerif),i] ,
  EauxReconsVerif[!complete.cases(EauxReconsVerif),i+6])
+ print(a1$stat)
+ print(a1$p.value)
+ print(a1$estim)
+    }
…
> for (i in 1:6  )
+    {
+ a1<-t.test(EauxReconsVerif[complete.cases(EauxReconsVerif),i] ,
  EauxReconsVerif[!complete.cases(EauxReconsVerif),i+12])
+ print(a1$stat)
+ print(a1$p.value)
+ print(a1$estim)
+    }
```

On peut voir que les moyennes des échantillons ne sont pas différentes pour l'estimation par moyenne pondérée, mais sont à la limite de signification pour Cl et Na avec la seconde estimation. Évidemment ceci est valable pour l'exemple choisi. La prise en compte d'une information supplémentaire (eau plate ou eau gazeuse) pourrait sans doute améliorer les résultats.

TABLEAU 8F - **Eaux2010** : résultats de la comparaison de l'échantillon complet aux deux estimations (n=26, moyenne pondérée ou médiane).

Variable	HCO_3	SO_4	Cl	Ca	Mg	Na
Echantillon complet (n=87)	495	96	43	90	22	95
Estimation par moyenne pondérée	377	121	27	100	24	39
t *(Student)*	1.00	-0.40	1.23	-0.38	-0.31	1.79
P	0.323	0.692	0.221	0.707	0.758	0.076
Estimation par médiane	374	116	18	99	24	34
t *(Student)*	1.02	-0.31	1.99	-0.32	-0.22	1.96
p	0.311	0.756	0.049	0.751	0.824	0.053

5.4 L'imputation multiple

5.4.1. Idées générales : l'imputation simple n'est pas toujours une bonne manière de procéder. En effet, les analyses qui en découlent ne font pas de distinction entre données réellement observées et données imputées, c'est-à-dire qu'elles ne prennent pas en compte la variabilité des valeurs imputées. De ce fait, l'incertitude liée aux données manquantes est complètement occultée. Ceci conduit en général à sous-estimer les variances. L'imputation multiple est une méthode d'analyse de données multidimensionnelles incomplètes qui va apporter une solution à ce problème en créant plusieurs valeurs possibles d'une valeur manquante. En faisant cette opération, nous allons tenir compte de l'incertitude liée aux valeurs manquantes et des corrélations entre les variables. Une des library bien développée sous *R* est **Amelia** (Honaker et *al*., 2002 ; Honaker & King , 2010, Honaker et *al*., 2010), c'est celle que nous avons utilisée.

L'algorithme d'**Amelia** prend un ensemble de données de dimension équivalente à celui de départ ; il en effectue un ré-échantillonnage par *bootstrap* ; il en estime les statistiques par *EM*, puis il remplace les valeurs manquantes. Il suppose que les données sont *MAR*. Ce processus est répété *m* fois ; il produit donc *m* ensembles de données complètes, où les valeurs observées sont les mêmes et les valeurs non observées sont tirées de leurs distributions *a posteriori*. L'algorithme conserve les caractéristiques importantes de la distribution et les relations entre variables. Mais il n'a pas pour objectif de prédire les données manquantes avec une précision maximale. Le nombre de tirages n'est pas très élevé, la valeur par défaut de *m* est 5.

On peut considérer qu'il y a trois étapes :

Etape 1 : remplacement des valeurs manquantes par *m* valeurs provenant d'une distribution ad hoc, cf. Fig.7A.

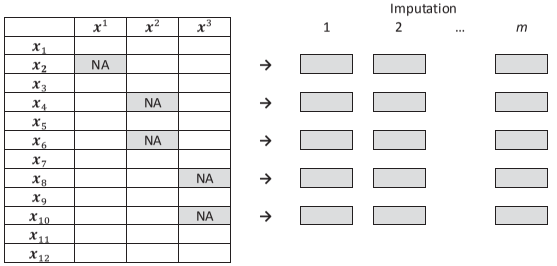

FIGURE 7A. *Imputation multiple : remplacement des valeurs manquantes.*

Etape 2 : analyse statistique de chacun des *m* corpus de données complètes, cf. Fig.7B.

Corpus de données avec valeurs observées et estimées

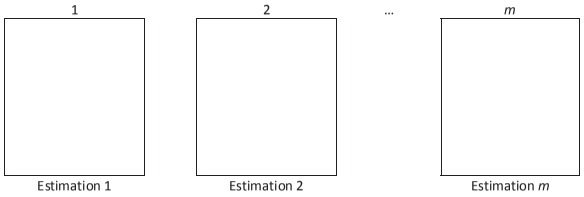

FIGURE 7B. *Imputation multiple : analyses indépendantes des corpus reconstitués.*

Etape 3 : combinaison des résultats obtenus qui reflètent la variabilité supplémentaire due aux données manquantes, cf. Fig.7C.

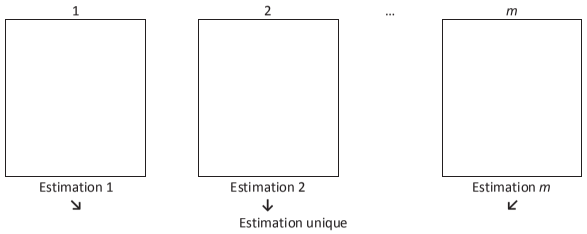

FIGURE 7C. Imputation multiple : combinaison des résultats des analyses indépendantes.

En conclusion on peut dire que c'est une approche très générale. Un ensemble de m imputations permet d'effectuer plusieurs analyses. Le résultat final introduit l'incertitude liée à l'absence de certaines données, sans qu'on sache la mesurer précisément. Il n'est pas nécessaire d'avoir une valeur de m élevée.

5.4.2. Suppositions et algorithme : le modèle utilisé dans **`Amelia`** suppose que l'ensemble des données (celles qui sont observées et celles qui ne le sont pas) suivent une loi multiNormale. Si nous notons le corpus de données X (la partie observée étant X^{obs} et la partie non observée X^{man}), la supposition est :

$$X \sim N_p(\mu, \Sigma)$$

Si nous appelons M la matrice indicatrice des données manquantes de même dimension que X, à savoir $m_{ij} = 1$ si $d_{ij} \in X^{man}$ et $m_{ij} = 0$ sinon. La supposition *MAR* s'écrit :

$$p(M|X) = p(M|X^{obs})$$

Dans le modèle d'imputation, nous considérons les paramètres sur les données complètes $\theta = (\mu, \Sigma)$. La vraisemblance des données est $p(X^{obs}, M|\theta)$. La supposition *MAR* permet de la séparer en deux :

$$p(X^{obs}, M|\theta) = p(M|X^{obs})p(X^{obs}|\theta)$$

Comme nous ne pouvons inférer que sur les paramètres des données complètes, nous pouvons écrire que la vraisemblance est :

$$L(\theta|X^{obs}) \propto p(X^{obs}|\theta)$$

qui peut être réécrite en utilisant la *loi des espérances itérées* :

$$p(X^{obs}|\theta) = \int p(X|\theta)dX^{man}$$

Avec cette vraisemblance nous pouvons voir que la distribution *a posteriori* est :

$$p(\theta|X^{obs}) \propto p(X^{obs}|\theta) = \int p(X|\theta)dX^{man}$$

La difficulté principale dans l'analyse des données incomplètes provient du tirage au sort dans cette probabilité. L'algorithme *EM* (Dempster et *al.* 1977.) est une approche calculatoire pour en trouver le mode. L'algorithme *EMB* d'**`Amelia`** combine le classique *EM* avec l'approche *bootstrap*.

5.4.3. Les données `Eaux2010` retraitées : sans entrer dans les détails, nous donnons quelques graphiques permettant de mieux juger la qualité et la fiabilité de la reconstitution des données manquantes. Dans le fichier `Eaux2010`, il y a 6, 10, 12, 2, 3 et 6 données manquantes pour les variables HCO_3, SO_4, `Cl`, `Ca`, `Mg` et `Na`. Un graphique comme celui de la figure 8A permet de contrôler la distribution des imputations.

```
> library(Amelia)
> Eaux2<-Eaux2010[,1:6]
> Eaux2.Amelia.out<-amelia(x=Eaux2[,1:6])
> plot(Eaux2.Amelia.out)
```

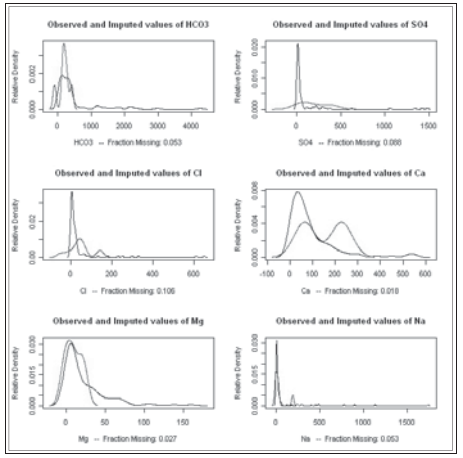

FIGURE 8A. **`Eaux2010`** *: distribution des imputations moyennes tracée sur la distribution des valeurs observées.*

```
> overimpute(Eaux2.Amelia.out,var=1,main="Observation/Estimation pour la
  variable HCO3",xlab="Valeurs observées",ylab="Valeurs estimées")
> overimpute(Eaux2.Amelia.out,var=4,main="Observation/Estimation pour la
  variable Ca",xlab="Valeurs observées",ylab="Valeurs estimées")
```

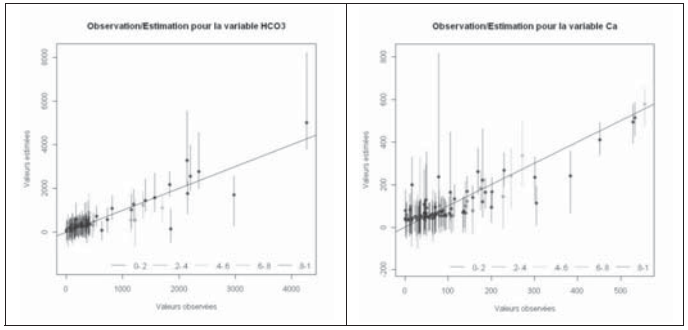

FIGURE 8B. **Eaux2010** *: étude de l'ajustement du modèle d'imputation.*

La fonction dans **Amelia** traite chaque observation comme si elle était manquante et estime sa valeur sur la base du modèle choisi. Les points indiquent les valeurs estimées et les lignes verticales l'intervalle de confiance à 90 %. La diagonale est la droite $y = x$. Si les estimations étaient parfaites, tous les points seraient sur la diagonale ; pour que le modèle soit bon, 90 % des lignes doivent couper cette diagonale. On utilise une technique de **sur-imputation**[23] développée pour juger de l'ajustement du modèle d'imputation. En fait, il est impossible de dire si la prédiction moyenne est proche ou non de la valeur non observée, puisqu'on ne connaît pas cette dernière. Par nature même, elle n'existe pas pour faire cette comparaison ! Si elle existait le problème ne se poserait pas. Par contre, on peut s'interroger sur sa précision. La sur-imputation implique le traitement séquentiel des valeurs observées comme si elles avaient été manquantes. On peut regarder graphiquement si les valeurs observées tombent bien à l'intérieur de l'intervalle qui aurait été estimé (Fig.8B). On peut même faire un diagnostic visuel de la convergence *EM*. Pour plus de détails voir Honaker et *al.* (2010). En conclusion : nous avons les moyens d'estimer des valeurs qui nous manquent. Mais, soyons sans illusion : ce qui nous manque ne peut être que partiellement restitué ; il ne faudra jamais l'oublier.

6. BILAN

Il est aussi possible d'utiliser les tests de permutation dans le cadre du modèle linéaire (cf. par exemple la library **lmPerm** sous *R*).

A condition d'avoir un nombre raisonnable (moins de 10 %) de données manquantes, l'analyse des cas complets peut être envisagée. Dans le cas d'études longitudinales, elle est réalisée grâce au **modèle mixte** (Cf. Chapitres 9 et 14). Cette

[23] En anglais : *overimputing*.