STA211 : gestion des données manquantes

V. Audigier, N. Niang

04 février, 2025

1 Introduction

Les données manquantes constituent un problème quasiment incontournable dans l'application des techniques de data-mining. En effet, l'augmentation de la taille des données fait mécaniquement augmenter la probabilité de rencontrer des données manquantes. Les raisons à ces données manquantes sont nombreuses. Par exemple, il peut s'agir de données qui n'ont pas été saisies informatiquement suite à un oubli, ou encore des questionnaires qui n'ont été que partiellement remplis parce que les individus interrogés ont oublié de répondre, ou ne savaient pas répondre, ou n'ont pas voulu répondre car certaines questions étaient embarrassantes (questions sur la consommation de drogues, questions portant sur les revenus, etc). Les données manquantes peuvent aussi apparaître lors de la fusion de jeux de données : en concaténant des jeux de données ont fait apparaître des blocs verticaux de données manquantes car les variables diffèrent potentiellement d'un jeu à l'autre. Or, les méthodes de data-mining ne sont généralement pas conçues pour s'appliquer sur des jeux incomplets. Il convient donc de gérer ces données manquantes pour pouvoir y appliquer une analyse. Deux approches sont notamment envisageables : effectuer un pré-traitement pour se ramener à des données complètes, ou utiliser des méthodes avancées qui peuvent s'appliquer sur les données incomplètes.

L'objet de ce document est de présenter les manières classiques de gérer les données manquantes en insistant tout particulièrement sur les techniques dites d'*imputation* consistant à remplacer les données manquantes par des valeurs plausibles. Le propos sera à nouveau illustré à partir du jeu de données German Credit sur lequel des données seront supprimées de façon artificielle.

2 Les données manquantes

Face à ces données incomplètes, l'attitude généralement adoptée par les utilisateurs est de limiter leur analyse aux individus complets, méthode appelée *suppression par liste* ou *étude des cas complets* (Little and Rubin (2002)). Cette méthode est également utilisée par défaut dans la plupart des logiciels. Or, cette façon de procéder n'est pas satisfaisante pour au moins deux raisons. La première est que les individus complets ne constituent pas nécessairement un échantillon représentatif du jeu de données. Ceci implique que l'inférence menée via la méthode du cas complet est généralement biaisée. La seconde raison est que le nombre d'individus complets tend à être rapidement petit dès lors que le nombre de variables est grand. En effet, dans la mesure où chaque variable est sujet à être incomplète, la probabilité qu'un individu soit complet est faible. Par exemple, supposons que les données manquantes soient disposées complètement au hasard et que 5% des valeurs soient manquantes pour chaque variable, alors un jeu composé de 50 variables contient en moyenne 8% d'individus complets. Ainsi, la méthode du cas complet n'est généralement pas raisonnable en data-mining et les données manquantes doivent faire l'objet d'une attention particulière.

Les données (incomplètes) auxquelles on s'intéresse ici peuvent être définies comme des données ''fictives'' pour lesquelles tous les individus seraient renseignés, mais dont seule une partie est visible. Par exemple, une donnée *possession d'un véhicule* non-renseignée pour un individu rentre dans ce cadre : la donnée fictive peut être *oui* (ou *non*) et cette donnée n'est pas visible pour l'individu. En revanche, si un individu ne possède pas de véhicule, alors la donnée relative à la marque de sa voiture ne peut pas être renseignée. Dans ce cas, il n'y a pas de donnée ``fictive'' pour laquelle cette donnée serait renseignée. La gestion de ce type de données manquantes peut se faire en construisant de nouvelles variables issues de la fusion de variables d'origine. Par exemple, plutôt que de considérer les deux variables *possession d'un véhicule* et *marque de la voiture* (pour laquelle des données ne sont pas renseignées), on considérera la variable fusionnée dont les modalités seront les différentes marques et la modalité *pas de véhicule*. Toutes les données de cette nouvelle variable seront ainsi observées et il n'y a donc plus de problème de données manquantes. La gestion de ce type de données incomplètes ne sera pas plus développée dans ce document, nous nous intéresserons ici uniquement aux données manquantes pour lesquelles des données fictives pourraient exister.

2.1 Vocabulaire et notations

La gestion des données manquantes nécessite l'introduction d'un vocabulaire et de notations particulières. On note $\mathbf{X}_{n\times p}$ la matrice des données (cf Table 2.1). On appelle *données complètes*, les données fictives, en partie inconnues. La matrice des données complètes sera notée $\mathbf{X}_{n\times p}^{full}$. On appelle *dispositif des données manquantes*, la matrice $\mathbf{R}_{n\times p}$ composée de 1 et de 0 indiquant respectivement qu'une donnée est manquante ou présente (cf Table 2.2). On note x_i^{obs} le profil observé de l'individu i et x_i^{miss} le profil manquant.

Table 2.1: Extrait du jeu de données German Credit sur lesquelles des données ont été enlevées

	Données brutes				
	Status	Duration	History	Purpose	Credit.Amount
1	lt.0	6	critical	Radio.Television	1169
2	0.to.200	48	paidDuly	Radio.Television	5951
3	Statusnone	12		Education	2096
4	lt.0	42	paidDuly	Furniture.Equipment	7882

	Données brutes				
	Status	Duration	History	Purpose	Credit.Amount
5	lt.0	24	delay	NewCar	4870
6	Statusnone	36	paidDuly	Education	9055
7	Statusnone	24	paidDuly	Furniture.Equipment	
8	0.to.200	36	paidDuly	UsedCar	6948
9	Statusnone	12	paidDuly	Radio.Television	3059
10	0.to.200	30	critical	NewCar	5234

Table 2.2: Extrait des données complètes (à gauche) et extrait du dispositif des données manquantes (à droite)

	Données complètes					Dispositif de données manquantes				
	Status	Duration	History	Purpose	Credit.Amount	Status	Duration	History	Purpose	Credit.Amount
1	lt.0	6	critical	Radio.Television	1169	0	0	0	0	0
2	0.to.200	48	paidDuly	Radio.Television	5951	0	0	0	0	0
3	Statusnone	12	critical	Education	2096	0	0	1	0	0
4	lt.0	42	paidDuly	Furniture.Equipment	7882	0	0	0	0	0
5	lt.0	24	delay	NewCar	4870	0	0	0	0	0
6	Statusnone	36	paidDuly	Education	9055	0	0	0	0	0
7	Statusnone	24	paidDuly	Furniture.Equipment	2835	0	0	0	0	1
8	0.to.200	36	paidDuly	UsedCar	6948	0	0	0	0	0
9	Statusnone	12	paidDuly	Radio.Television	3059	0	0	0	0	0
10	0.to.200	30	critical	NewCar	5234	0	0	0	0	0

$\mathbf{X}_{n\times p}$, $\mathbf{R}_{n\times p}$, x_i^{obs} et x_i^{miss} peuvent être vus comme les réalisations de variables aléatoires. On note $X = (X_1, \ldots, X_p)$ et $R = (R_1, \ldots, R_p)$ les variables aléatoires associées à $\mathbf{X}_{n\times p}$ et $\mathbf{R}_{n\times p}$. R est appelé *mécanisme des données manquantes*. Les variables aléatoires associées aux profils observés ou manquants sont notées respectivement X^{obs} et X^{miss} de sorte que $X = \left(X^{obs}, X^{miss}\right)$.

Les méthodes utilisées pour gérer les données manquantes reposent en particulier sur le lien entre R et X.

2.2 Taxonomie des mécanismes

Les mécanismes à l'origine des données manquantes peuvent être classés en trois groupes : les données générées complètement au hasard dites *MCAR* pour *missing completely at random*, les données générées au hasard dites *MAR* pour *missing at random*, et les données non générées au hasard, dites *MNAR* pour *missing not at random* (Rubin (1976), Little (1995)). Les méthodes employées pour gérer les données manquantes sont conditionnées par le type de mécanisme qui affecte le jeu de données.

2.2.1 Mécanisme MCAR

On appelle *données manquantes générées complètement au hasard* des données manquantes dont la probabilité d'occurrence est sans lien avec les données complètes. Sous l'hypothèse MCAR, le mécanisme R est donc indépendant de X^{obs} et X^{miss}. Ce type de mécanisme est typiquement rencontré dans les enquêtes où des individus sont amenés à remplir un questionnaire. En effet, il se peut que certains individus aient oublié de répondre à des questions ou que certaines réponses n'aient pas été saisies manuellement au moment de la numérisation de questionnaires papier. Cette hypothèse sur le mécanisme, même si elle est légitime dans certains cas, reste assez forte.

Un mécanisme MCAR permet de considérer les individus complets comme un sous-échantillon représentatif des individus du jeu de données. Ainsi, les données manquantes complètement au hasard permettent d'appliquer les méthodes usuelles sur le jeu de données restreint à ses individus complètement observés sans engendrer de biais. Toutefois, cela conduit à réduire la taille de l'échantillon considéré et donc à construire des estimateurs avec des variances plus grandes.

2.2.2 Mécanisme MAR

De façon intuitive, les données générées au hasard correspondent à un dispositif des données manquantes non causé par les données non-observées, mais pouvant être dû à la partie observée. Un exemple de données MAR est le cas d'une enquête de satisfaction auprès d'actifs sans emploi vis-à-vis du service offert par l'Agence pour l'emploi : une première série de questions est posée en Janvier aux bénéficiaires. On suppose que toutes les personnes ont répondu à ces questions. Les personnes ayant répondu *non* à la question *êtes-vous globalement satisfait du service proposé ?* sont alors soumises une nouvelle fois à cette série de questions en Février, les autres personnes ne sont pas réinterrogées. Le jeu de données est constitué des réponses aux questions posées en Janvier et Février. L'apparition de données manquantes sur les données de Février dépend de la réponse en Janvier à la question *êtes vous globalement satisfait du service proposé*. Le mécanisme est donc MAR. Sous ce mécanisme, les individus complets ne sont plus un sous-échantillon représentatif de la population et l'inférence sur ces individus peut être biaisée.

L'hypothèse MAR généralise l'hypothèse MCAR, mais est moins restrictive. Elle peut parfois être évidente, comme dans l'exemple précédent, quand on connaît le mécanisme générant les données manquantes, mais en général les données manquantes ne dépendent pas de l'expérimentateur et on ne sait pas si cette hypothèse est vérifiée. De plus, il est impossible de vérifier cette hypothèse [p.9]Fitzmaurice et al. (2014). Ainsi, à défaut de pouvoir la vérifier, il est recommandé d'inclure des *variables auxiliaires* dans le jeu de données. On entend par là des variables (avec peu ou pas de données manquantes) qui ne sont pas d'un intérêt scientifique, mais qui permettent d'expliquer la présence de données manquantes et rendre ainsi l'hypothèse MAR valide. Par exemple, si on souhaite évaluer le lien entre le nombre d'agents dans l'Agence pour l'emploi en Février et la satisfaction vis-à-vis du service proposé par l'agence ce même mois, on ne se limitera pas à imputer le jeu constitué des deux variables *êtes vous globalement satisfait du service proposé* et *nombre d'agents*, mais on inclura la variable *êtes-vous globalement satisfait du service proposé* pour Janvier. Cette dernière servira uniquement à imputer les données, mais ne sera pas analysée par la suite.

2.2.3 Mécanisme MNAR

Par opposition au mécanisme MAR, un mécanisme est dit MNAR si la probabilité d'apparition de données manquantes est en partie causée par les données non-observées X^{miss}. Ces mécanismes sont fréquents dans le cadre d'enquêtes sur des sujets sensibles comme les revenus, la consommation d'alcool, l'usage de drogues, etc. Par exemple, un individu fortuné aura plutôt tendance à ne pas répondre à une question portant sur ses revenus. La probabilité d'apparition d'une donnée manquante est ici liée directement à la partie non-observée des données. Dans ce cas la modélisation est plus complexe car il faut non seulement spécifier une distribution pour les données complètes (comme dans le cadre 'usuel' sans données manquantes) mais aussi une distribution pour le mécanisme, et les distributions conditionnelles de ces deux

distributions (i.e. spécifier le lien entre les données complètes et le mécanisme). En effet, il serait illusoire d'espérer inférer (sans biais) sur le revenu médian d'une population en ne disposant que d'un échantillon des revenus des personnes les moins fortunées. Pour mener une inférence sans biais, on a besoin de connaître, en plus des réponses des individus les moins fortunés, la probabilité qu'un individu donne son revenu en fonction de ce même revenu, afin de tenir compte de la représentativité de l'échantillon observé.

Bien que ce mécanisme soit le plus général, les travaux pour gérer les données manquantes dans ce cadre sont assez spécifiques à leur application [p.5]Allison (2002). Par la suite nous présentons quelques méthodes permettant de gérer les données manquantes de type MAR uniquement.

2.2.4 Illustration

Nous illustrons ces 3 mécanismes en Figure 2.1 en considérant les variables *Duration* et *Credit Amount* et en ajoutant des données manquantes sur la première uniquement. Le mécanisme MCAR choisi est tel que les données sont manquantes indépendamment des données complètes ; le mécanisme MAR est tel que la probabilité d'observer une donnée manquante sur un individu est plus élevée quand la variable *Credit Amount* (sans données manquantes) prend de grandes valeurs ; le mécanisme *MNAR* est tel que la probabilité d'observer une donnée manquante sur un individu augmente avec la variable *Duration* elle même (qui est incomplète).

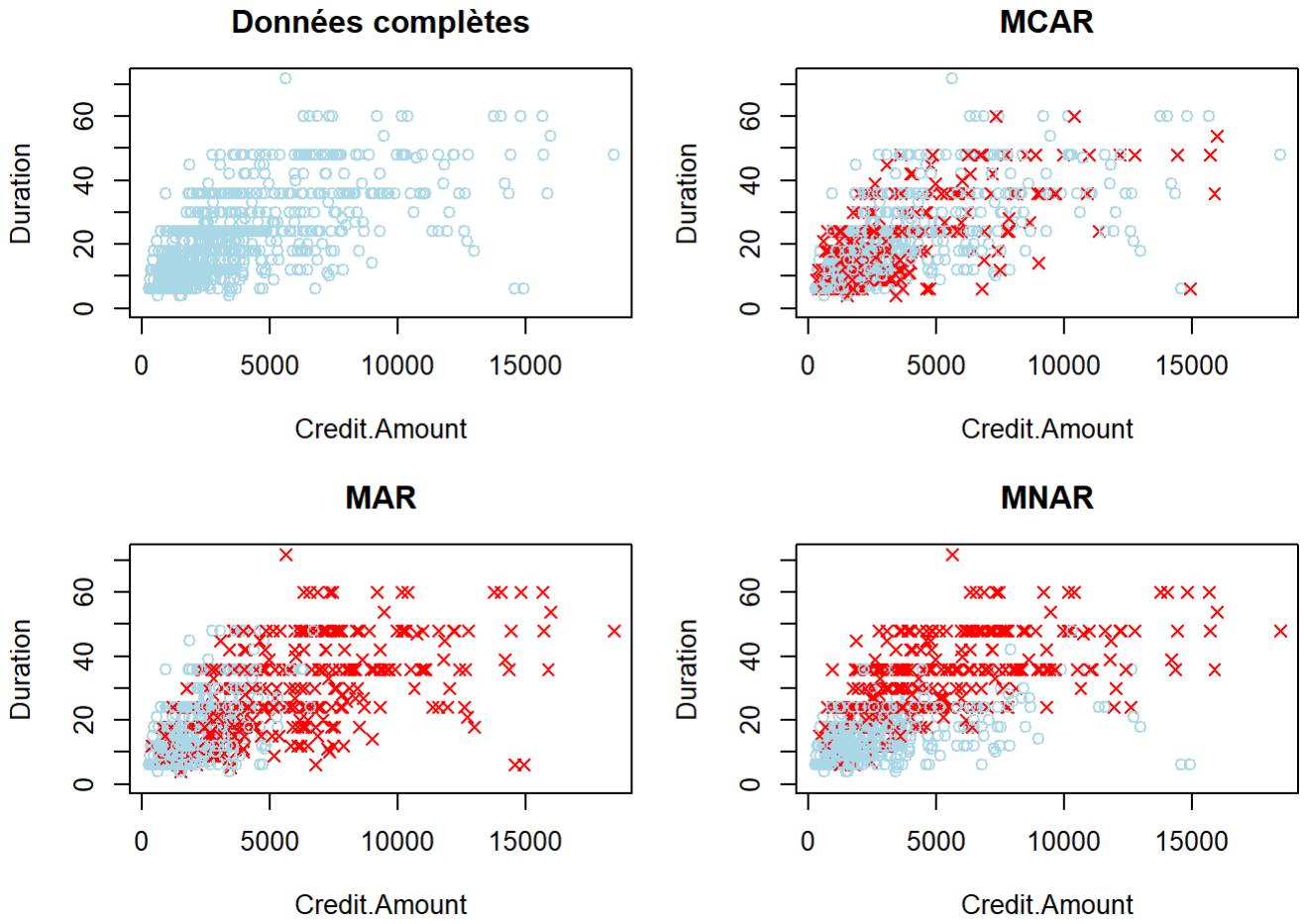


Figure 2.1: Exemple de données manquantes de type MCAR, MAR et MNAR à partir des variables Duration et Credit.Amount. Les points bleus correspondent aux individus complets, et les points rouges aux individus pour lesquels la variable Duration est manquante.

2.3 Identification des mécanismes

Etant donné que la méthode qui sera utilisée pour gérer les données manquantes nécessite de faire une hypothèse sur le mécanisme des données manquantes, il est nécessaire d'avoir une bonne connaissance de celui-ci. Pour cela, on utilisera généralement les méthodes d'analyse exploratoire univariées, bivariées et

multivariées.

Ces analyses exploratoires peuvent tout d'abord être appliquées au dispositif des données manquantes $\mathbf{R}$. L'analyse univariée permettra alors d'identifier la proportion de données manquantes pour chaque variable. L'analyse bivariée elle permettra d'identifier si des données manquantes apparaissent de façon simultanée sur deux variables. Enfin, une analyse multivariée permettra d'identifier des variables qui ont tendance à être manquantes simultanément et à identifier des groupes d'individus dont le profil de non-réponse est similaire. Pour cela on pourra utiliser une ACM, complétée par une classification (non-supervisée). Il conviendra alors de s'interroger sur les raisons de liaisons entre absence de données, et sur l'origine de groupes d'individus aux profils de non-réponse similaires. Suite à cette analyse, certaines données pourront éventuellement être retrouvées en allant interroger les responsables des données. Il est en effet très improbable que le sexe d'un individu ne soit pas renseigné, ce type d'information (généralement importante) doit pouvoir se retrouver facilement. De même, certaines variables sont recueillies de façon systématique car définie dans un protocole d'étude, leur absence dans le jeu de données est simplement liée à un problème de saisie. Des variables très incomplètes seront elles amenées à être écartées de l'analyse. De même, pour les individus très peu renseignés, il sera alors nécessaire de veiller à la représentativité des individus conservés par rapport à l'étude menée.

On peut également appliquer ces analyses exploratoires sur les données brutes en considérant le caractère manquant comme une modalité particulière (dans le cas de données manquantes sur une variable quantitative, on effectuera alors un découpage en classes). Des analyses bivariées ou multivariées permettront de faire le lien entre les données observées sur certaines variables et la présence des données manquantes sur d'autres.

Néanmoins, à elles seules, ces analyses ne permettront pas de trancher entre les différents mécanismes. Pour l'illustrer, on reprend les données de la section 2.2.4 et on représente en Figure 2.2 la distribution de la variable *Credit Amount* en fonction du dispositif des données manquantes sur la variable *Duration*.

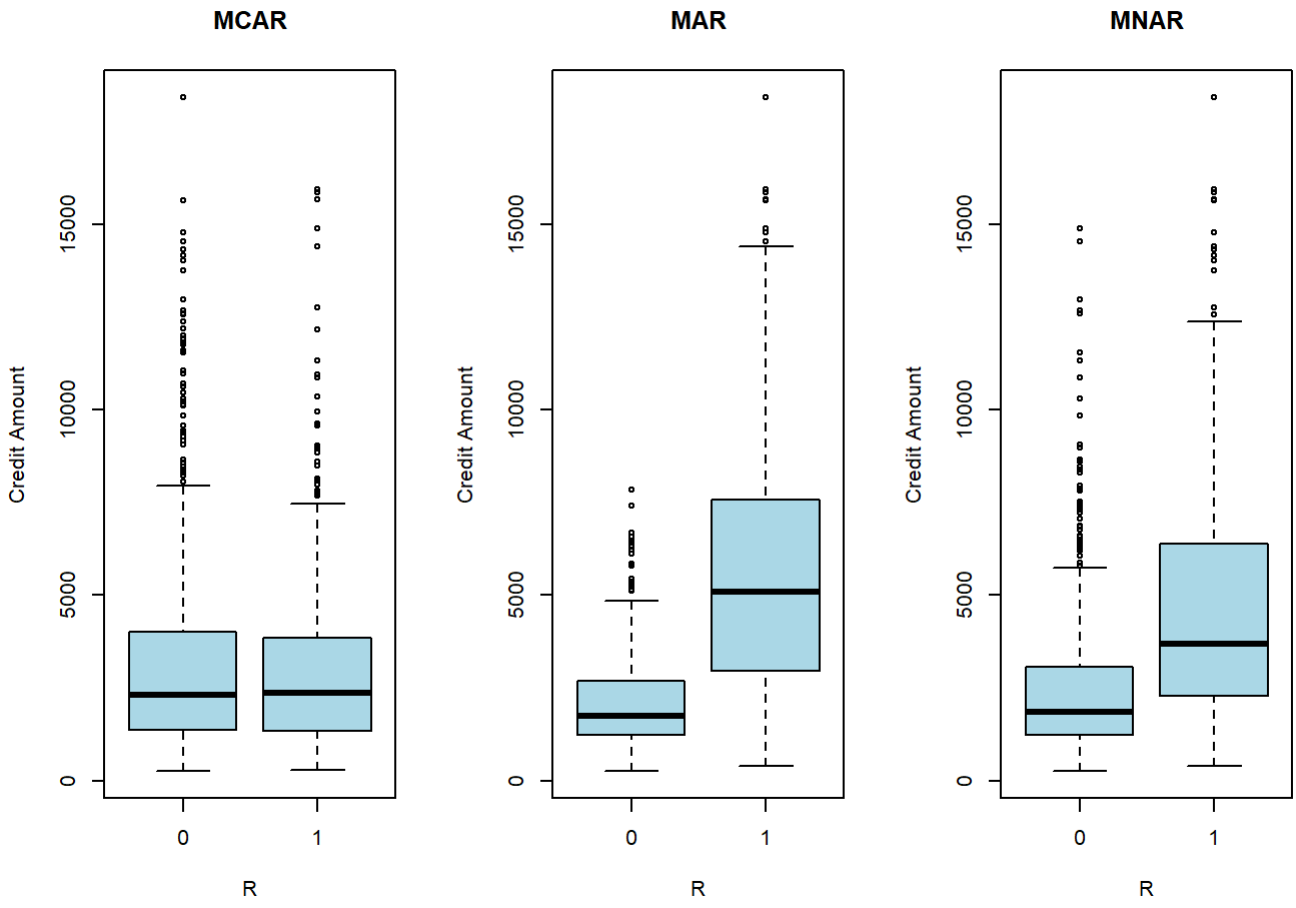


Figure 2.2: Influence du mécanisme sur la distribution des données : variables Duration et Credit Amount

On constate que pour le mécanisme MCAR, la distribution de la variable *Credit Amount* est la même pour les individus observés sur la variable *Duration* que pour les individus incomplets. Dans le cas MAR et le cas MNAR ces distributions sont clairement différentes, ce qui met en évidence un lien entre la variable *Credit Amount* et le mécanisme R (et amènerait donc à se poser la question de ce lien). Pourtant, dans le cas MNAR, il n'y a aucun lien de cause à effet entre la variable *Credit Amount* et le mécanisme. Par ailleurs, il n'est pas possible d'identifier la cause de ce dispositif car pour cela il faudrait disposer des données complètes pour la variable *Duration*.

3 Traitement des données manquantes par imputation

La façon la plus populaire de traiter les données manquantes est de procéder par imputation, i.e. de remplacer les données non-observées par des valeurs plausibles de façon à se ramener à un jeu de données dépourvu de données manquantes.

3.1 Imputation simple

L'imputation simple consiste à remplacer chaque donnée manquante par une unique valeur, celle-ci pouvant néanmoins varier d'un individu à l'autre ou d'une variable à l'autre.

3.1.1 Typologies des méthodes

Les méthodes d'imputation les plus basiques imputent chaque variable indépendamment les unes des autres par la moyenne, la médiane ou en tirant au hasard des données parmi celles observées. Ces approches ne sont généralement pas satisfaisantes notamment parce qu'elles "détruisent" les relations entre variables. Il sera généralement préférable d'utiliser des méthodes d'imputation préservant mieux ces relations comme l'imputation par régression. Il s'agit d'ajuster un modèle de régression à partir des données observées puis de prédire les données non-observées à partir de ce modèle. On ajoutera de préférence une perturbation aléatoire sur la prédiction de façon à mieux respecter la distribution des données (on parle d'*imputation stochastique*). La Figure 3.1 illustre l'utilisation de quelques unes de ces méthodes sur la variable *Duration* dans le cas du mécanisme MAR précédent.

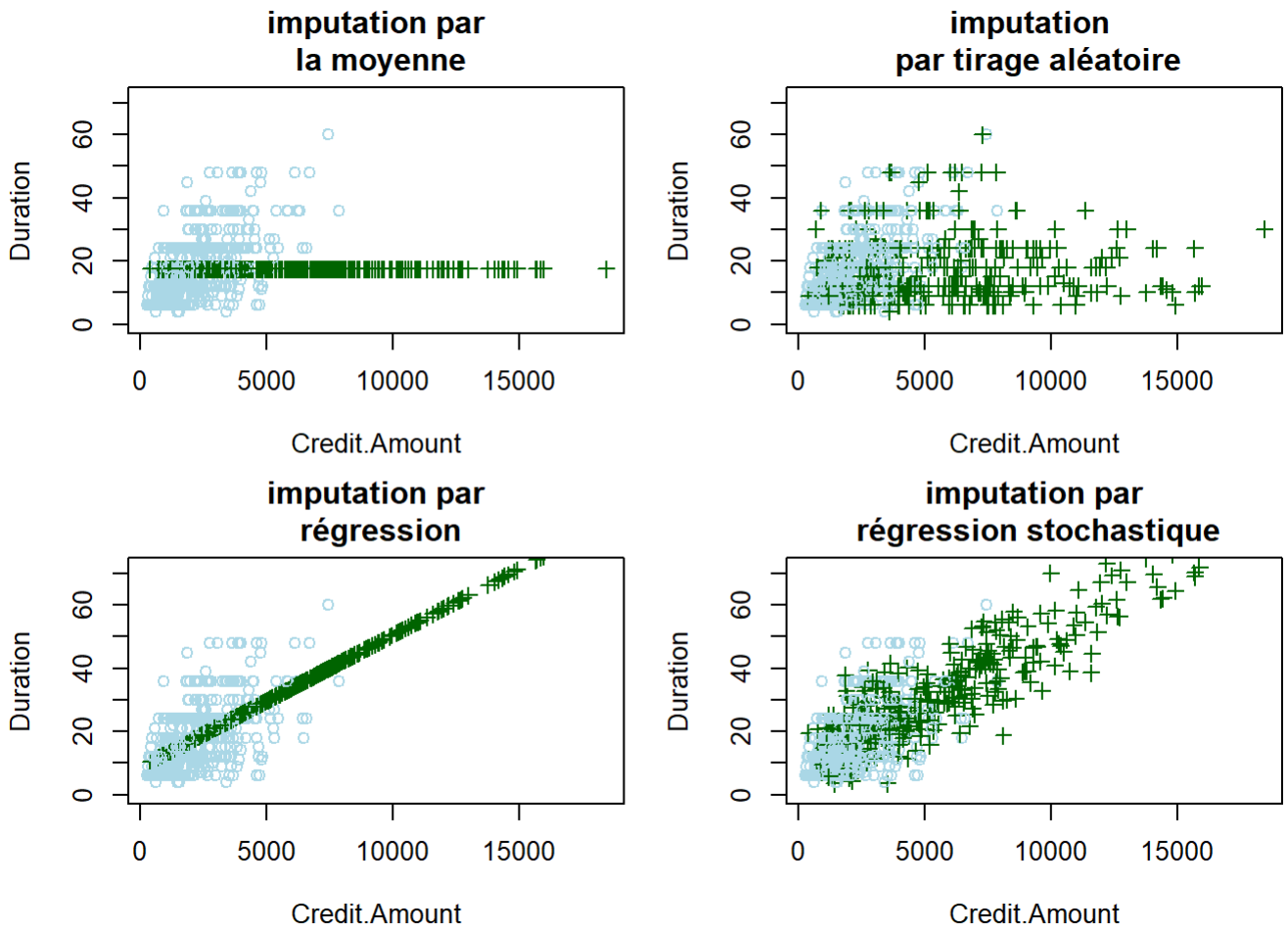


Figure 3.1: Différentes méthodes d'imputation simple. Les individus imputés sont représentés en vert, tandis que les individus complets sont en bleu.

On peut également préserver les relations sans spécifier de modèle mais en utilisant des approches non-paramétriques, par exemple en imputant par plus proches voisins : on identifie les k individus complets les plus proches au sens d'une certaine distance de l'individu à imputer, puis on en tire un au hasard parmi les k. On imputera alors l'individu en utilisant la valeur du voisin tiré. Ce type de méthode est appelé méthode *hot-deck*. Un de leurs avantages est que les valeurs imputées sont toujours comprises entre le minimum et le maximum des valeurs observées, ce qui permet d'éviter des incohérences comme par exemple l'imputation par des valeurs négatives de variables toujours positives (e.g. l'age, où la Durée d'un crédit) ce qui peut potentiellement arriver pour des méthodes paramétriques. On peut aussi utiliser des approches intermédiaires dites *semi-paramétriques* qui consistent à ajuster un modèle de régression, puis à identifier les k individus les plus proches en termes de prédiction vis-à-vis du modèle. On impute alors l'individu par la valeur observée d'un des k voisins.

Comme en statistique classique, sans données manquantes, les approches paramétriques nécessiteront peu d'observations, mais seront très dépendantes du modèle choisi, alors qu'à l'inverse, les approches non-paramétriques ne dépendront pas de cette modélisation mais nécessiteront elles davantage d'observations. Les approches semi-paramétriques constituent un compromis intéressant ce qui expliquent qu'elles soient souvent utilisées par défaut dans les logiciels.

Il est clair que les résultats obtenus suite à l'application d'une méthode de data-mining sur les données imputées seront dépendantes de la méthode d'imputation utilisée, il convient donc de choisir des méthodes d'imputation adaptées. Le choix est vaste, il existe potentiellement autant de méthodes d'imputation simple que de modèles pour expliquer une variable réponse à partir de variables explicatives (régression logistique, réseaux de neurones, arbres de décision, analyse discriminante linéaire, ...).

Une règle générale pour choisir les modèles d'imputation est de choisir des modèles au moins aussi complexes que la méthode qui sera appliquée sur les données imputées (Schafer (2003)), appelée *modèle d'analyse* (par opposition au *modèle d'imputation* qui lui ne sert qu'à compléter les données). Typiquement,

si on souhaite appliquer un modèle de régression logistique avec des termes d'interaction entre certaines variables explicatives, il faut que ces effets d'interaction soient préservés lors de l'imputation. Notons que l'imputation par forêts aléatoires (les forêts aléatoires seront abordées dans un prochain cours en tant que méthodes de prédiction) présentent par exemple un intérêt pour la gestion des interactions. De la même façon, si on veut appliquer un modèle de régression logistique sans interaction, on aura besoin de préserver au moins les relations entre la variable réponse et les variables explicatives, on n'imputera donc pas les variables marginalement car cela détruit les liaisons, etc.

3.1.2 Imputation de plusieurs variables

L'imputation est facile à mettre en oeuvre quand seule une seule variable est incomplète, mais en pratique, il y en a généralement plusieurs ce qui rend difficile l'ajustement des modèles d'imputation. Pour imputer plusieurs variables incomplètes, une méthode très populaire est l'imputation séquentielle (van Buuren (2012)). Elle consiste à définir la méthode d'imputation qui sera utilisée pour chacune des variables (par exemple un modèle de régression logistique pour les variables binaires, un modèle de régression linéaire pour les variables quantitatives, etc), puis d'imputer chaque variable tour à tour selon le modèle dédié. Plus précisément l'algorithme est le suivant :

1. ordonner les variables de celle ayant le moins de données manquantes à celle en ayant le plus
2. initialiser les valeurs manquantes par des valeurs quelconques (par exemple en faisant des tirages dans les données observées) de sorte à obtenir une première imputation du jeu de données
3. pour j de 1 à p

- ajuster le modèle d'imputation choisi pour imputer la variable j en utilisant le jeu de données restreint aux individus observés sur la variable j
- mettre à jour les données imputées sur la variable j en utilisant ce modèle

4. répéter l'étape 3. jusqu'à convergence, i.e. jusqu'à ce que la distribution des variables (imputées) se stabilise.

Les logiciels proposent généralement des méthodes par défaut pour imputer chaque variable en fonction de leur nature en utilisant le modèle linéaire généralisé (régression linéaire pour une variable quantitative, régression logistique pour une variable binaire, etc.). Néanmoins, ces choix sont propres aux logiciels et il est donc nécessaire d'une part, de savoir précisément quelles sont les méthodes d'imputation utilisées, et d'autre part, de les modifier si nécessaire.

3.1.3 Validation du modèle

Si le nombre de variables est important, il est difficile de choisir chaque modèle d'imputation avec précision. Généralement, le choix des modèles se fait de façon itérative, en imputant les données selon un choix par défaut, puis en comparant les distributions des données imputées et des données observées. Si des différences majeures sont observées, alors on pourra envisager de modifier le modèle d'imputation utilisé. Bien sûr, sous un mécanisme MAR on ne s'attend pas à trouver des distributions identiques pour les individus complets et pour les individus incomplets. Au contraire, une différence de tendance centrale dans la distribution est attendue, mais une différence dans la forme de la distribution serait beaucoup plus difficile à expliquer et constituerait un indicateur d'une modélisation inadaptée. Aussi, si seule la tendance centrale diffère, mais que cette différence est majeure alors il est plus vraisemblable que cela soit dû à une modélisation non adaptée qu'à un mécanisme très marqué. La Figure 3.2 représente les distributions des valeurs observées et imputées par régression stochastique dans les cas MCAR et MAR précédents.

Figure 3.2: Comparaison des distributions des valeurs observées et des valeurs imputées par régression stochastique pour la variable Duration dans les configurations MCAR et MAR

On voit que les distributions marginales des valeurs imputées et observées sont très différentes, notamment en termes de forme, y compris dans le cas MCAR. Ceci illustre que le choix de la méthode d'imputation n'est pas pertinent. En revanche, une imputation par plus proches voisins paraît ici bien plus satisfaisante (cf Figure 3.3).

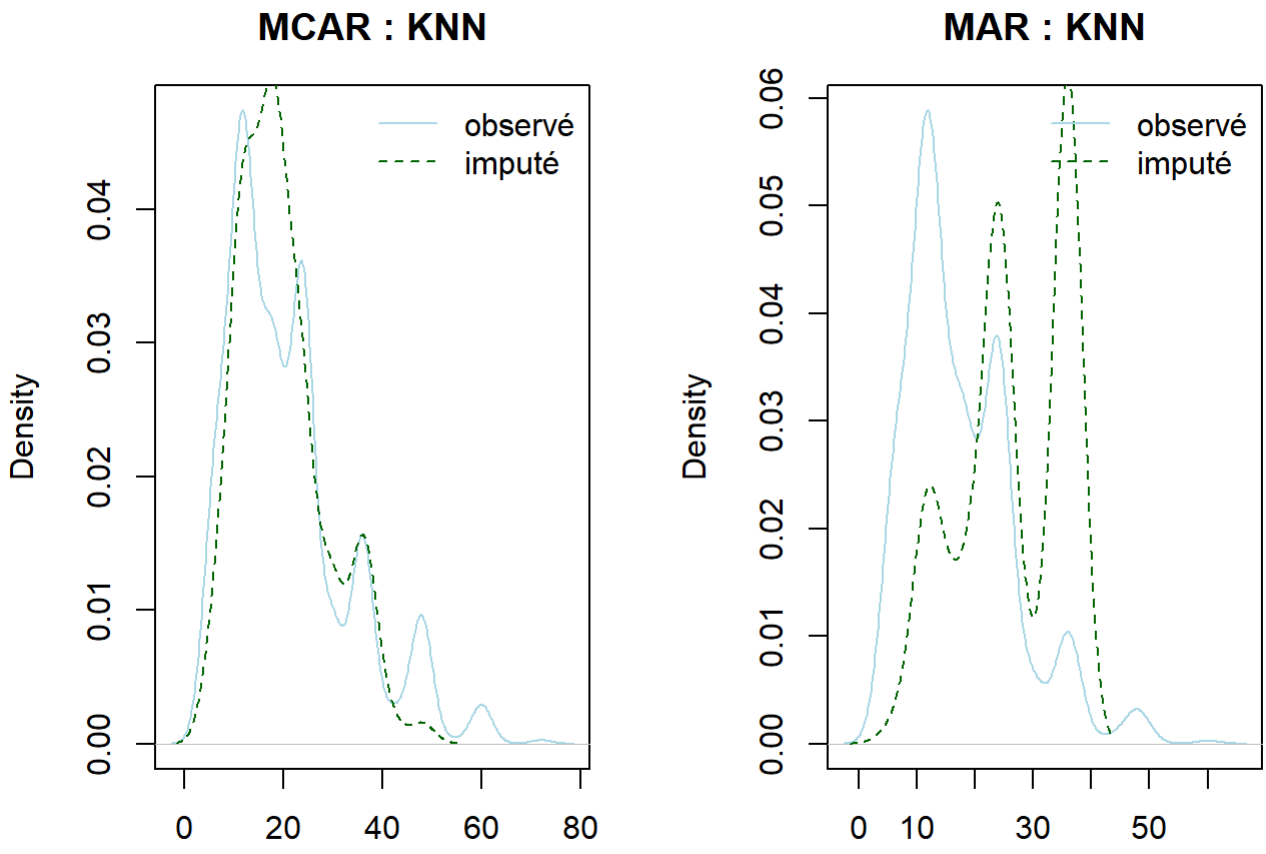


Figure 3.3: Comparaison des distributions des valeurs observées et des valeurs imputées par plus proches voisins (KNN) pour la variable Duration dans les configuration MCAR et MAR

On pourra aussi effectuer des analyses bivariées ou multivariées pour comparer les distributions jointes (i.e. reposant sur des ensembles de variables) des valeurs imputées de celles des valeurs observées.

3.1.4 Modèles joints

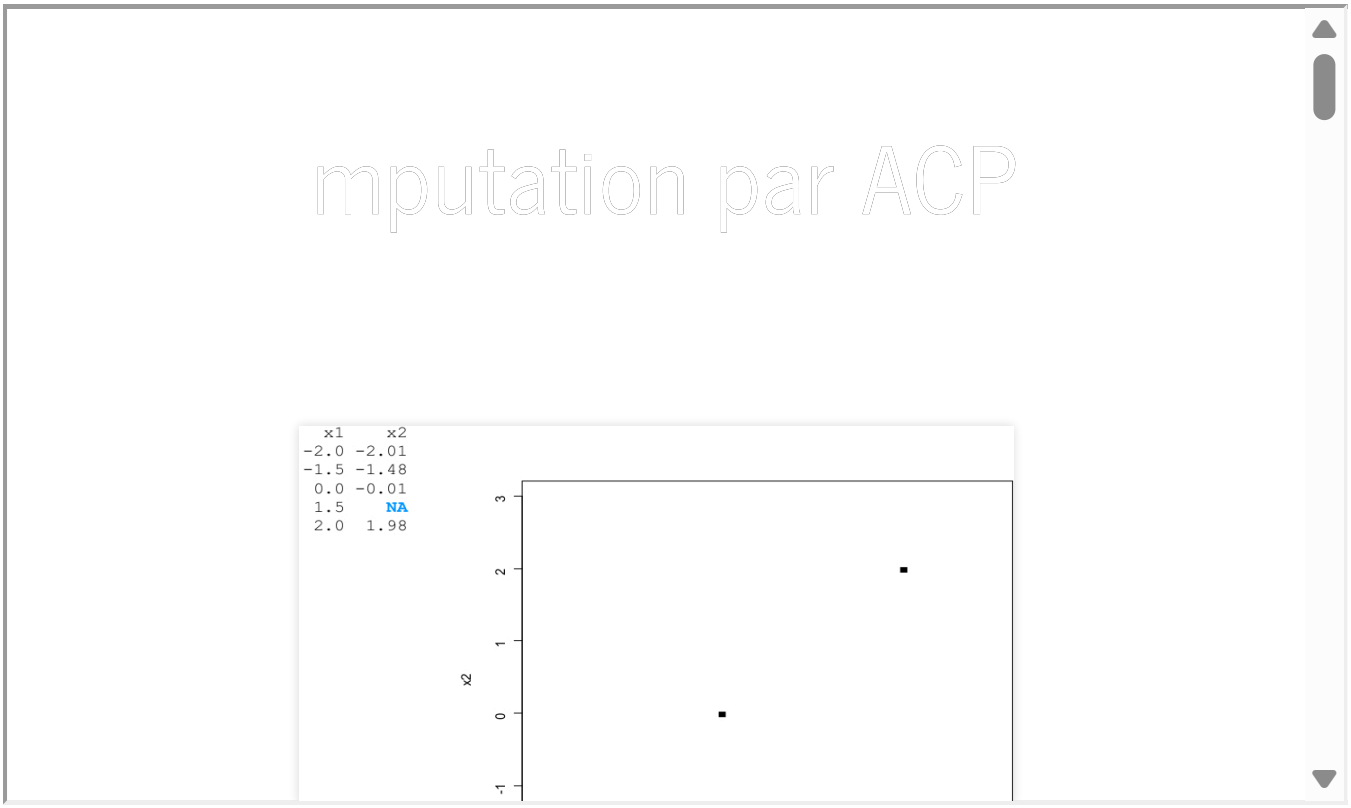
Les approches séquentielles présentent différents défauts, notamment un manque de garanties théoriques sur la convergence et un coût algorithmique non-négligeable. La problématique du temps d'exécution de l'imputation séquentielle est particulièrement vrai quand le nombre de variables est grand. Dans une telle situation, des méthodes dites *jointes* pourront être préférables. Ces méthodes d'imputation consistent à spécifier un modèle pour la distribution jointe à l'ensemble des variables (plutôt que de spécifier un modèle par variable à imputer). Par exemple, quand les variables sont toutes continues, on peut potentiellement faire l'hypothèse que les données sont issues d'une distribution Gaussienne multivariée (éventuellement après avoir effectué des transformations des données pour se ramener à la normalité). Il existe aussi des modèles d'imputation joints pour les données qualitatives ou mixtes, mais leur nombre est assez limité. Citons toutefois parmi les plus classiques

- le modèle log-linéaire ou le modèle à classes latentes pour variables qualitatives
- le "general location model" pour les données mixtes

Par conséquent, il est souvent difficile de trouver un modèle d'imputation joint bien ajusté aux données (contrairement aux approches séquentielles beaucoup plus flexibles). Néanmoins, contrairement aux méthodes séquentielles, les méthodes d'imputation jointes imputent toutes les variables simultanément, ce qui peut grandement diminuer les temps d'exécution.

Ceci est particulièrement vrai pour les méthodes d'imputation par analyse factorielle. Ces méthodes d'imputation par modèle joints consistent à prédire les valeurs manquantes par la projection des individus sur les sous-espaces factoriels. Par exemple, dans le cas de données quantitatives et d'un sous espace de

dimension 1, il s'agit de prédire les valeurs manquantes selon la droite d'ACP. Le principe de ces méthodes est illustré ci-dessous :



Il est également possible de généraliser à un espace de dimension 2 (plan) ou autre, mais aussi à des données qualitatives, en imputant par ACM ou mixte en imputant par AFDM.

Ces méthodes sont particulièrement intéressantes dans les cas suivants :

- multicolinéarité
- grande dimension (i.e. nombre d'individus inférieur au nombre de variables)
- grand nombre d'observations
- nombre élevé de variables
- présence de données mixtes

Pour mieux comprendre le fonctionnement de ces méthodes, on pourra consulter [chap.3]Audigier (2015), tandis que l'on consultera Josse and Husson (2016) et [Appendix]Audigier (2015) pour leur mise en application via le package R missMDA.

3.1.5 Analyse de sensibilité à l'hypothèse sur le mécanisme

L'avantage des méthodes d'imputation précédentes est qu'elles ne nécessitent que de modéliser la distribution des données complètes $\mathbf{X}^{full}$, mais pas le mécanisme à l'origine des données manquantes (on notera par exemple qu'aucune hypothèse sur la distribution de R n'a été spécifiée dans les exemples précédents). Ceci n'est en effet pas nécessaire sous l'hypothèse MAR (ce résultat est admis). Néanmoins, cette hypothèse étant invérifiable, il est souhaitable d'évaluer la robustesse des résultats obtenus à un écart vis-à-vis de cette hypothèse. Cette analyse appelée *analyse de sensibilité* est généralement effectuée une fois le modèle d'imputation et le modèle d'analyse bien arrêtés. Pour effectuer l'analyse de sensibilité, une stratégie classique est d'ajouter une constante aux données imputées et de voir comment varient les résultats de l'analyse en fonction [p289-290]Enders (2010). Notons, qu'il est indispensable de choisir cette constante de façon à rester dans des cas crédibles, ce qui nécessite une bonne connaissance des données. Cette analyse doit donc être effectuée avec soin pour être vraiment utile.

3.2 Imputation multiple

Le défaut des méthodes d'imputation simples quelles qu'elles soient est qu'une fois les données imputées, on ne fait plus de distinction entre les données imputées et les données observées. Or, les données imputées souffrent d'une certaine incertitude et celle-ci n'est pas reflétée au travers de l'imputation. Ceci constitue essentiellement un problème quand on souhaite construire des intervalles de confiance pour des paramètres d'un modèle une fois les données imputées. En effet, en ignorant l'incertitude sur les données imputées, on sous-estime la variabilité des estimateurs ce qui conduit à obtenir les intervalles de confiance trop courts, i.e. dont le taux de couverture est en dessous de celui attendu. Par exemple, si on construit un intervalle classique à 95% pour une moyenne à partir d'une variable imputée par imputation simple, alors cet intervalle aura en réalité un niveau de confiance inférieur à 95%. Pour le vérifier, on pourra effectuer la simulation suivante sous R :

```
nsim <- 2000 #nombre de simulations effectuees
res <- matrix(NA, nsim, 2)
for(ii in 1:nsim){
  #creation d'un jeu de données avec deux variables x et y
  x <- rnorm(500)
  y <- 2*x+rnorm(500, sd = .5)
  #ajout de données manquantes sur y selon un mécanisme mar
  ismar<-sapply(x,
                FUN = function(xx){
                  yy <- xx
                  prob <- pnorm(1.2*yy-.5)
                  res <- sample(c(T,F), size = 1, prob = c(prob, 1-prob))
                  return(res)})
  ymar <- y;ymar[ismar] <- NA

  #imputation simple par régression stochastique de y
  res.lm <- lm(ymar[!ismar]~x[!ismar])
  ystoch <- ymar
  ystoch[ismar] <- cbind(1, x[ismar])%*%res.lm$coefficients+rnorm(sum(ismar), sd=summary(res.
lm)$sigma)

  #calcul de la moyenne de y après imputation et comparaison de la moyenne théorique (0) aux
bornes de l'intervalle de confiance  à 95% de l'espérance de y
  res[ii,] <- c(mean(ystoch),
              (0<=(mean(ystoch)+qt(p = 0.975,df=499)*sd(ystoch)/sqrt(500))&(
                0>=(mean(ystoch)-qt(p = 0.975,df=499)*sd(ystoch)/sqrt(500)))))
}

#évolution de la moyenne empirique de l'estimateur de la moyenne en fonction du nombre de sim
ulations
# et évolution du taux de couverture de l'intervalle de confiance construit à partir des donn
ées imputées en fonction du nombre de simulations
par(mfrow = c(1,2), mar = c(4, 5, 3, 1) + 0.1)
plot(cumsum(res[,1])/(1:nsim),
     ylab = expression(bar(y[imp])),
     xlab = "nombre simulations",
     cex = .2)
abline(h = 0, col = 2)
plot(cumsum(res[,2])/(1:nsim),
     ylab = "taux de couverture",
     xlab = "nombre simulations",
     cex = .2)
abline(h = 0.95, col = 2)
```

Pour remédier au problème des intervalles de confiance avec un taux de couverture trop faible, on peut utiliser les méthodes d'imputation multiple.

Le principe de l'imputation multiple est de proposer plusieurs tableaux imputés (au moins 3). On applique ensuite la méthode d'analyse souhaitée sur chaque tableau, puis on agrège les résultats entre eux selon des règles bien spécifiques dites *règles de Rubin* (voir illustration Figure 3.4).

Tableau incomplet Imputation Analyse Agrégation

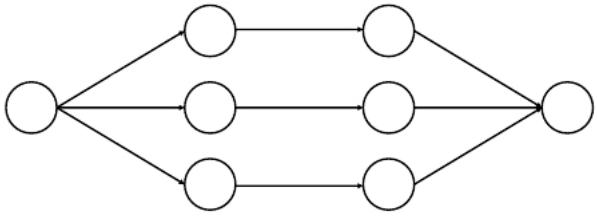


Figure 3.4: Schéma de l'imputation multiple

Pour générer plusieurs tableaux imputés, on définit d'abord un modèle d'imputation (e.g. un modèle de régression), puis on génère M jeux de paramètres pour ce modèle (e.g. en ré-échantillonnant M fois les données par bootstrap puis en ajustant le modèle sur régression sur chaque réplication). On impute alors les données selon chacun de ces M paramètres. On obtient ainsi M tableaux imputés. **ATTENTION** : on voit ici que l'imputation multiple ne se résume pas à une succession d'imputations simples car les paramètres du modèle d'imputation diffèrent selon chaque tableau imputé !

Ainsi, l'imputation multiple permet de refléter, au travers des M valeurs imputées d'une même donnée manquante, l'incertitude sur les données imputées. Une fois les données complétées, n'importe quelle méthode d'analyse peut être appliquée, mais il n'est pas forcément possible d'agréger les résultats obtenus. En effet, les règles de Rubin s'appliquent essentiellement aux modèles linéaires généralisés (i.e. régression linéaire, régression logistique, analyse de la variance, analyse de la covariance) afin d'obtenir une unique estimation ponctuelle des coefficients de régression et une unique estimation de la variabilité associée. Il est aussi possible de les appliquer sur des estimations de moyennes, ou de proportions (on pourra se reporter à Marshall et al. (2009) pour l'agrégation d'autres quantités). Appelant $\hat{\theta}_m$ les coefficients de régression estimés sur le tableau m, et $\widehat{Var}\left(\hat{\theta}_m\right)$ l'estimation de la variabilité associée, on obtient une unique estimation de θ selon la moyenne des différentes estimation (Équation (3.1)), tandis que la variance de $\hat{\theta}$ s'obtient en sommant une variance intra-tableau et une variance inter-tableaux (Équation (3.2)). Le terme correcteur (1+1/M) permet de tenir compte de la variabilité due à la simulation, il tend vers 1 quand M tend vers l'infini.

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{3.1}$$

$$\widehat{Var}\left(\hat{\theta}\right) = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\theta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2 \tag{3.2}$$

On recommande généralement d'utiliser $M = 3$ ou $M = 5$ tableaux imputés, mais on aura intérêt à prendre M grand de façon à diminuer la variabilité liée à la simulation et donc avoir une variance estimée $\widehat{Var}\left(\hat{\theta}\right)$ plus petite.

Remarque : l'imputation multiple offre un moyen supplémentaire de validation des modèles d'imputation. Certains logiciels proposent en effet de générer des valeurs plausibles également pour les données observées (on parle d'*overimputation* (Blackwell, Honaker, and King (2015))). Ainsi, en générant plusieurs centaines de tableaux, on peut construire un intervalle de prédiction à 95% pour les valeurs observées (des variables quantitatives uniquement) selon la méthode des percentiles. Si le modèle d'imputation est bien choisi, alors on s'attend à ce que 95% des valeurs observées soient contenues dans leur intervalle (en réalité un peu plus car les données observées ont servi à construire les modèles d'imputation). La figure 3.5 donne un exemple de représentation de quelques uns de ces intervalles pour la variable incomplète *Duration* imputée par régression.

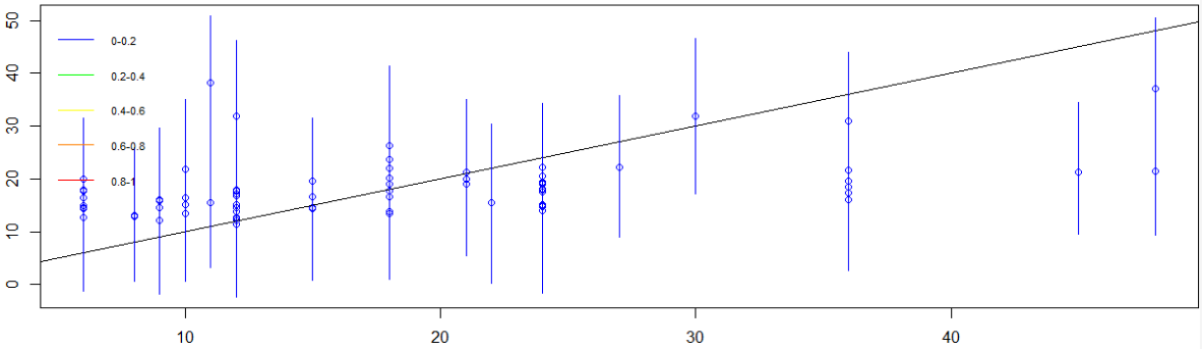


Figure 3.5: Evaluation de l'ajustement du modèle par overimputation. En abscisse on positionne 50 valeurs observées (prises au hasard) d'une variable et en ordonnée l'intervalle de prédiction à 95% pour ces valeurs. Si le modèle d'imputation est bien ajusté, alors on s'attend à ce que 95% des intervalles coupent la première bissectrice.

4 Autres méthodes de traitement

L'imputation n'est pas la seule façon de gérer les données manquantes. Il est aussi possible d'adapter les méthodes basées sur la vraisemblance, utilisées dans le cadre usuel sans données manquantes (e.g. estimation paramétrique par maximum de vraisemblance), à la présence de données manquantes. Pour cela on utilise des algorithmes itératifs qui reviennent grosso modo à alterner des étapes d'imputation des données manquantes et d'estimation des paramètres du modèle (e.g. par maximum de vraisemblance) sur les données imputées. Ceci est aussi possible dans un cadre Bayésien. Dans ce cas on alterne génération des données imputées, et tirage des paramètres dans leur distribution a posteriori. Ces algorithmes portent respectivement les noms d'algorithme EM (Expectation-Maximization) et DA (data-augmentation). Néanmoins, il n'est pas toujours possible de développer de tels algorithmes en fonction des modèles d'analyse considérés.

Remarque : La différence avec les approches d'imputation précédentes (Section 3), est que le modèle qui sert à imputer les données est ici le même que celui dont on analysera les paramètres in fine, tandis qu'en imputation, on peut choisir des modèles d'imputation et d'analyse différents. Par exemple, on peut imputer par plus proches voisins et appliquer une régression linéaire sur les données imputées ce qui n'est pas possible pour des algorithmes EM ou DA.

En plus des méthodes basées sur la vraisemblance, les approches par pondération permettent aussi de gérer les données manquantes. Elles consistent à attribuer un poids aux individus sans données manquantes de façon à corriger le biais observé dans la mise en oeuvre de la méthode du cas complet. Cette idée est issue de la théorie des sondages où, pour limiter le nombre de personnes interrogées, la probabilité qu'un individu fasse partie de l'étude n'est pas uniforme mais définie selon un plan, dit plan de sondage. Par la suite, pour inférer sur la population, les individus sont repondérés de façon inversement proportionnelle à la probabilité qu'ils avaient de faire partie de l'échantillon. L'extension à la gestion des données manquantes est directe : on établit pour chaque individu complet la probabilité qu'il a d'être manquant et on lui affecte un poids inverse à cette probabilité. On applique ensuite la méthode d'analyse souhaitée en tenant compte de ces poids. Ainsi, les méthodes de pondération nécessitent de modéliser le mécanisme à l'origine des données manquantes car c'est lui qui permet de définir les poids utilisés. En revanche, il n'est pas nécessaire de spécifier la distribution des données. Toutefois, les raisons pour lesquelles les données sont manquantes sont souvent mal connues ce qui rend l'estimation des poids très difficile en pratique.

On pourra utilement consulter ce livre Gégout-Petit et al. (2022), écrit en français, afin d'approfondir ces autres techniques de gestion des données manquantes.

5 Conclusion

Les données manquantes sont incontournables en data-mining. Leur gestion nécessite d'abord de comprendre la position des données manquantes et leurs relations avec les données observées. Parmi les méthodes à disposition, les méthodes d'imputation sont certainement les plus pratiques pour les gérer. En particulier, les méthodes d'imputation séquentielles permettent d'utiliser des modèles d'imputation assez souples et donc de proposer des données imputées dont la distribution est assez proche de celle (inconnue) des données manquantes. Les méthodes d'imputation multiple présentent l'avantage de pouvoir construire des intervalles de confiance pour certains paramètres, mais si la méthode que l'on veut appliquer une fois les données imputées ne nécessite pas de calculer d'intervalle de confiance, ces méthodes ne présentent plus nécessairement d'intérêt majeur vis-à-vis des méthodes d'imputation simple.

Choisir une méthode d'imputation est tout aussi délicat que de choisir un modèle pertinent pour analyser les données complétées, les problématiques sont les mêmes que dans le cadre usuel sans données manquantes (validation de modèle, prise en compte des interaction, gestion des valeurs aberrantes, sélection de variables...). Ce choix sera d'autant plus important que la quantité de données manquantes sera élevée.

L'hypothèse MAR, est l'hypothèse faite par défaut par la plupart des méthodes de gestion des données manquantes. L'analyse de sensibilité est toujours délicate, et il sera en général préférable de passer plus de temps à choisir des modèles d'imputation adaptés à la distribution des données que de faire une analyse de sensibilité superficielle ([p.93]van Buuren (2012)). Dans le cas où l'hypothèse MAR ne paraît pas raisonnable, des solutions ont également été proposées (voir imputation par Pattern mixture models ou Selection models).

Enfin, notons que les méthodes d'imputation offrent un moyen de gestion des valeurs aberrantes. On peut en effet considérer ces données comme manquantes puis utiliser la donnée imputée dans l'analyse plutôt que sa valeur observée.

Il existe de nombreuses solutions logicielles pour l'imputation des données (voir ici (https://cran.r-project.org/web/views/MissingData.html) pour une liste relativement complète des possibilités sous R). En particulier, les packages R *mice*, *VIM*, *mi* proposent des méthodes d'imputation séquentielles. On notera que le package *VIM* a la particularité de proposer des méthodes d'imputation gérant les valeurs aberrantes et des outils d'analyse exploratoire pour les données manquantes. Le package *micemd* pourra également être utilisé pour paralléliser les méthodes d'imputation multiple disponibles dans le package *mice* afin de diminuer les temps de calcul. Des approches jointes sont également disponibles dans les packages R *Amelia* (imputation par modèle Gaussien), *missMDA* (imputation par analyse factorielle), *cat* (imputation par modèle log-linéaire pour données qualitatives), *mix* (imputation par *General location model* pour données mixtes). On notera que la communauté des utilisateurs d'Amelia est très active et qu'il est possible de s'abonner à une mailing list pour poser directement des questions aux auteurs, ou consulter les questions archivées. Citons enfin la fonction *proc MI* du logiciel SAS qui propose des méthodes d'imputation séquentielles et jointes.

Références

Allison, P. D. 2002. *Missing Data*. Thousand Oaks, CA: Sage.

Audigier, Vincent. 2015. "Multiple imputation using principal component methods : A new methodology to deal with missing values." Theses, Agrocampus Ouest. https://pastel.archives-ouvertes.fr/tel-01336206 (https://pastel.archives-ouvertes.fr/tel-01336206).

Blackwell, M., J. Honaker, and G. King. 2015. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods and Research*, 1–39.

Enders, C. K. 2010. *Applied Missing Data Analysis*. Methodology in the Social Sciences. Guilford Publications.

Fitzmaurice, G. M., M. G. Kenward, G. Molenberghs, G. Verbeke, and A. A. Tsiatis. 2014. "Missing Data: Introduction and Statistical Preliminaries." In *Handbooks of Missing Data Methodology*, edited by G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, 3–22. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. New York: CRC Press.

Gégout-Petit, Anne, Myriam Maumy-Bertrand, Gilbert Saporta, and Christine Thomas-Agnan. 2022. *Données manquantes*. Editions Technip. https://hal-cnam.archives-ouvertes.fr/hal-03696270 (https://hal-cnam.archives-ouvertes.fr/hal-03696270).

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1–47. https://doi.org/10.18637/jss.v045.i07 (https://doi.org/10.18637/jss.v045.i07).

Josse, J., and F. Husson. 2016. "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software, Articles* 70 (1): 1–31. https://doi.org/10.18637/jss.v070.i01 (https://doi.org/10.18637/jss.v070.i01).

Little, R. J. A. 1995. "Modelling the Drop-Out Mechanism in Repeated Measures Studies." *Journal of the American Statistical Association* 90: 1112–21.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. New-York: Wiley series in probability; statistics.

Marshall, A., D. G. Altman, R. L. Holder, and P. Royston. 2009. "Combining Estimates of Interest in Prognostic Modelling Studies After Multiple Imputation: Current Practice and Guidelines." *Bmc Medical Research Methodology* 9 (5): 57.

Nakache, Gueguen, J.-P. 2005. "Analyse Multidimensionnelle de Données Incomplètes." *Revue de Statistique Appliquée* 53 (3): 35–62. http://eudml.org/doc/106567 (http://eudml.org/doc/106567).

Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–92.

Schafer, J. L. 2003. "Multiple imputation in multivariate problems when the imputation and analysis models differ." *Statistica Neerlandica* 57 (1): 19–35.

van Buuren, S. 2012. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. Hardcover; Chapman; Hall/CRC. https://stefvanbuuren.name/fimd/ (https://stefvanbuuren.name/fimd/).