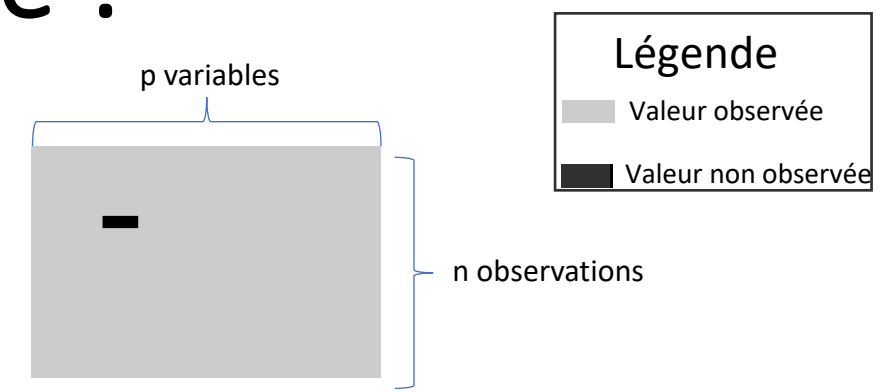
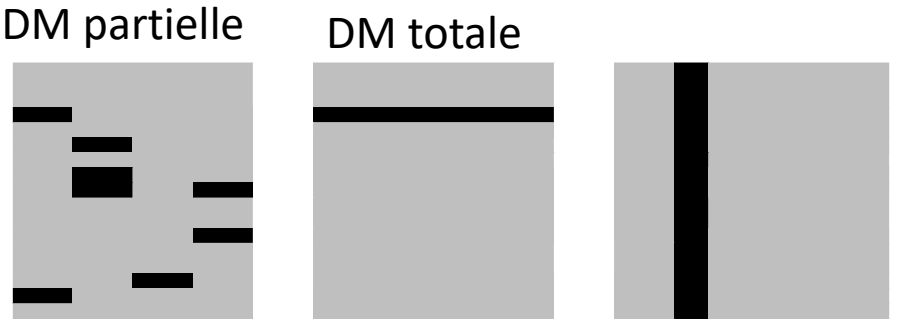
Source image : https://ledidi.com/academy/missing-data-mechanisms-and-how-to-handle-it

Qu’est-ce qu’une donnée manquante ?

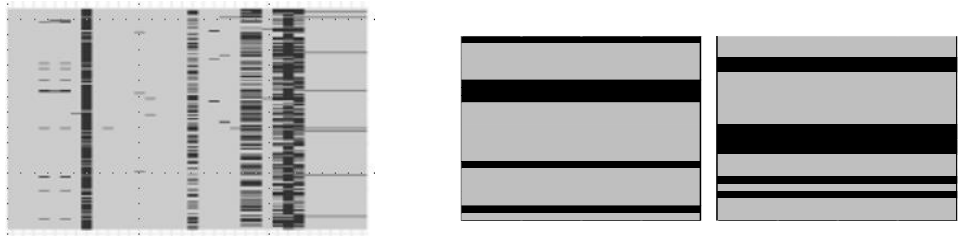
- Une **donnée manquante** (DM) : valeur non observée pour un couple variable-observation



- **DM partielle** (*item non-response*): pour un individu donné, seules quelques valeurs non observées
- **DM totale** (*unit non-response*): toutes les variables d’un individu donné sont non observées

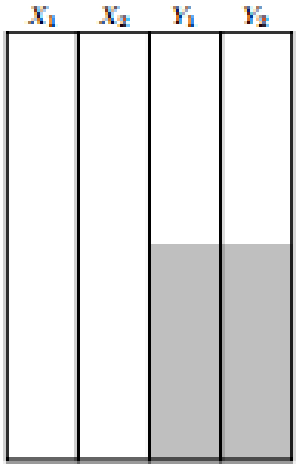


- Lorsqu’on parle de données manquantes dans un sens plus général, il s’agit de l’ensemble des données manquantes



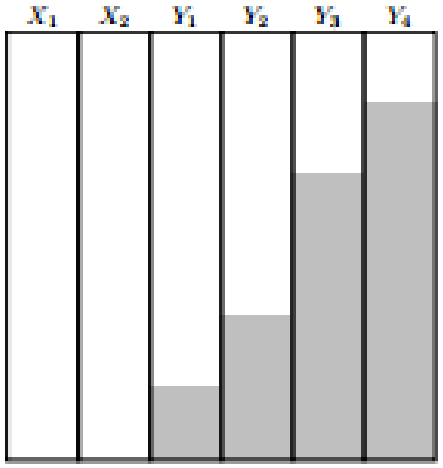
- Sous R, mentionné par le terme NA (*Not Available*) ≠ NaN (Not A Number) ➛division par 0

Répartition de données manquantes



Structure univariée

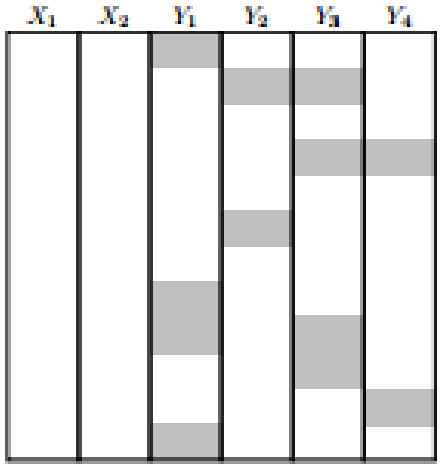
Mêmes individus ont des DMs pour les mêmes d < p variables



Monotone

Si les variables peuvent être ordonnées de telle sorte que lorsque l'observation est manquante pour un individu i pour une variable, alors toutes les autres variables suivantes sont aussi manquante

➛ fréquent données longitudinales



Sans structure

DMs réparties sans structure particulière dans le jeu de donnée

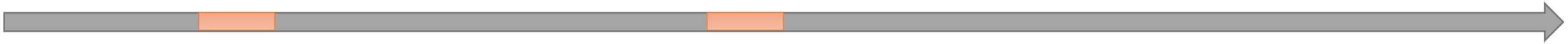
Valeur manquante

→ **Adapter notre stratégie de traitement**

- Exclure des variables ou des individus
- Collecter de nouvelles données (quand cela est possible)
- Estimer/remplacer les données manquantes : **imputation**, quelle(s) méthode(s) choisir pour cela ?

Cas des données longitudinales

Y complètement observé. L'individu s'est présenté à toutes les visites



Y manquante intermittente. L'individu a manqué quelques visites.

Y manquante monotone. L'individu est perdu de vue (ne vient plus à partir d'une certaine date).

Valeurs manquantes

Bien comprendre ses données

- Valeurs manquantes ➛ pas pu être observées, perdues, incohérentes
- Pas toujours « NA », ou une case vide dans les jeux de données
  1. Données omiques, spectrométrie de masse ➛ présence de 0
     - Soit la variable (protéine, métabolite, etc.) est vraiment absente ➛ vrai 0
     - Soit < seuil de détection ➛ présente mais en **très faible quantité**
     - Problème technique : exemple en protéomique, MS/MS la protéine « se coupe mal », peptide non identifié ➛ valeur mise à 0 ➛ **?**
     - Etc.

     ➔ **Méconnaissance du mécanisme réel de la donnée manquante**
  2. Données météorologiques : -999, code des données manquantes pour la température
- Autre exemple particulier : données longitudinales
  - L'individu absent à un temps t ➛ Pourquoi ?
  - L'individu quitte l'étude (notion de censure ➛ données manquantes particulières)

Quelques notations

- Soit un vecteur $Y = (Y_1, \dots, Y_p)$ de p variables aléatoires (numériques ou catégorielles)
- **R matrice indicatrice des valeurs manquantes** $(r_{ij})_{i=1,\dots,n, j=1,\dots,p}$

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

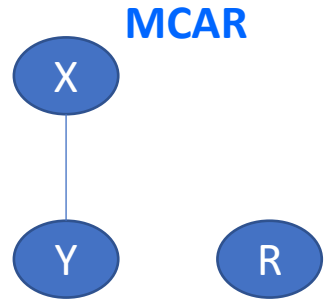
- Y_{obs} : parties observées de Y
- Y_{miss} : parties manquantes de Y

$$Y = RY_{obs} + (1 - R)Y_{miss}$$

- **Mécanisme des données manquantes** : $f(R|Y)$ distribution conditionnelle de R sachant Y
  - Peut dépendre de paramètre ψ ➛ $f(R|Y;\psi)$
  - Présence de covariables $(X_j)_{j=1,\dots,q}$ complètement observées ➛ $f(R|Y,X;\psi)$

Mécanisme des données manquantes

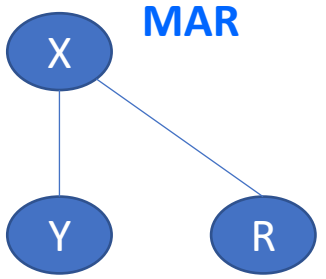
Little & Rubin (2002)



- **Missing Completely At Random (MCAR)**

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

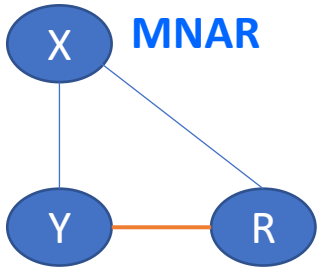
Exemple : Une personne oublie accidentellement de répondre à une question pendant une enquête ou sous-échantillon pour des mesures trop coûteuses



- **Missing At Random (MAR)**

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

Exemple : personnes âgées (X = âge) ont plus tendance à refuser de donner leur revenu (Y)



- **Missing Not At Random (MNAR)**

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

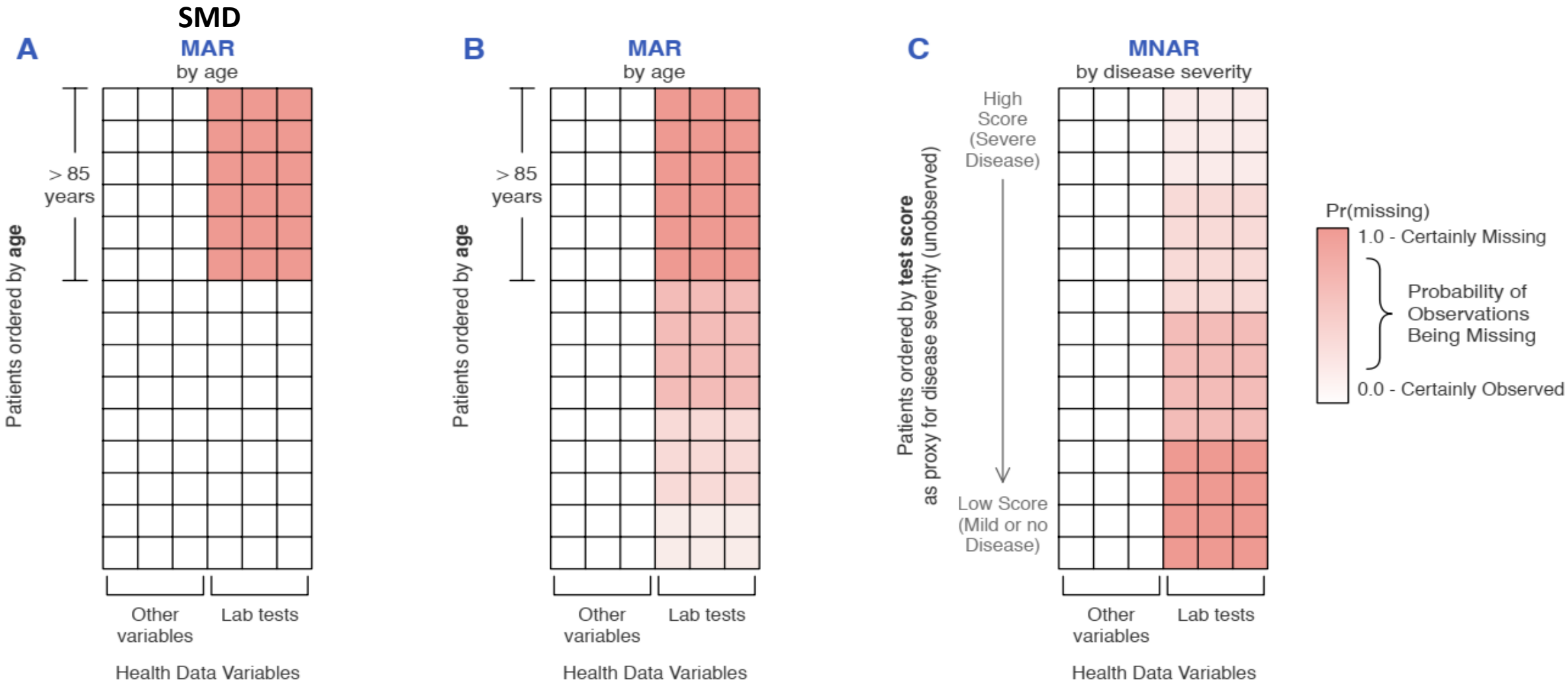
Exemple : Question sensible

- Mécanisme **ignorable** : données manquantes aléatoirement + les paramètres régissant le mécanisme de génération de DM et des données doivent être « distinguables »

- **Structurally missing data (SMD)** (*Mitra et al., 2023*) **:** fait référence à des motifs multivariés non aléatoires d'absence de données dans un ensemble de données

Exemple : âge du premier enfant alors que l'individu n'a pas d'enfant, réunit plusieurs bases de données

Illustration MAR vs MNAR



Source : figure 2 de *Mitra et al. (2023)*

Pourquoi doit-on gérer les données manquantes ?

« The best thing to do with missing data is to not have any »

Gertrude Mary Cox

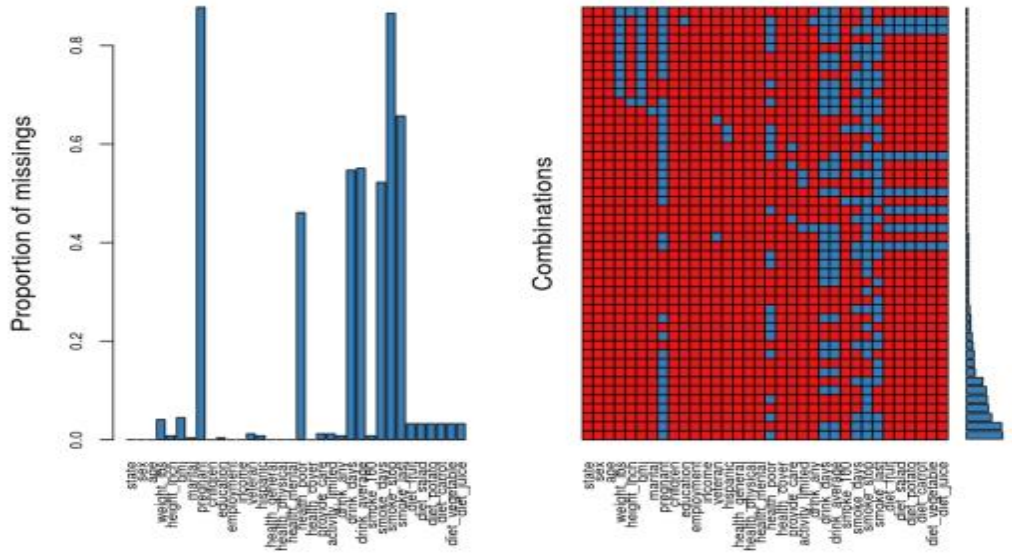
Objectif :

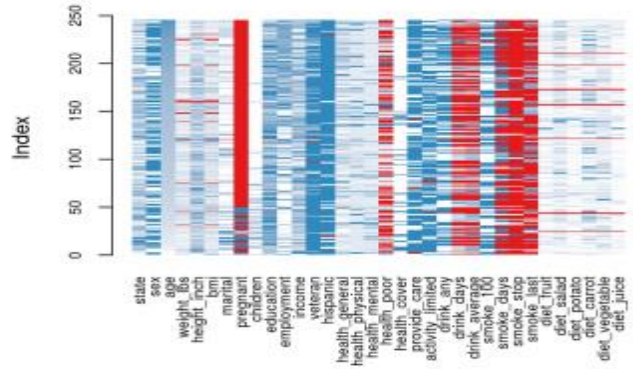
- Limiter la perte de précision
- Limiter les biais d'estimation dans les méthodes d'inférence
- Préserver les caractéristiques essentielles des données : distribution des variables, relations entre les variables
- pas retrouver les vraies valeurs

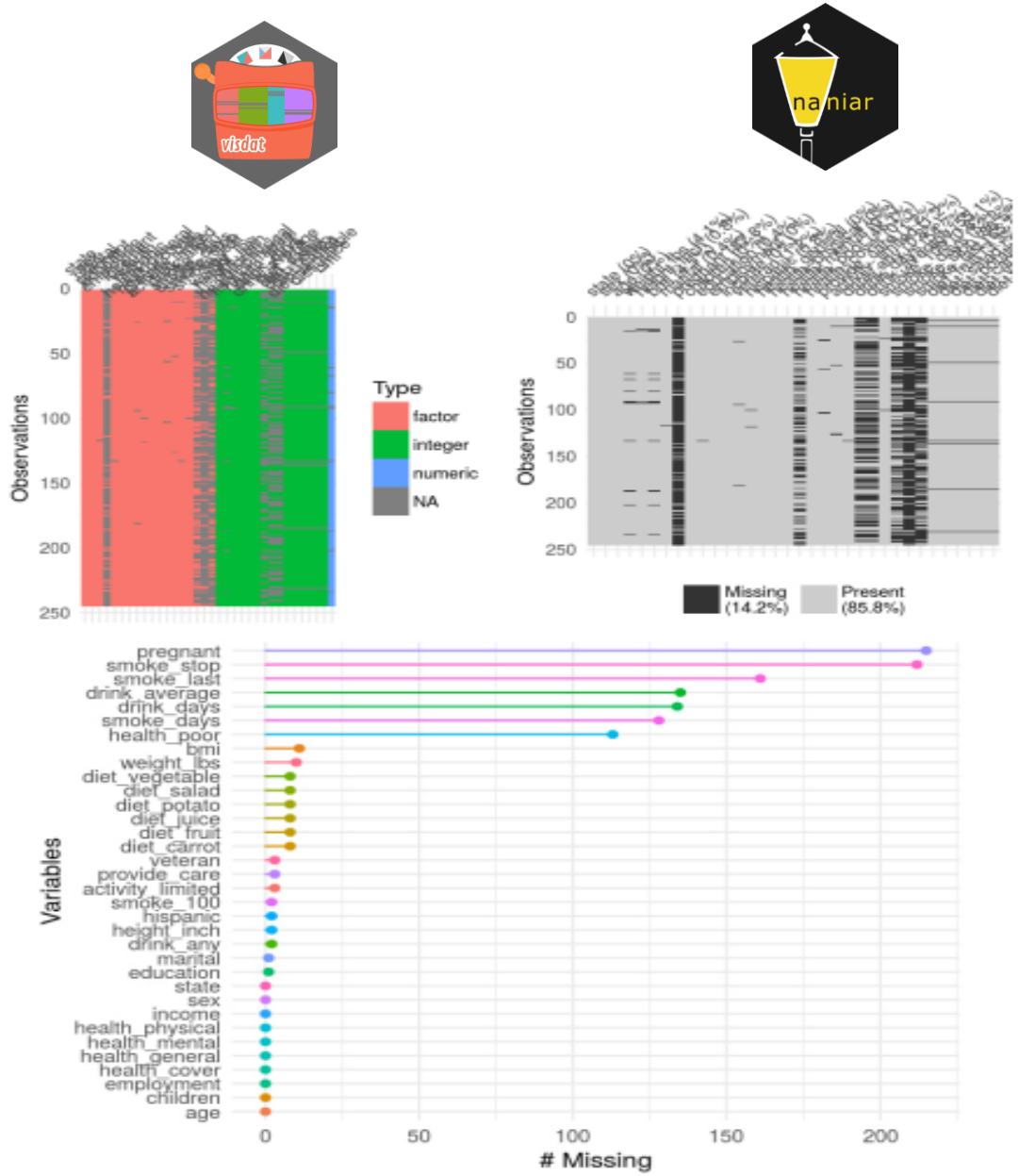
Note : première façon d'éviter les données manquantes : bien construire son plan expérimental

- ➢ particulièrement vrai pour des études longitudinales, ou des enquêtes auprès de personnes

Visualiser les données manquantes





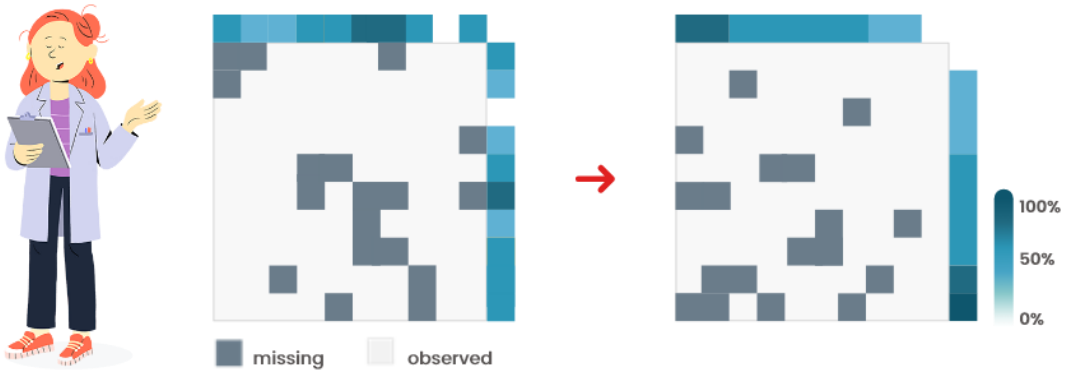


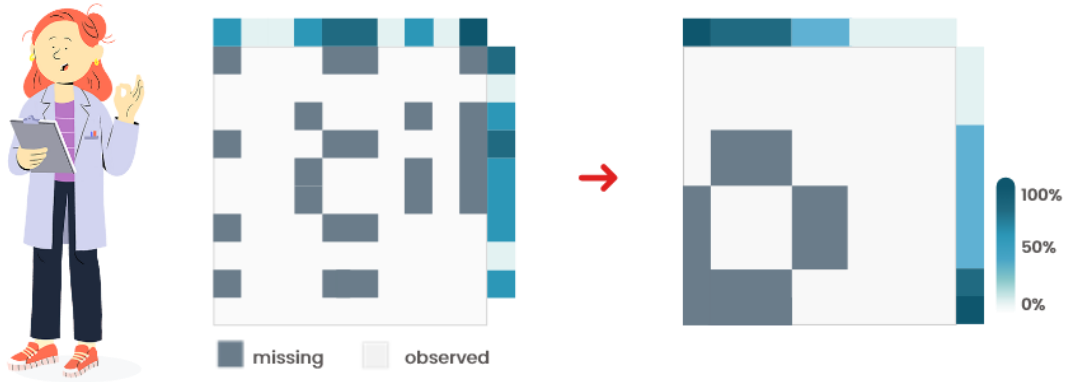
R package ***VIM***

Permet anomalies ou erreurs dans les données imputées

Heatmap

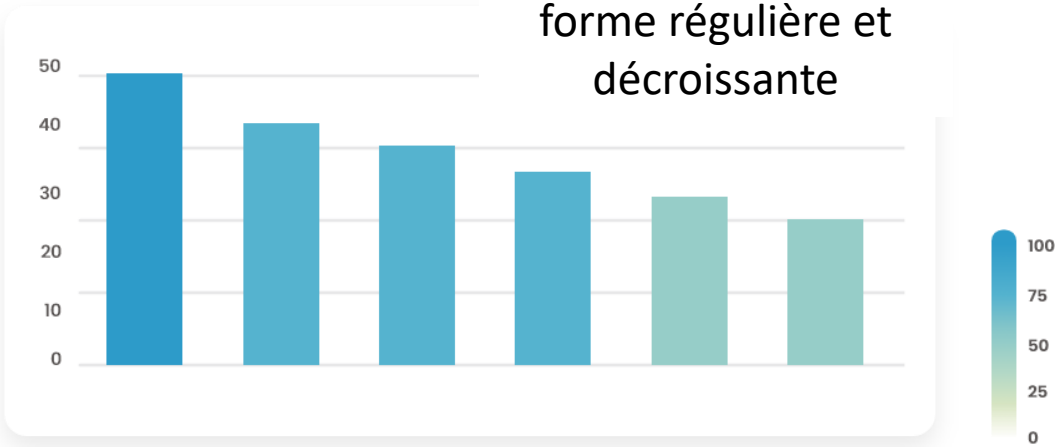
Source image





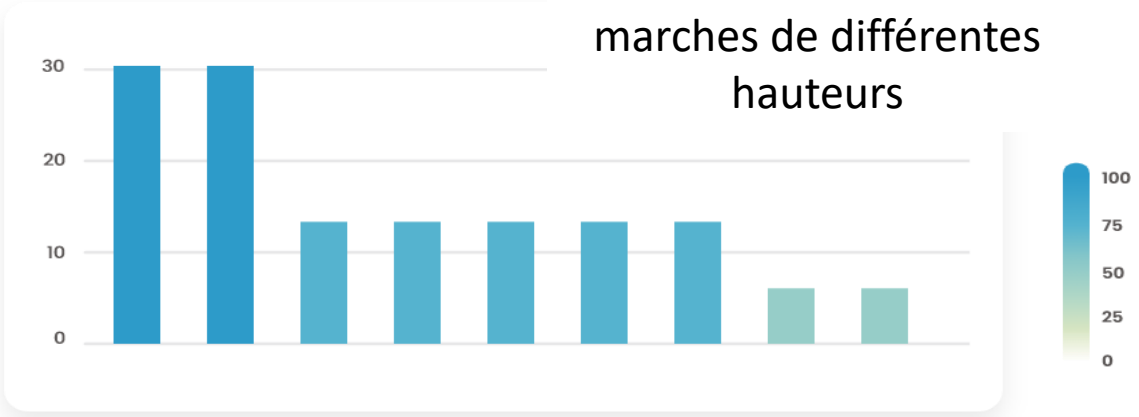
Pattern particulier des données manquantes : potentiellement présence de MNAR

Barplots



forme régulière et décroissante

Envisager d'imputer les valeurs manquantes (si la proportion de valeurs manquantes n'est pas trop élevée, inférieure à 30 % par exemple).



marches de différentes hauteurs

Imputation ou supprimer ligne/colonne ➛ peut introduire des biais. Le pattern a-t-il un sens ? Est-il possible d'établir un lien entre les schémas d'absence et un facteur expérimental ? Si oui, envisagez d'imputer par les niveaux de ce facteur.

Filtrer : supprimer des variables et/ou individus

- Enlever les variables et individus avec trop de données manquantes

- Au-delà d'un certain seuil (20-30%), les méthodes d'imputation ne vont pas bien fonctionner
  - ➢ introduction de biais très fort,
  - ➢ résultats difficilement interprétables (#données « fictives/artificielles » > #données observées)

  - ➢ Attention si DM MNAR ➛ introduction de biais

Bien connaître ses données

- Se poser aussi les bonnes questions

Exemple : spectrométrie de masse, choix arbitraire : suppression de toutes les variables avec > 20% de données manquantes

➢ Configuration suivante pour une variable (2 groupes d'intérêt, 12 individus)

	In1d_g1	Ind2_g1	Ind3_g1	Ind4_g1	Ind5_g1	Ind6_g1	In1d_g2	Ind2_g2	Ind3_g3	Ind4_g4	Ind5_g2	Ind6_g2
$omic_i$	110,038	NA	251,457	161,326	244,02	144,133	NA	NA	34,958	NA	NA	NA

➢ Ici sur l'ensemble des individus : 50% de NA ➛ si on applique le filtre, on supprime la variable

➢ Filtre par groupe : g1: 16,7% ; g2: 83,3% de NA ➛ potentiel biomarqueur

➢ **Question** : comment imputer les données manquantes?

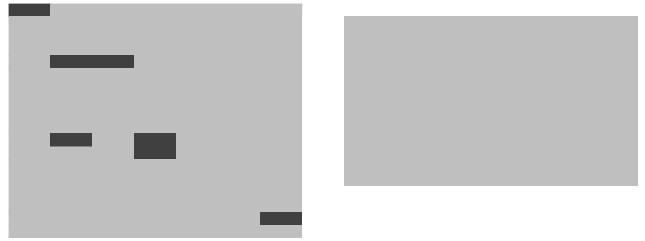
- Groupe 2 : supposer que la variable n'est pas présente pour le groupe 2 (ou en très faible quantité, i.e. < seuil de détection) ➛ remplacer le NA par 0 ou une toute petite valeur
- Groupe 1 : Utiliser une autre approche pour imputer la DM (exemple : knn)
- Exemple de packages R : ***DAPAR, imp4p***

Choisir sa stratégie

- **Méthodes fondées uniquement sur les données observées** :
  - Analyse des cas complets, analyse des cas disponibles, etc.
- **Imputation simple**
  - Complétion stationnaire
  - Basée sur des similarités entre individus
  - Approches par prédiction
  - Approches fondées sur les analyses factorielles
  - Etc.
- **Méthodes pour les approches paramétriques d'inférence statistiques**
  - Modélisation jointe : approches EM ou bayésiennes
- **Méthodes pour estimer la variabilité et fiabilité de l'imputation**
  - Évaluer la fiabilité de l'imputation (erreur dans l'estimation de la valeur imputée) : outil de diagnostique
  - Estimation de la variabilité liée au processus d'imputation
    - Imputation multiple, estimation de l'incertitude dans les modèles EM
- **Données MNAR** : méthodes spécifiques
  - modèles de sélection, modèles par mélange de profils, modèles à paramètres partagés

Méthodes fondées uniquement sur les données observées

- Analyse des cas complets *Little & Rubin (2002)*
  - Ne considérer que les individus pour lesquels toutes les données sont observées
  - **Souvent par défaut dans les logiciels (dont R)**
  - *Graham (2009)* ➛ déconseillé si DM > 5%
  - ☺ facile à mettre en œuvre, pas de modèle à spécifier
  - ☹ Valable pour MCAR, perte possible de beaucoup d'individus



FACTOMINER R

```
Warning message:
In PCA(orange) :
  Missing values are imputed by the mean of the variable: you should use the imputePCA function of the missMDA package
```

- Pondération par probabilité inverse (IPW) *Seaman & White (2011)*
  - Réduire les biais d'estimation ➛ repondérer les cas complets disponibles
  - R package : **ipw**
  - ☺ facile à mettre en œuvre, pas de modèle à spécifier
  - ☹ Valable pour MCAR et MAR, souvent - efficace imputation multiple

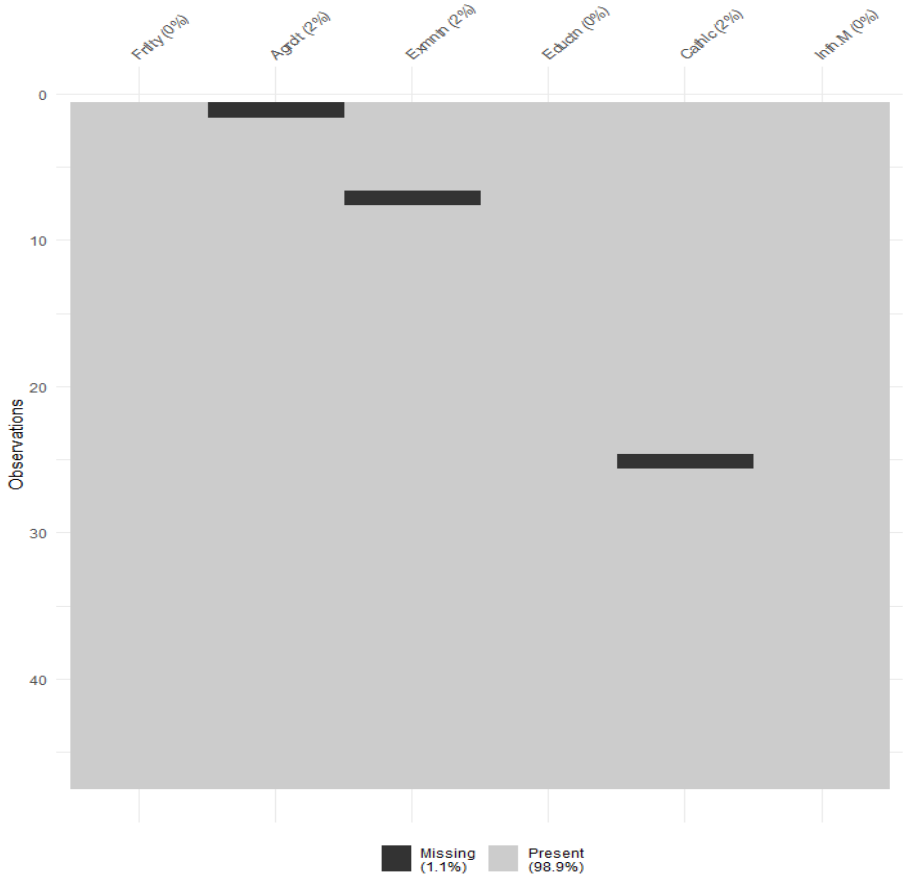
Méthodes fondées uniquement sur les données observées

- Analyse des cas disponibles *Allison (2001), Pigott (2001)*
- estimer ≠ aspects du problème avec ≠ sous-échantillons
  - Matrice de corrélation, covariance
    - Fonction R cor, cov
  - Apprentissage des arbres

```
cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

- ☺ : facile à mettre en œuvre, pas de modèle d'imputation correct à spécifier, permet de prendre en compte + d'individus
- ☹ : principalement valable MCAR
  - favorise/défavorise de manière artificielle certaines variables selon %DM, notamment si variable pertinente pour partitionner l'échantillon (erreurs en apprentissage majorées)
  - interprétation difficile car sur des sous-échantillons (pas forcément représentatif), erreur standard des estimateurs habituels difficiles à obtenir

Illustration, corrélation




```
> cor(swM, use="everything")
           Frtlty Agrclt Exmntn      Eductn Cathlc      Infn.M
Frtlty  1.0000000     NA      NA -0.66378886      NA  0.41655603
Agrclt         NA      1      NA          NA      NA          NA
Exmntn         NA     NA       1          NA      NA          NA
Eductn -0.6637889     NA      NA  1.00000000      NA -0.09932185
Cathlc         NA     NA      NA          NA       1          NA
Infn.M  0.4165560     NA      NA -0.09932185      NA  1.00000000
```

```
> cor(swM, use = "all.obs")
Error in cor(swM, use = "all.obs") :
  observations manquantes dans cov / cor
```

cas complets

```
> cor(swM, use="complete.obs")
           Frtlty      Agrclt      Exmntn      Eductn      Cathlc      Infn.M
Frtlty  1.0000000  0.37821953 -0.6548306 -0.67421581  0.4772298  0.38781500
Agrclt  0.3782195  1.00000000 -0.7127078 -0.64337782  0.4014837 -0.07168223
Exmntn -0.6548306 -0.71270778  1.0000000  0.69776906 -0.6079436 -0.10710047
Eductn -0.6742158 -0.64337782  0.6977691  1.00000000 -0.1701445 -0.08343279
Cathlc  0.4772298  0.40148365 -0.6079436 -0.17014449  1.0000000  0.17221594
Infn.M  0.3878150 -0.07168223 -0.1071005 -0.08343279  0.1722159  1.00000000
```

cas disponibles

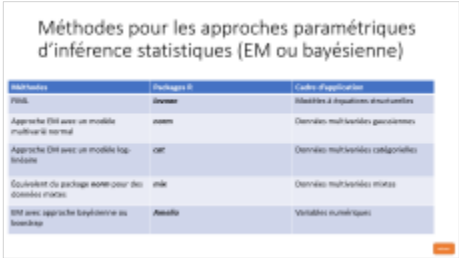
```
> cor(swM, use="pairwise.complete.obs")
           Frtlty      Agrclt      Exmntn      Eductn      Cathlc      Infn.M
Frtlty  1.0000000  0.39202893 -0.6531492 -0.66378886  0.4723129  0.41655603
Agrclt  0.3920289  1.00000000 -0.7150561 -0.65221506  0.4152007 -0.03648427
Exmntn -0.6531492 -0.71505612  1.0000000  0.69921153 -0.6003402 -0.11433546
Eductn -0.6637889 -0.65221506  0.6992115  1.00000000 -0.1791334 -0.09932185
Cathlc  0.4723129  0.41520069 -0.6003402 -0.17913339  1.0000000  0.18503786
Infn.M  0.4165560 -0.03648427 -0.1143355 -0.09932185  0.1850379  1.00000000
```

Méthodes pour les approches paramétriques d'inférence statistiques (EM ou bayésienne)

Schafer (1997)

- **Objectif** : inférence ➛ approches fondées sur la modélisation paramétrique de la distribution multivariée des données, $f(Y;\theta)$ permettant 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 DMs soit vérifiée

- **Approche fréquentiste**
  - basée sur des approches EM
  - Maximum de vraisemblance à information incomplète : **FIML** (*Full Information Maximum Likelihood*)
  - Peut servir à l'imputation, θ estimé ➛ échantillonner selon la loi $f(Y;\theta)$ de pour compléter les DMs MAIS pas besoin si objectif est d'estimer θ
- **Approche bayésienne**
  - Loi a priori définie sur θ, $p(\theta)$ ➛ utilisée pour déterminer la loi a posteriori
  - Approche par augmentation de données, itération de 2 étapes
    - Étape I (Imputation) : M tableaux complets générés selon la loi $f(Y_{miss}|Y_{obs};\theta)$ courante
    - Etape P (postérieure) obtention de la loi a posteriori



Imputation simple (1/3) : Complétion stationnaire *Enders, 2010*

Type de variables	Numérique	Catégorielle	longitudinale
Méthode	Moyenne, médiane	Mode	LOCF (dernière valeur connue)

- ☺ simple et rapide
- ☹ même pour un %DM bas ➛ ignore les relations entres variables et entre individus, sous-estime variabilité/corrélation, déforme les distributions, variance doit être ajustée, sensible aux valeurs aberrantes

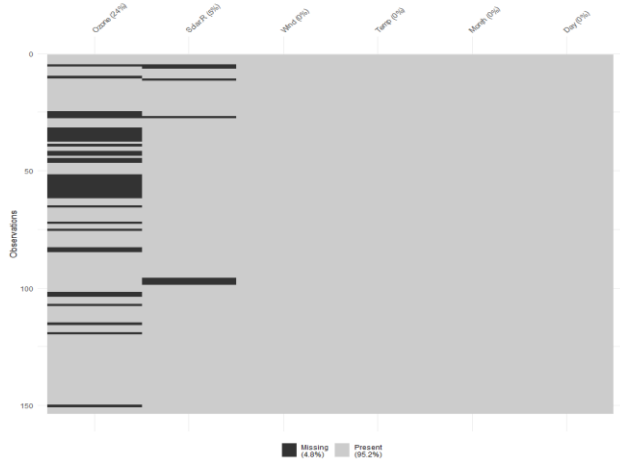
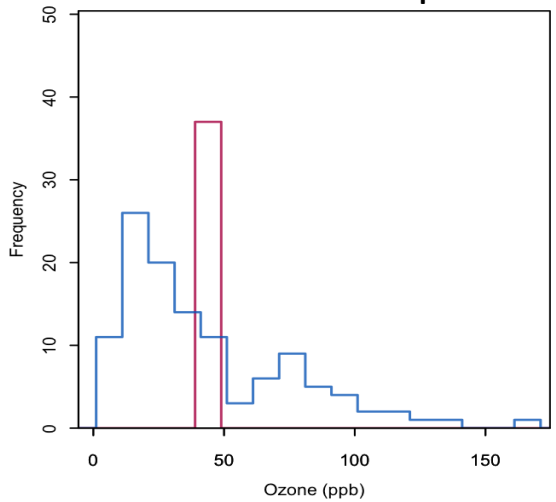
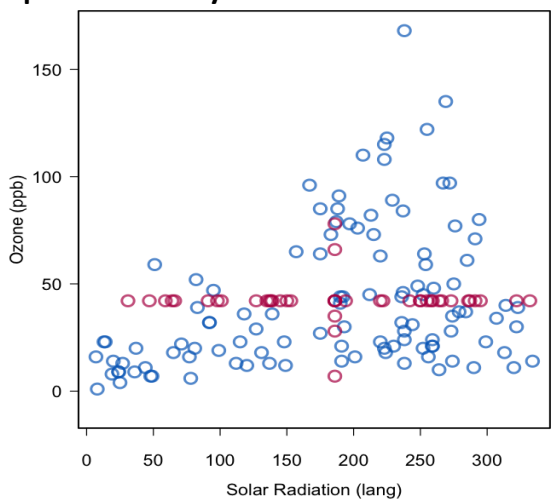


Illustration imputation par la moyenne





Imputation simple (2/3) : similarité entre individus

Utiliser des valeurs observées des individus similaires à l'individu pour lequel une valeur est manquante

kNN	Hot-deck
basé sur une distance Choix de k (*Jönsson & Wohlin 2004*)	fondé sur le concept de donneurs *Andridge & Little (2010)*
☺ facile, basée sur des mesures de similarités, pas d'hypothèse sur la distribution, souple (choix de la distance), non paramétrique	
qd k >1 : améliore erreur et erreur quadratique moyenne de l'estimation de diverses stat.	préservent distribution (MCAR), estimateur sans biais (MAR), données imputées réalistes, si échantillon grand : moins sensible à une mauvaise spécification des hypothèse qui sous-tendent l'imputation
☹ tendance à déformer les corrélations entre variables (imputation individus en entier ➛ besoin d'avoir *n* nombre conséquent d'individus)	
k ↗ tend à déformer la distribution univariées des variables (et donc la variance)	- estimateurs biaisées dans certains cas ➛ pas adaptée à l'estimation des mesures d'associations entre variables, - estimateur moyenne sous estimée (➛ imputation multiple), - résultat dégradés si peu échantillon petit (dépend des donneurs)

Imputation simple (3/3)

- Approches par prédiction
  - Régression locale (LOESS) *Cleveland & Devlin (1988)*
  - Régression linéaires, pénalisées (lasso, ridge, elastic net),etc.
  - Régression non paramétrique : arbre de régression (Breiman et al. 1984), forêt aléatoire (Breiman 2001, Stekhoven & Bühlmann 2012)
  - Mieux adaptées pour la modélisation jointe des variables
  - ☺ : jeu complet, flexibles (large choix d'approche de régression)
  - ☹ : valable MAR, + difficile à mettre en œuvre, bonne spécification de la méthode de régression, requièrent bonne prédictibilité des variables avec DM par les autres variables

- Approches fondées sur les analyses factorielles
  - Imputation basée sur de l'ACP, ACM, AFDM : *nipals* (DM faible, Wold 1966), ACP itérative (ACP-EM, Kiers 1997), ACP itérative régularisée (Verbanck et al 2015), ACP bayésienne (Ilin & Raiko 2010, Verbanck et al 2015)
  - ☺bien adapté analyse exploratoire, variantes adaptées a la grande dimension et au grand volume
  - ☹ valable MAR, cadre théorique restreint aux modèle de génération des données fondés sur les modèles à effets fixes ou mixtes; problème si relation non linéaire entre les variables continues

Prendre en compte l'incertitude liée à l'imputation

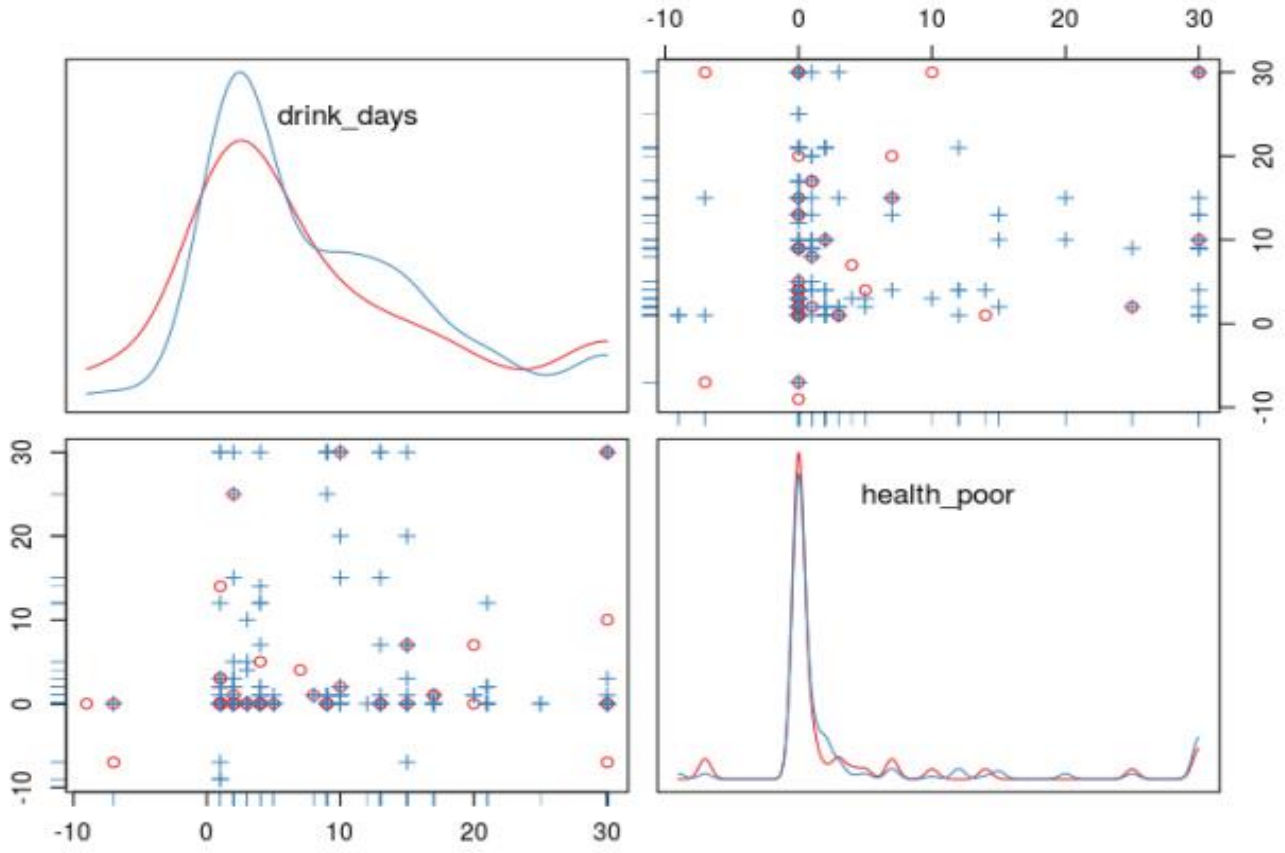
- Pour évaluer la fiabilité des résultats
  - Incertitude sur les données observées (bruit)
  - Incertitude provenant de l'imputation
  - 2 composantes qui sont confondues, erreur globale estimée

- **Outils de diagnostique** (*Abayomi et al 2008, Stuart et al, 2009*)
- **Imputation multiple** *(Rubin 1987, Schafer 1999, Rubin 2012)*
- *SEM*: imputation est prise en charge par une hypothèse paramétrique nécessitant l'estimation d'un paramètre θ (Meng & Rubin 1991)

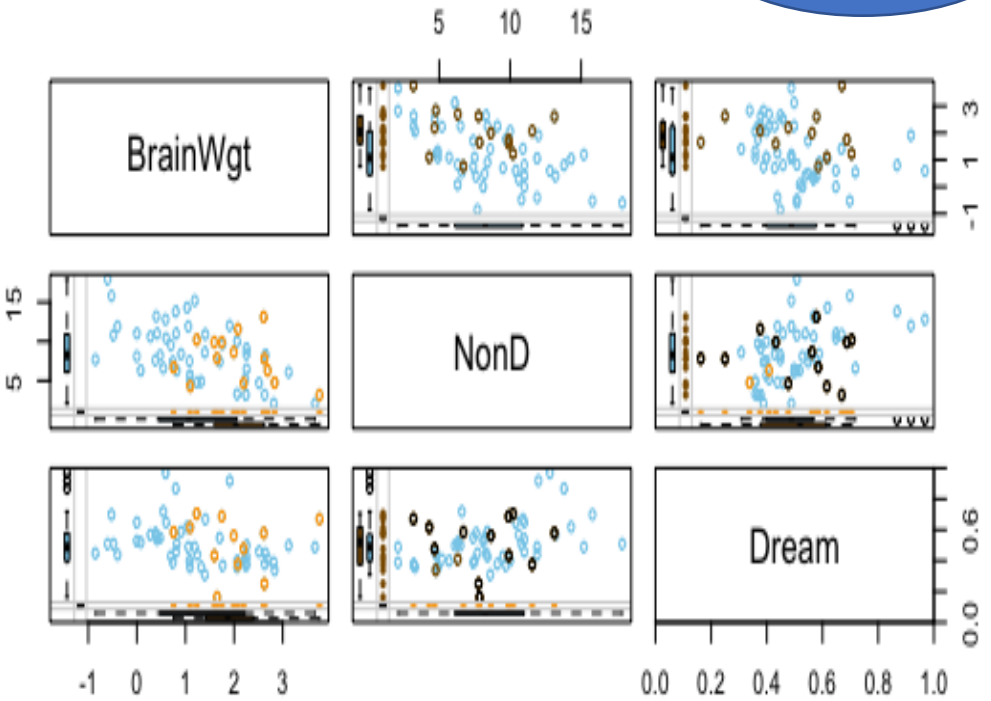
Outils de diagnostiques (1/2)

1) Représentation graphique

- Couleurs ≠ pour données imputées et observées
- **Densité semblable ? Valeur atypique dans les données imputées ? Répartition spécifique des points (valeurs imputées) sur le nuage de points ?**



R package:
VIM



Outils de diagnostiques (2/2)

R package:
mi

2) Comparer densités entre valeurs imputées et observées

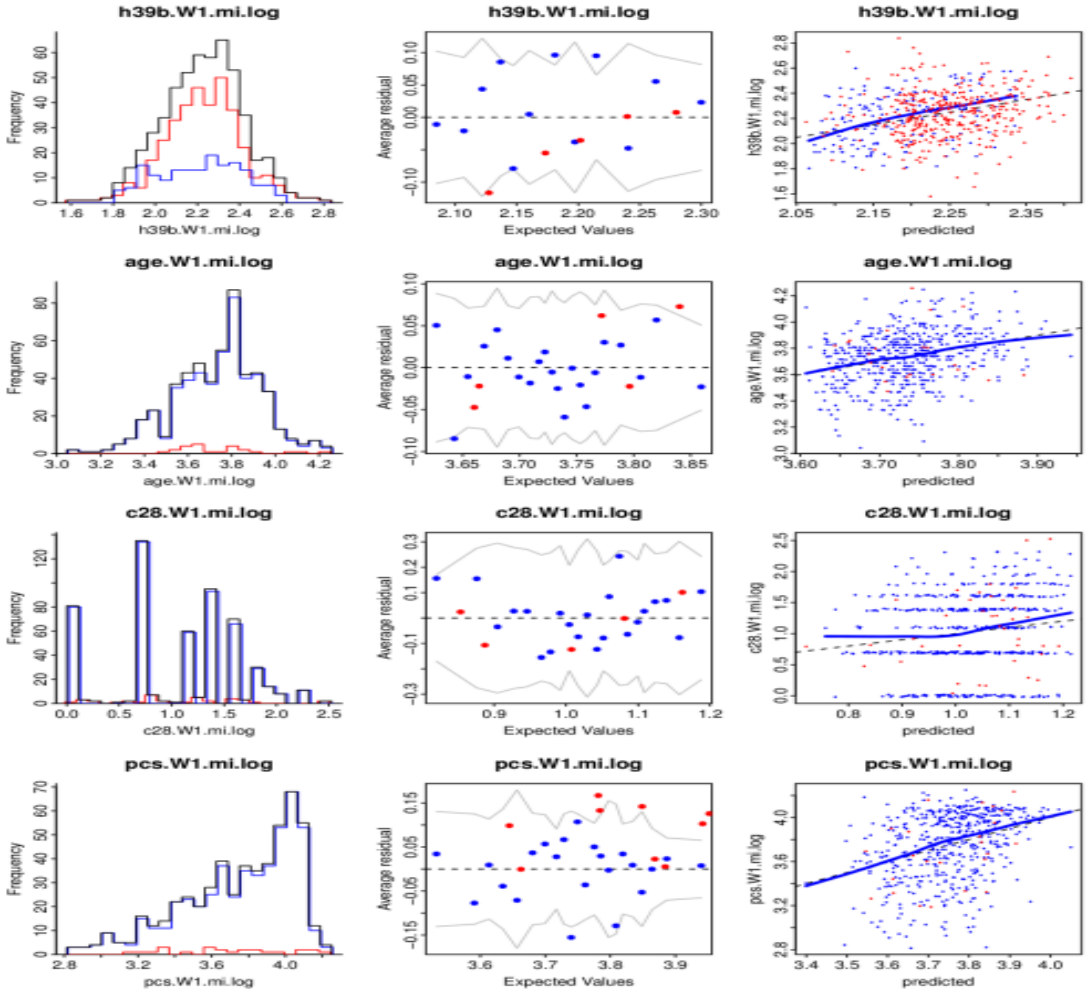
- Test de kolmogorov-Smirnov
- Graphique : histogramme, courbe de densité

3) Pour les imputations générées par des modèles ajustées sur données observées

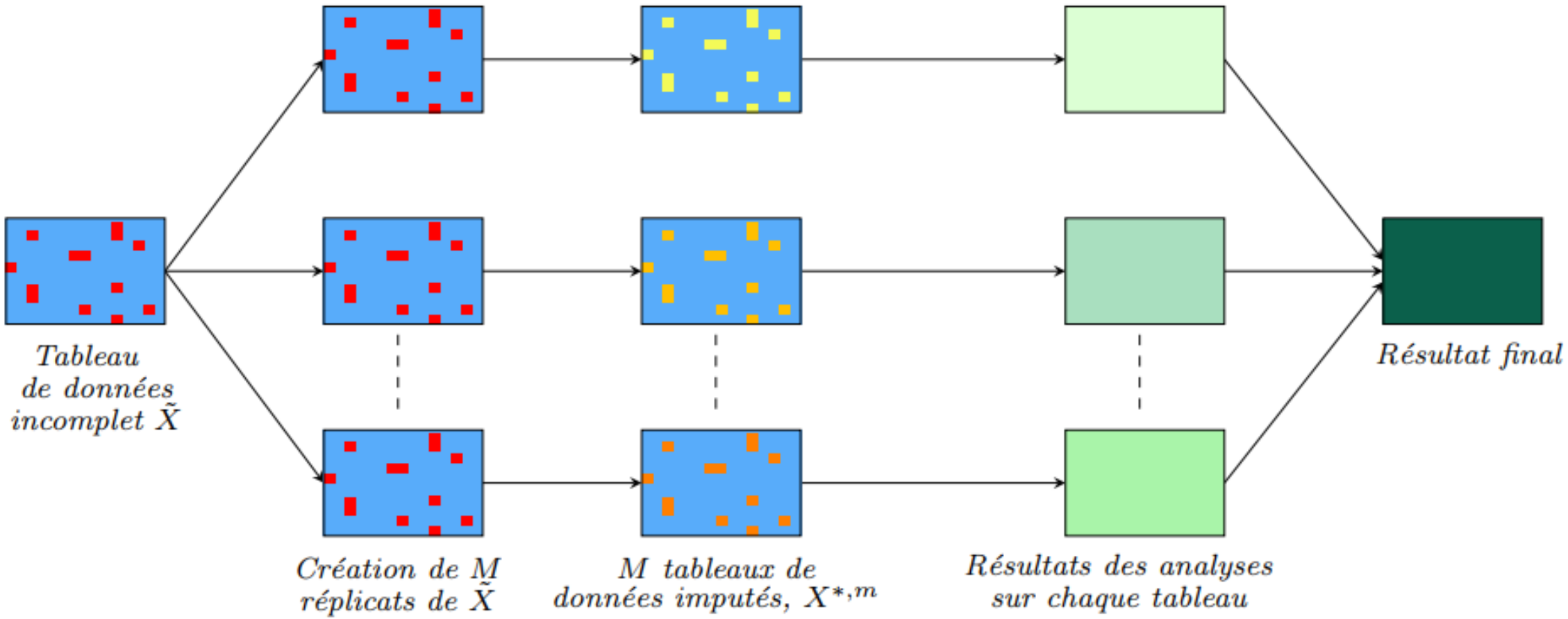
- Graphique résidus, qqplot pour un modèle linéaire
- Se rapproche de la sur-imputation

<u>Conclusion</u> :

- problème imputation ➛ choisir une autre méthode
- sous-groupe de la population avec plus de DMs mis en avant ➛ les étudier

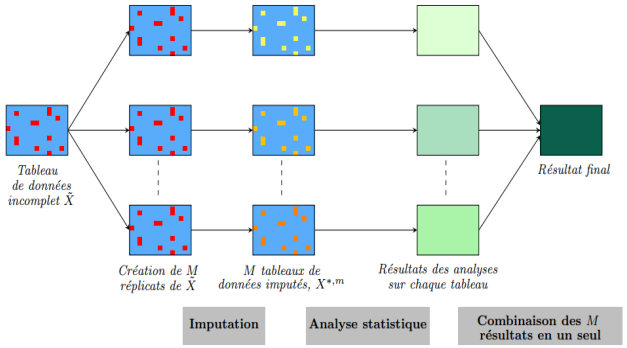


Imputation multiple, schéma



Phase d'imputation

- Différentes méthodes : ré-échantillonnage, type hot-deck, bayésienne, etc.
- M, nombre d'imputations conseillé :
  - *Rubin (1987)* : entre 3-5, 3-10,
    ↳ 20-100 devenu assez commun
  - *White, I.R. et al.(2010)* : M↗ quand taux de DM ↗ + Monte-Carlo, nombre M nécessaire
  - *Von Hipel et al. (2020)* : formule + poussée pour obtenir des estimations de SE reproductibles (SAS macro %mi_combine, Stata, R package ***howManyImputations***)



Combiner les résultats

- Estimation d'un paramètre numérique ➛ estimateur moyen
- Résultats sous forme + complexe, quelques exemples :
  - ***missMDA*** : ellipse de confiance autour de la projection des individus, *Josse et al. 2012*
  - Projection consensuelle : projection la + corrélée aux M projections obtenues lors d'AFM, *Voillet et al. 2016* (missRows)
  - Inférence de réseau : fréquence de prédiction d'une arête, *Imbert et al. 2018* (***RNASeqNet***)
  - Imputation multiple pour faire de l'analyse différentielle (données protéomiques): ***mi4p,*** *Chion et al. 2021*

Quelques approches pour tenter de prendre en compte les données MNAR

- **Problème** : la loi de Y_{miss} n'est pas indépendante de la loi de R
  - Approches habituelles (estimer la loi multivariée Y; $f(Y;\theta)$) ➛ estimateurs ou valeurs imputées biaisés

- Factorisation de la loi jointe
  - **Modèle de sélection** (Heckman 1976, Diggle & Kenward 1994)
    - ☹ : hypothèse paramétrique sur f(R|Y;psi), sensible à une mauvaise spécification de cette loi
  - **Modèle de mélange de profils** (Rubin 1977)
    - ☹ : sensible aux hypothèses de restriction (non vérifiables), compromis à effectuer, loi marginale de Y pas disponible directement
  - ☺ adapté au cas où la non réponse directement liée aux variables observées (ex: question sensible)
  - Si DM attribuable à un processus sous-jacent (ex: progression maladie) : modèles à paramètres partagés

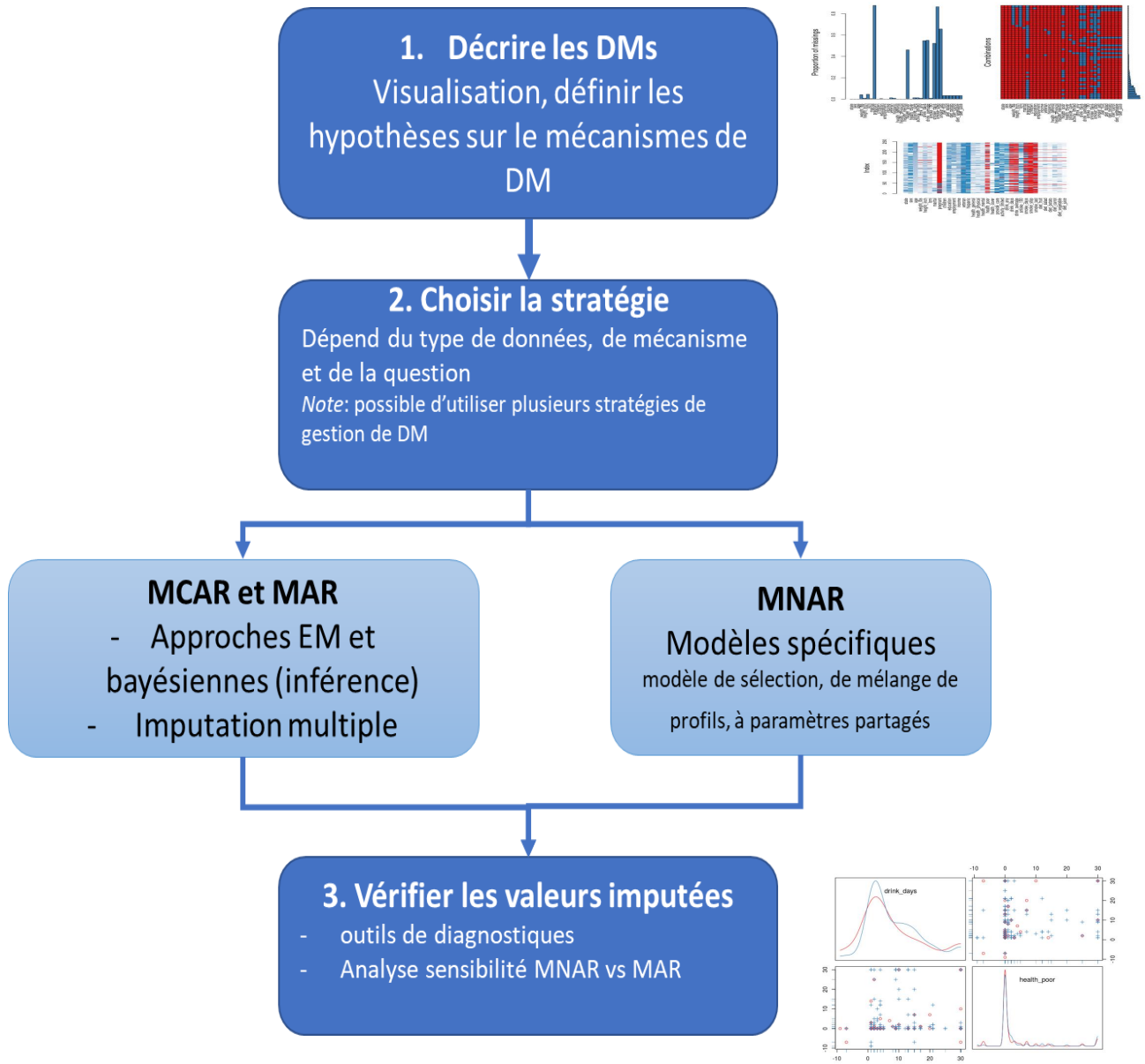
- **Modèles à paramètres partagés** (Little 1995; Hogan & Laird 1997) : estimer les dépendances entre Y et R au moyens de variables aléatoires latentes

- **Analyse de sensibilité** : fondée sur une perturbation des données en direction hypothèse MNAR pour vérifier pertinence du modèle MAR
  - Comparaison de ≠ jeux imputés issus de modèles d'imputation ≠ (R package ***mice***)
  - *VERBEKE, et al (2001)* proposent d'utiliser des modèles de mélanges de profil pour effectuer analyse de sensibilité

Conclusion

- Quelle que soit la stratégie utilisée :
  - Cela ne pourra **jamais** remplacer un jeu de données complet et consistant
  - Toujours une perte d'information, possibilité d'introduction de biais qui ne reflète pas nécessairement les relations régissant les données
  - Bien de tester plusieurs méthodes et voir celle(s) qui fonctionne(nt) ➛ outils de diagnostique

Recommandations générales :



Liens utiles

- https://cran.r-project.org/web/views/MissingData.html
- https://rmisstastic.netlify.app/

- Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes Imbert, A. & Vialaneix, N., Journal de la Société Française de Statistique (2018)

- Quelques livres :
  - https://stefvanbuuren.name/fimd/
  - The missing book Nicholas Tierney & Allison Horst (2022)
  - https://argoshare.is.ed.ac.uk/healthyr_book/identification-of-missing-data.html (chap 11)

Me: I should do multiple imputation

Other me: just impute using the mean and be done with it

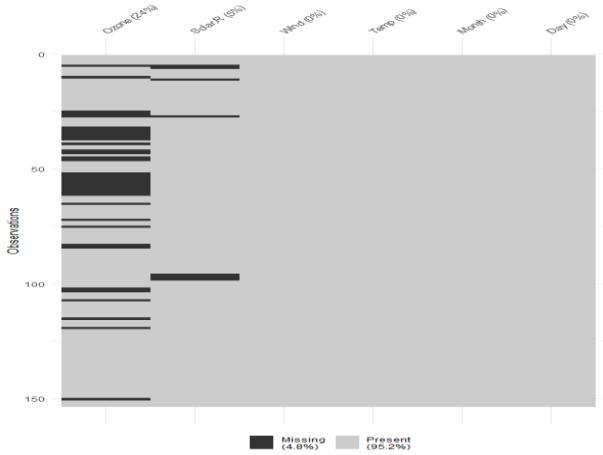
FIN

Source : https://www.brodrigues.co/blog/2019-11-02-mice_exp/

Test sur le «mécanisme» de données manquantes

- MAR vs MNAR : pas de test connu, meilleure option est d'utiliser les connaissances spécifiques du domaine sur les données
  - msImpute : outils de diagnostique MAR/MNAR (MS, niveau peptidique)
- MCAR vs MAR : test de Little MCAR permet d'évaluer les preuves en faveur ou à l'encontre de ces deux paramètres
  - P-value faible suggèrent données MAR, Pvalue élevée suggèrent
  - R package :
    - ***naniar,*** mcar_test(.)
    - ***misty***, na.test(.)

```
> mcar_test(airquality)
# A tibble: 1 x 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1      35.1    14 0.00142                4
```



Enders, 2010:

- n'identifie pas les corrélats potentiels de l'absence de réponse
- Based on multivariate normality
- Étudie différence de moyenne en supposant que le modèle partage une matrice de covariance commune : ne peut pas détecter les écarts MCAR basés sur la covariance provenant d'un modèle MAR ou MNAR (qui peuvent produire des sous groupes de DM avec des moyennes égales)
- Simulation -> test suffers from low statistical power surtout quand ppeu de variables viole l'hypothèse MCAR, relation entre variable et missingness faible ou MNAR
- Peut seulement rejeter MCAR mais pas le prouver
- MNAR ne peut pas être exclu quelque soit le résultat du test (Ho: données MCAR ou MNAR, résultats significatif : MAR ou MNAR

Exemple de pourquoi s'intéresser aux DMs

Cas simple, inférence statistique, estimer l'espérance

- Approche naïve : cas complet
- Objectif : estimer l'espérance de Y_1, $\mu_1 = \boldsymbol{E}(Y_1)$
- Estimateur habituel : $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}$
  - Pas nécessairement observé sur certaines valeurs de Y_1 sont manquantes
  - $\tilde{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n} r_{i1} y_{i1}$ avec $n_1 = \sum_{i=1}^{n} r_{i1}$, nombre de valeurs observées de Y_1
- MCAR, R et Y indépendantes ➛ $\tilde{\mu}_1$ estimateur sans biais de μ_1
- MAR ou MNAR : R et Y pas indépendantes ➛ exemple si Y_1 liée à Y_2
  - $R_1 = \begin{cases} 0 \; si \; Y_2 \leq a \\ 1 \; sinon \end{cases}$ pour $a \in ensemble \; des \; réels$
  - $\boldsymbol{E}(\tilde{\mu}_1) = \boldsymbol{E}(Y_1 \mathbf{1}_{\{Y_2 > a\}})$ ➛ biais dans l'estimation de μ_1

Mécanisme ignorable

- *Rubin 1976* : conditions minimales requises qui permettent d'ignorer le processus de génération des données manquantes dans l'inférence statistique ➛ mécanisme ignorable

1) Données manquantes aléatoirement (MAR + MCAR)

2) les paramètres régissant le mécanisme de génération de DM et des données doivent être « distinguables », i.e. φ = (ψ, θ) où :
  - ψ désigne les paramètres qui régissent la distribution de R et
  - θ sont les paramètres qui régissent celle de Y
- Paramètres distinguables lorsqu'ils vivent dans de espaces en produits cartésiens
  - ↳ MAR : possible de factoriser la densité des données observées :

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

  - Donc : vraisemblance des données observées proportionnelle à la vraisemblance ignorant le mécanisme à l'origine des données manquantes $L(\theta|Y_{obs})$

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

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 Faible taux de DM (<5%) Mécanisme MCAR	Facile à mettre en œuvre; Pas besoin de modèle d'imputation	Valable ssi MCAR; Perte d'information; perte de précision
Analyse des cas disponibles	***Regtools*** Option dans certaines fonctions Cor(…, ***method=« pairwise »)***	Numériques et catégorielles	Facile à mettre en œuvre; Pas besoin de modèle d'imputation; + d'individus pris en compte	Ok si MCAR (Dé) favorise artificiellement des variables Stat. Sur des population différente -> difficilement comparable
Pondération par probabilité inverse (IPW)	***ipw***	Numériques et catégorielles		

retour

Méthodes pour les approches paramétriques d'inférence statistiques (EM ou bayésienne)

Méthodes	Packages R	Cadre d'application
FIML	***lavaan***	Modèles à équations structurelles
Approche EM avec un modèle multivarié normal	***norm***	Données multivariées gaussiennes
Approche EM avec un modèle log-linéaire	***cat***	Données multivariées catégorielles
Équivalent du package ***norm*** pour des données mixtes	***mix***	Données multivariées mixtes
EM avec approche bayésienne ou boostrap	***Amelia***	Variables numériques

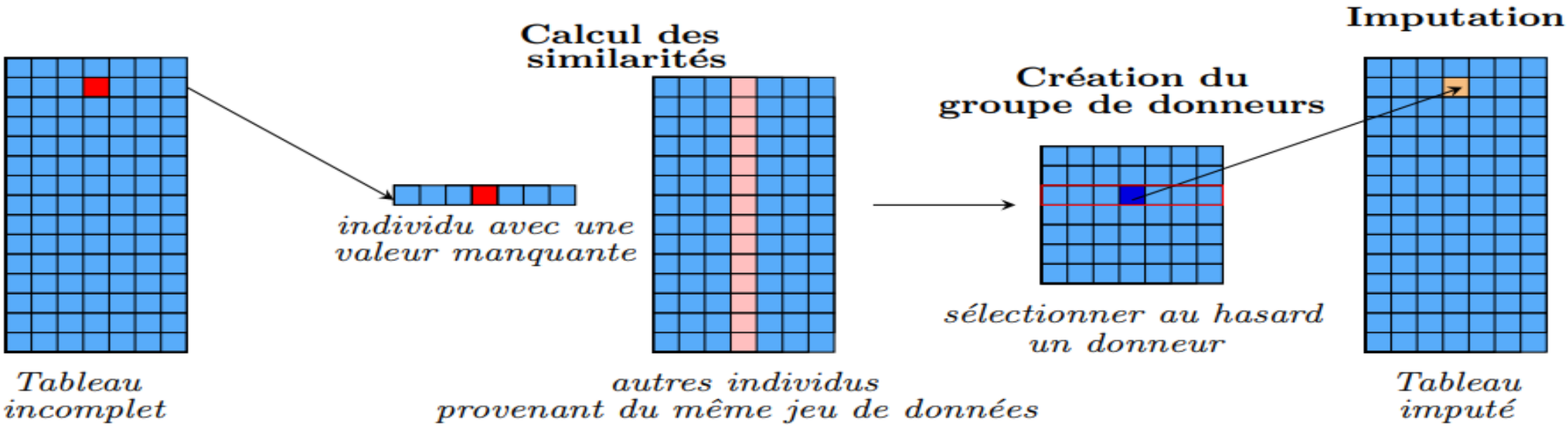
retour

Imputation simple

Méthodes	Packages R	Cadre d'application
Moyenne, médiane	***Hmisc, simputation***	Numériques
Mode	***Hmisc***	Catégorielles
LOCF	***zoo***	Longitudinales
K plus proches voisins	***Impute, VIM, yaImpute***	Numérique et/ou catégorielle (cf distance)
Hot-deck	***hot.deck, simputation,VIM, HotDeckImputation***	Mixte
Régression	***simputation, VIM***	Numériques
Régression LOESS	***locfit***	Numériques
Régression stochastique	***mice (m=1)***	Numériques
Arbres et forêts aléatoires	***missForest***	Mixtes
NIPALS	***ade4, pcaMethods, mixOmics***	Numériques
Analyses factorielles	***missMDA***	Catégorielles et/ou numériques
ACP probabiliste, ACP bayésienne	***pcaMethods***	
Interpolation, ajustement d'une courbe de lissage, estimation de régression longitudinales	***forecast, imputeTS, spaceTime, timeSeries, xts, zoo***	Séries temporelles

retour

Hot-deck, principe



- Hot-deck métrique ou plus proche voisin
  - distance
- Hot-deck avec score d'affinité
- Hot-deck hiérarchisé
  - Ordre naturel entre les variables
- Hot-deck aléatoire avec ou sans remise
  - Nécessite individus avec un profil homogène
  - Hot-deck par classe (classification)
- Hot-deck séquentiel

retour

missForest, Stekhoven&Bühlmann (2012)

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_{imp}^{old} = matrice précédemment imputée
   (b) **Pour** s dans k **faire**
      i. Ajuster $y_{obs}^{(s)} \sim x_{obs}^{(s)}$ par forêt aléatoire
      ii. Prédire $y_{mis}^{(s)}$ avec les régresseurs $x_{mis}^{(s)}$
      iii. Y_{imp}^{new} est la nouvelle matrice complétée par les valeurs prédites $y_{mis}^{(s)}$
   (c) mettre à jour le critère γ

Source : wikistat

- Imputation naïve: moyenne, médiane, mode
- Critère arrêt : différence entre la matrice imputée nouvelle et celle de l'itération précédente ↗

Approches d ’évaluation de la variabilité due aux NA ou à l’imputation

Méthodes	Packages R	Cadre d’application
fondée sur une approche par modélisation jointe (EM et bayésiennes)	***Amelia***	Numériques
Version multiple hot-deck	***Hot.deck***	mixtes
Approches de modélisation jointe (EM et bayésiennes) multi-niveau	***Jomo, pan***	Mixtes
Équations chaînées	***mi, mice*** mifa (use mice + combined covariance matrices)	Mixtes
Analyses factorielles	***missMDA***	Numérique, mixtes
Combinaison générique	***mitools***	Mixtes stratifiés en classe
Imputation multiple par AFM	***missRows***	Imputation individus en entier
Imputation multiple : wrapper de ***mice***, de ***imp4p*** et *impute.knn*	***mi4p***	Imputation multiple pour de l’analyse différentielle protéomique
Outils de diagnostique		
Calculs d’erreur	***Amelia, missMDA, yaImpute***	Mixtes
Graphique	***mi, VIM***	mixtes

missCompare: pipeline d’imputation pour guider les utilisateurs à gérer leurs DMs et sélectionner la méthode la + adaptée à leurs données via des outils de diagnostique

retour

MICE : Multivariate Imputation by Chained Equations

- Introduction de l'aléa réalisé via l'approche d'équations chaînées, approché bayésienne fondées sur la méthode FCS, Stef Van Buuren & Oudshoorn (1999)
- Bâtir un modèle conditionnel pour chaque variable de façon séquentielle jusqu'à convergence

Table 6.1: Built-in univariate imputation techniques in the `mice` package. Methods marked by * are default for scale type.

Method	Description	Scale Type
`pmm`	Predictive mean matching	Any*
`midastouch`	Weighted predictive mean matching	Any
`sample`	Random sample from observed values	Any
`cart`	Classification and regression trees	Any
`rf`	Random forest imputation	Any
`mean`	Unconditional mean imputation	Numeric
`norm`	Bayesian linear regression	Numeric
`norm.boot`	Normal imputation with bootstrap	Numeric
`norm.nob`	Normal imputation ignoring model error	Numeric
`norm.predict`	Normal imputation, predicted values	Numeric
`quadratic`	Imputation of quadratic terms	Numeric
`ri`	Random indicator for nonignorable data	Numeric
`logreg`	Logistic regression	Binary*
`logreg.boot`	Logistic regression with bootstrap	Binary
`polr`	Proportional odds model	Ordinal*
`polyreg`	Polytomous logistic regression	Nominal*
`lda`	Discriminant analysis	Nominal

FCS : *Fully Conditionnal Specification* (spécification de lois conditionnelles)

missMDA, Josse&Husson (2016)

Imputation multiple: MIPCA

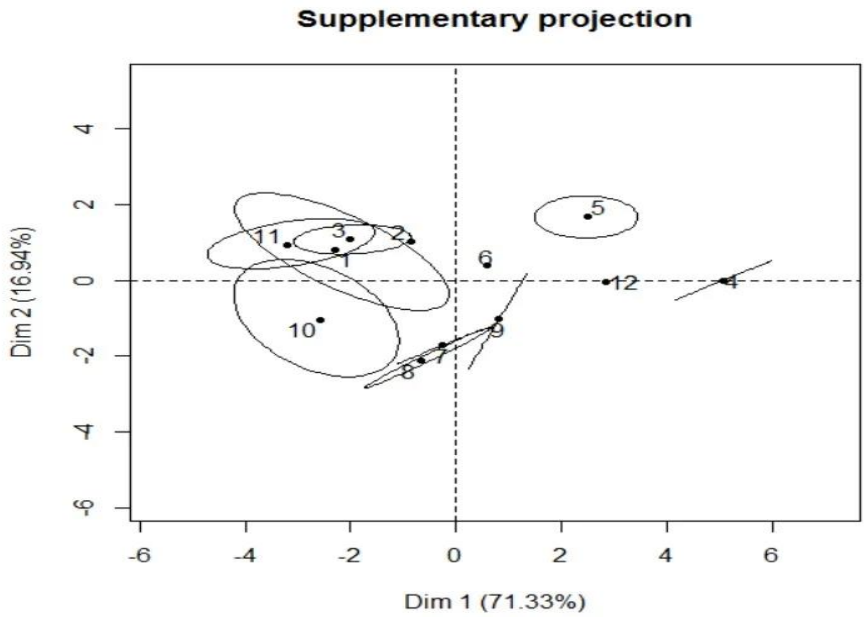
- Générer M jeux de données dans lesquels seules les valeurs imputées ≠ -> approche type « bootstrap sur les résidus »
- ACP itérative sur les M jeux de données
- Représentation graphique globale

ACP itérative

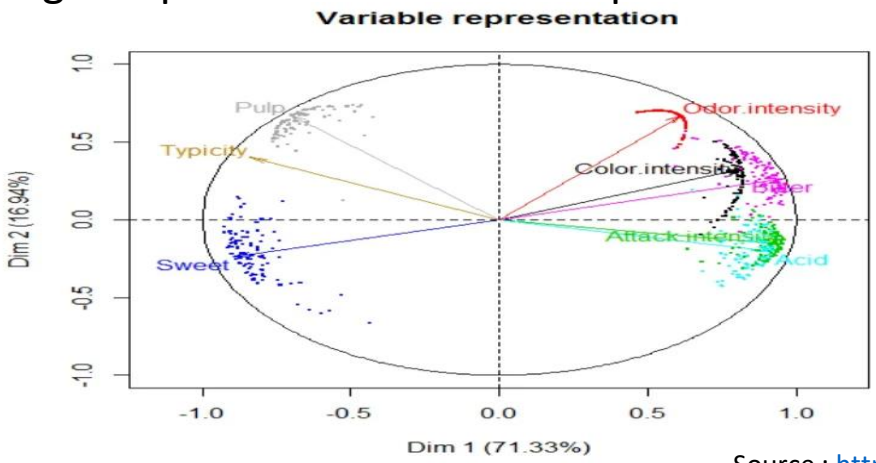
1. Initialisation : remplacer la DM par la moyenne
2. ACP réalisée, chercher à trouver la droite qui minimise les distances entre les points et leur projection perpendiculaire sur cette droite
3. Prédire la DM -> nouvelle valeur imputée
4. Nouvelle ACP
5. Répétition de ces étapes jusqu'à convergence

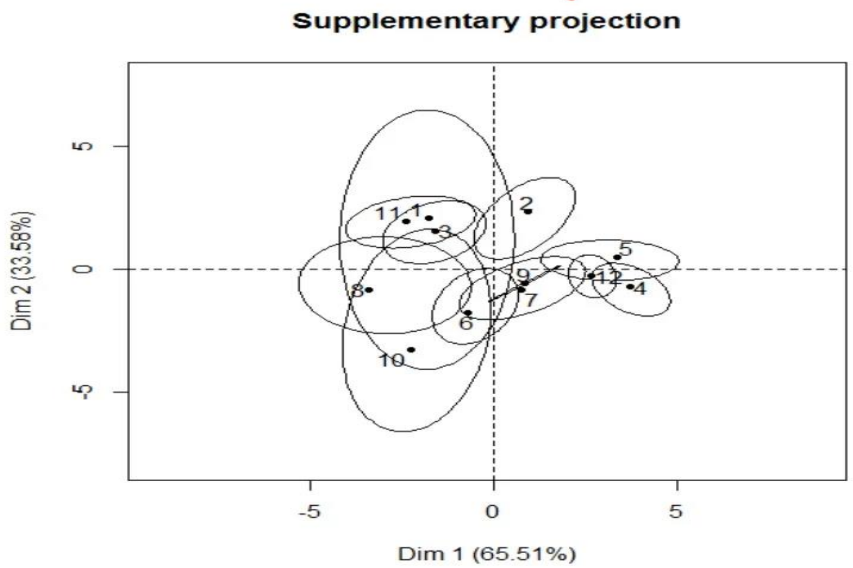
missMDA, Comment lire les graphiques

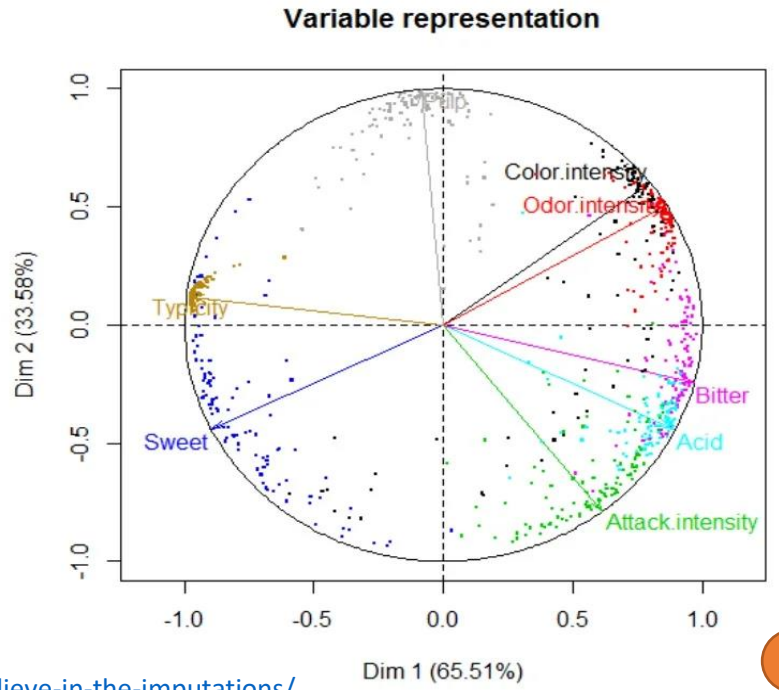
Ellipse: témoin de la variabilité inter-imputation



Nuage de points: incertitude des prédictions



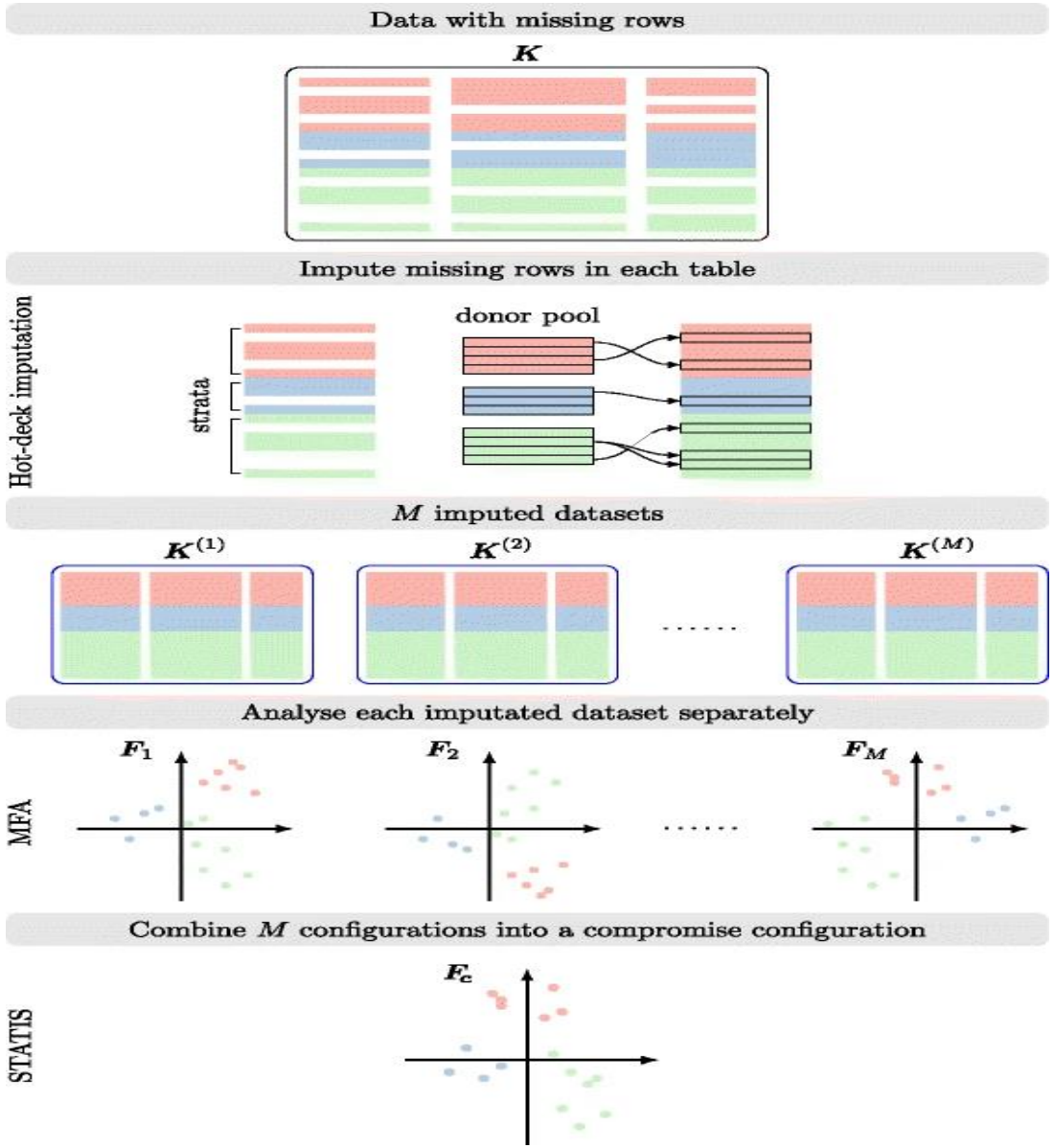




Source : https://francoishusson.wordpress.com/2017/08/05/can-we-believe-in-the-imputations/

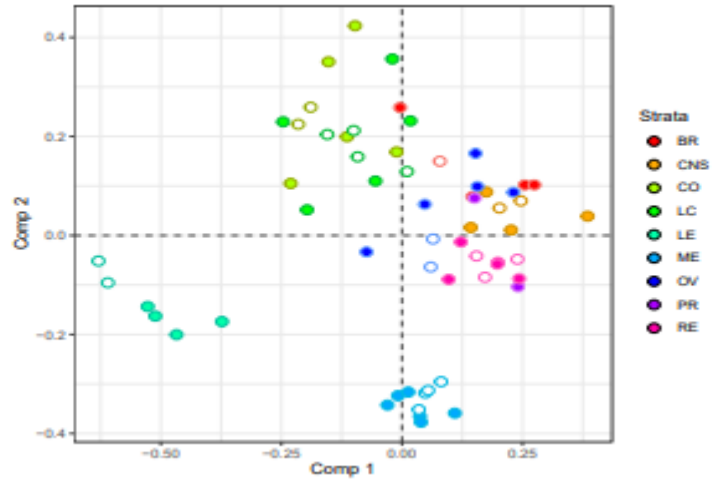
retour

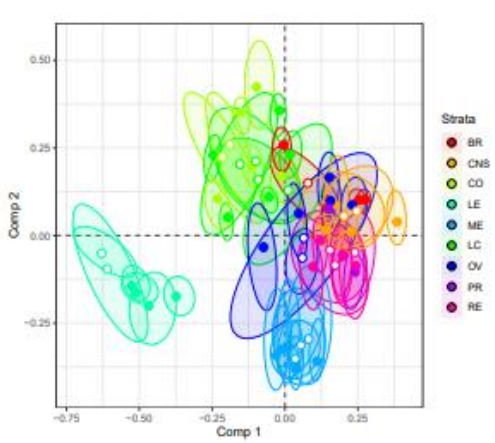
missRows, MIMFA

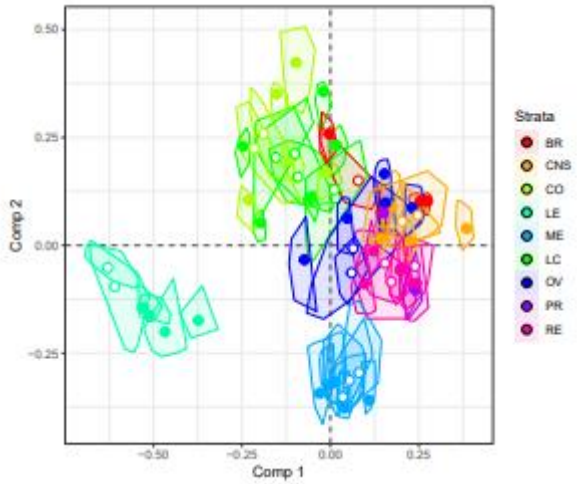


[Voillet et al (2016)]

Vignette , missRows



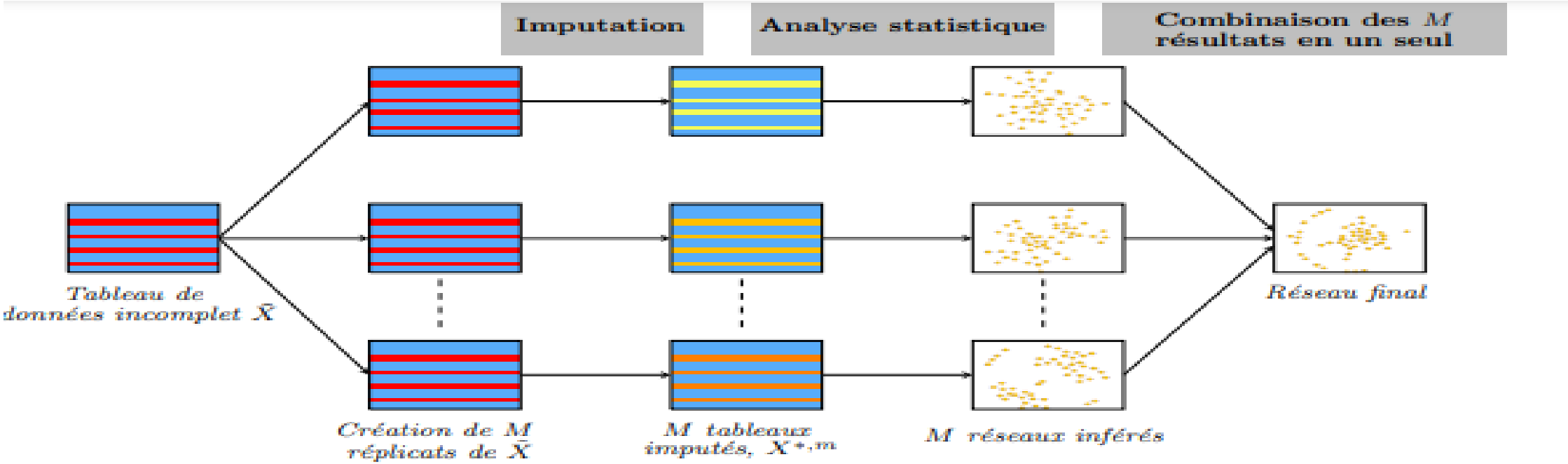


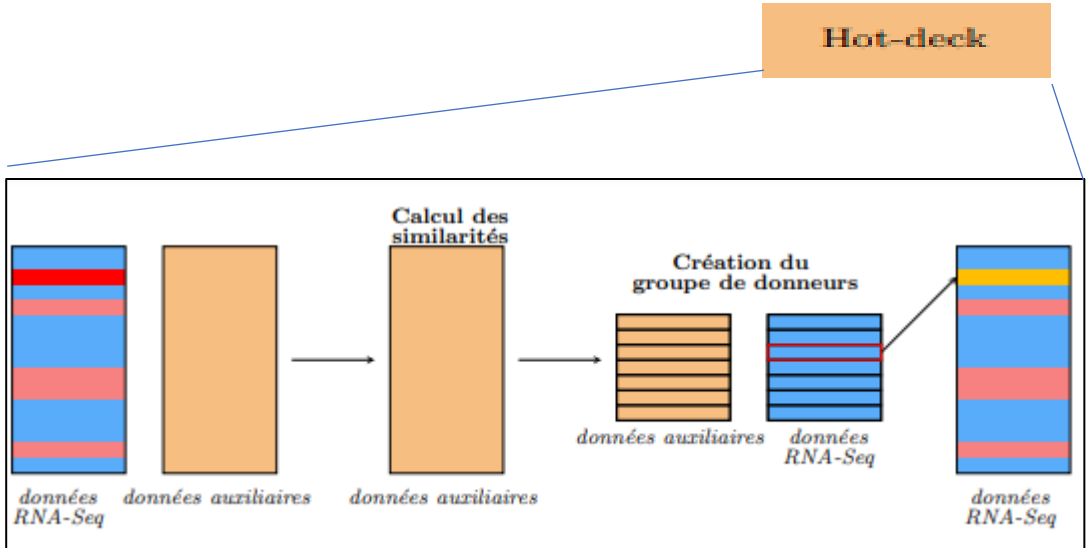


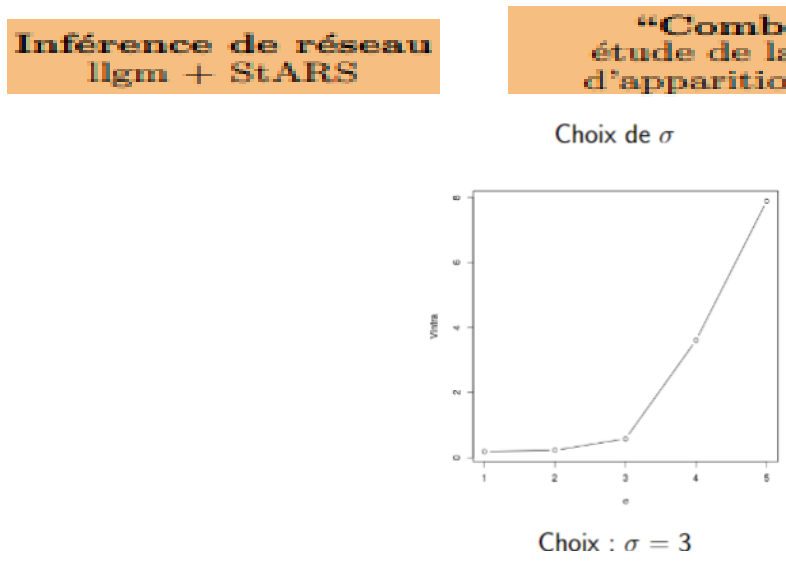
+ ellipse (convex hull) est large, + incertitude sur la position exacte de l'individu ➛ attention à l'interprétation

retour

Inférence de réseau : RNAseqNet







Estimation de l'incertitude dans les modèles EM

- SEM (*Supplemental EM*)
- Exprimer l'erreur quadratique moyenne de θ en fonction de 2 quantités estimables
  - Erreur quadratique moyenne de θ sur les données observées
  - Taux de convergence de l'algo EM
- Obtenue directement avec FIML, nécessite une étape supplémentaire (SEM)
- Principal limite : hypothèse paramétriques + adaptation de l'approche pour chaque cadre d'hypothèse
  - Imputation multiple (IM) + simple pour évaluer incertitude liée à l'imputation
    - ☹ IM : parfois difficile de trouver règle de combinaison résultats satisfaisant les propriétés préconisées par Little et Rubin (2002)
  - Cadre inférence statistique : bcp papiers ➛ supériorité en terme de puissance stat. de l'approche EM, FIML, (vs IM)

Quelques packages R pour les données MNAR

Méthodes	Packages R	Cadre d'application
Équations chaînées	***miceMNAR***	Binaire ou continu
Modèle de sélection, profil mélangé, *hurdle model*	***missingHE***	Variables continues, longitudinale
Modèles de sélection de copules	***GJRM***	Résultat non gaussien
MAR : MLE (EM algo), SVD, knn; ***MNAR***: mindet, minprob, QRILC	***imputeLCMD***	Données censurées à gauche
KNC (k-nearest class means), knn, ***MAR*** : low-rank approximation, ***MNAR*** : random draw from Gaussian parameterised around lowest observed value in the sample	***msImpute***	Imputation pour l'intensité de peptides (spectrométrie de masse, lebel-free) ➛ variables continues
Imputation multiple; ***MAR*** : MLE, PCA, random forest, SLSA; ***MNAR*** : *igcda* (under a Gaussian Complete Data Assumption), *pa* (présence dans 1 condition/absence dans l'autre) : petites valeurs	***imp4p*** (certaines méthodes aussi dans ***DAPAR***)	Imputation peptides et protéines ➛ variables continues

retour