MASTER TRAITEMENT DE L'INFORMATION ET EXPLOITATION DES DONNÉES

—TRIED—

STAGE M2-TRIED (DATA SCIENCE)

THÈME

Consensus pondéré pour la gestion des données manquantes en classification non-supervisée

Rédigé par :
Mouhamadou Lamine NDAO

Sous la supervision de :
Vincent AUDIGIER,
Ndèye NIANG

Septembre 2021

Sommaire

REMERCIEMENTS ii

Liste des figures iii

Liste des tables iv

INTRODUCTION GENERALE 1

I Cadre méthodologique 3

1 Méthodes de classification 4

2 Consensus de partitions 8

3 Imputation multiple 14

II Expérimentations et résultats 21

4 Apport du consensus pondéré en présence de données manquantes 22

5 Application sur des données réelles 41

CONCLUSION 48

BIBLIOGRAPHIES A

ANNEXES I

Table des matières VI

REMERCIEMENTS

Je remercie toutes les personnes qui de près ou de loin m'ont aidé à réaliser ce travail. Je remercie plus particulièrement mes encadrants Mme. Ndèye Niang et M. Vincent Audigier.

Table des figures

3.1 Schéma d'une imputation multiple 15
3.2 Illustration du pas entre tableaux imputés 17

4.1 Evolution de taux de couverture en fonction du pas et selon le nombre de tableaux m 25
4.2 Evolution de l'erreur quadratique moyenne en fonction du pas et selon le nombre de tableaux m 27
4.3 Plan d'expérimentation pour $\delta = 1$ 30
4.4 L'évolution des indices de Rand ajustés moyens en fonction du pas et du nombre de tableaux considérés pendant l'IM ; approche NMF 34
4.5 Box-plot des indices de Rand ajustés des partitions obtenues par consensus simple NMF en fonction du nombre de tableaux m pour les 30 expériences . . 35
4.6 L'évolution des indices de Rand ajustés moyens en fonction du pas et du nombre de tableaux considérés pendant l'IM ; approche WNMF 38
4.7 Box-plots des indices de Rand ajustés des 30 expériences en fonction du nombre de tableaux considérés m et de la méthode de consensus NMF et WNMF . . . 39

5.1 Corrélation entre les polluants 42
5.2 Instabilité en fonction du nombre de classes K 44
5.3 L'indice de Rand entre les partitions obtenues sur les 20 tableaux 45
5.4 Visualisation des classes obtenues avec les deux approches de consensus de partitions sur le premier plan de l'ACP sur le tableau après imputation par la méthode factorielle 46
5.5 Les variables les plus discriminantes par rapport aux deux partitions obtenues 46
5.6 Distribution du maximum des moyennes pendant 30 minutes en termes de concentration en monoxyde carbone 47
5.7 Distribution du maximum des moyennes pendant 30 minutes et 1 heure en termes de concentration en monoxyde carbone II
5.8 Matrice des indices de Rand entre les 1000 partitions deux à deux issues des 1000 tableaux générées lors des 4 première V
5.9 Distribution des indices de Rand selon le pas pour m=20 tableaux des 4 premières expériences VI

Liste des tableaux

2.1 Matrice de connectivité M 9
2.2 Matrice d'indicatrices H 9

5.1 Les poids obtenus avec l'approche WNMF 45
5.2 Description univariée des données à mettre en annexe I
5.3 Description variables de polluant IV

INTRODUCTION GÉNÉRALE

L'un des problèmes récurrents en data mining est la présence de données manquantes. Le problème qui se pose est le fait que les méthodes statistiques standard ont été développées pour analyser des ensembles de données complètes. Ces données peuvent se présenter sous différentes formes dont la plus classique est celle rectangulaire contenant des données numériques (sous forme de tableau). La présence de données manquantes au sein de ces ensembles de données peut avoir plusieurs origines. Dans le cadre d'une enquête, elle peut être due à un refus de répondre à certaines questions jugées sensibles par l'enquêté (par exemple le salaire), par une non compréhension de la question par l'enquêté ou tout simplement par une erreur de saisie lors de la collecte des données. Les outils classiques pour l'analyse de ces données ne permettent pas généralement une prise en compte directe de ces données manquantes. Ainsi, il apparaît une nécessité d'outils pour la prise en compte de ces dernières. Pour cela, deux approches sont souvent adoptées. Une première approche utilise des techniques avancées en data mining. Selon cette approche, il existe deux stratégies classiques : une première basée sur l'algorithme Expectation Maximization (EM) dans un cadre fréquentiste et une deuxième basée sur l'algorithme de Data-Augmentation (DA) dans un cadre bayésien. Les deux consistent à adapter les méthodes statistiques en présence de données manquantes [1]. La deuxième approche dite "méthodes ad-hoc" consiste à faire un pré-traitement pour se ramener à une base complète. Selon cette approche, il existe un nombre de techniques qui sont souvent adoptées par les analystes. Par exemple, une des possibilités est de recoder les non réponses par des modalités comme « ne sait pas » ou « refus », une autre possibilité consiste à supprimer les observations concernées et ne travailler qu'avec les données complètes. Cette dernière méthode dite « complete case analysis » présente souvent des inconvénients tels les problèmes de représentativité, présence de biais dans les calculs d'estimateurs, etc. Pour palier à ces défauts, les auteurs [2] et [3] ont proposé des approches qui consistent à « retrouver » approximativement ces données manquantes : on parle de méthodes d'imputation. Classiquement, il existe plusieurs approches d'imputation que nous détaillerons dans la suite de ce rapport. Ces approches peuvent être regroupées en deux grands groupes : les méthodes d'imputation simple et les méthodes d'imputation multiple. Concernant le premier groupe de méthode, le principe consiste à remplacer chaque donnée manquante par une valeur unique. Cependant, ce sont des méthodes d'imputation qui présentent quelques défauts notamment lorsqu'il s'agit de calcul des intervalles de confiance pendant l'analyse. En effet, ces méthodes ne donnent pas une idée sur la variabilité des valeurs obtenues après imputation. Autrement dit, une fois les données imputées on ne fait plus la distinction entre valeurs observées et valeurs imputées alors que les valeurs imputées sont incertaines. C'est dans ce sens que l'on fait souvent appel aux méthodes d'imputation multiple. Le principe de cette technique consiste à proposer, dans un premier temps, plusieurs tableaux imputés. Chacun de ces tableaux est imputé de manière à refléter l'incertitude sur les données imputées. Ensuite, sera appliquée une méthode d'analyse souhaitée sur chaque

tableau, par exemple une régression. Enfin, les résultats obtenus sur les différents tableaux sont agrégés selon des règles bien spécifiques dites règles de Rubin [4] afin de disposer de l'estimation des paramètres du modèle et des intervalles de confiance associés.

Il faut noter qu'initialement, ces méthodes d'imputation multiple ont été développées dans le cadre supervisé (résultats d'une régression par exemple). Elles ont ensuite, été étendues dans le cadre de la classification non-supervisée. En particulier, l'agrégation des partitions obtenues sur chaque tableau est effectuée selon des techniques de consensus de partitions [5]. L'idée principale des méthodes de consensus est d'agglomérer les partitions obtenues à partir de chaque tableau (appelées partitions contributives) en une partition globale. Dans cette phase d'agrégation, deux approches sont possibles : on peut considérer que toutes les partitions sont de qualité identique et on leur attribut alors le même poids : on parle alors de consensus simple ; dans le cas contraire, on considère que les partitions contributives n'ont pas le même poids : on parle alors de consensus pondéré.

Lorsqu'on applique une méthode d'imputation multiple où les tableaux sont théoriquement imputés de façon indépendante, il est assez naturel de considérer que toutes les partitions doivent avoir le même poids. Or, en pratique cette indépendance n'est pas toujours facile à vérifier et il se peut que certaines partitions soient redondantes car certains tableaux imputés sont très similaires, ce qui peut détériorer la précision de la partition consensus obtenue par agrégation. On se demande alors, s'il ne serait pas plus intéressant d'affecter des poids aux partitions selon leur particularité.

Ainsi, l'objectif général de ce stage est d'évaluer l'apport du consensus pondéré lorsque les données manquantes sont gérées par imputation multiple. Plus précisément, il s'agit de voir la qualité du consensus de partitions lorsque l'indépendance entre les tableaux imputés varie, ensuite évaluer l'apport du consensus pondéré par rapport au consensus simple.

Pour cela, le rapport sera structuré en deux grandes parties. Dans la première partie, nous présenterons le cadre méthodologique de ce travail, en particulier les aspects théoriques des différentes méthodes que nous avons utilisées. Pour la présentation de chacune de ces méthodes, nous ferons une brève revue de l'approche en question. Dans la deuxième partie, il s'agira de présenter dans un premier temps les différentes simulations faites pour évaluer l'apport du consensus pondéré par rapport au consensus simple, puis l'application de ces techniques sur des données réelles.

Première partie

Cadre méthodologique

Chapitre 1

Méthodes de classification

Introduction

Ce chapitre a pour objectif d'expliquer la notion de classification en présentant quelques méthodes de classification que nous avons utilisées au cours de cette étude.

Le but de la classification est de découvrir des groupes (ou classes) cohérents d'observations dans un ensemble d'observations non-étiquetées. Les classes dégagées par une méthode de classification sont inconnues a priori. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (en anglais classification) pour lesquelles une typologie est a priori connue, au moins pour un échantillon d'apprentissage.

Les classes recherchées forment des ensembles d'observations homogènes au sens d'un critère de similarité telle que la distance entre deux observations. Lakhmi C. Jain. and al. [6] ont présenté une revue des méthodes de classification. Ces méthodes peuvent être divisées en deux grandes familles.

D'une part, on retrouve la famille des méthodes de classification à base de modèle de mélanges dans lesquelles on recherche la structure de classification à l'aide d'hypothèses probabilistes et statistiques. Ces méthodes conduisent souvent à des partitions floues qui définissent la probabilité d'appartenance d'une observation à une classe.

D'autre part, on retrouve la famille des méthodes basées sur la notion de distance entre observations. Ces méthodes essentiellement géométriques, n'émettent aucune hypothèse sur la structure des données. Les approches usuelles de classification dites géométriques peuvent être divisées en deux groupes : les méthodes hiérarchiques, qui fournissent une suite de partitions emboîtées représentées par des arbres ou dendrogrammes, et les méthodes de partitionnement direct en un nombre de classes qui doit être fixé a priori. Dans cette étude, on s'est intéressé à cette dernière approche de classification. Plus particulièrement, à la méthode des k-means parce qu'il s'agit d'une approche très classique.

1.1 Algorithmes de classification

La méthode des k-means [7] appartient à la famille des méthodes de classification géométrique. Dans cette famille de méthodes, le centre de classe correspond à un vecteur référent des observations de la classe. La méthode des k-moyennes cherche à optimiser le critère d'inertie suivant :

$$J(C) = \sum_{k=1}^{K} \sum_{z_i \in c_k} \|z_i - w_k\|^2 = \sum_{k=1}^{K} J_{c_k} \tag{1.1}$$

où w_k représente le référent de la classe c_k d'effectif n_k et $n_k J_{c_k} = \sum_{z_i \in c_k} \|z_i - w_k\|^2$ désigne l'inertie de la classe k. L'optimisation de la fonction objectif des k-means est basée sur un processus itératif alterné d'affectation des observations aux K classes et d'actualisation des vecteurs référents. La phase d'initialisation consiste à choisir arbitrairement K centres de classes et à répéter les deux étapes suivantes à chaque itération t jusqu'à la convergence et stabilisation des classes.

1. *Etape d'affectation* : à l'itération t affecter chaque observation z_i au centre le plus proche parmi les centres

   ${w_k}^t$:

   $$\in c_k^t \quad si \quad d(z_i, w_k^t) = \min_{k=1...K}(d(z_i, w_k^t)) \tag{1.2}$$

   où $d(x, y)$ correspond à la distance entre x et y

2. *Etape de minimisation* : il s'agit de calculer les nouveaux centres des classes de telle manière à minimiser la distance entre les observations appartenant à la classe et le centre de la classe.

Cet algorithme dépend fortement de l'initialisation des centres des classes et converge souvent vers des minima locaux plutôt que globaux.

Une variante des k-moyennes, permettant de surmonter les défauts de k-means, a été proposée dans global k-means [8] . Il s'agit d'une approche incrémentale de classification qui incrémente progressivement le nombre de classes recherché en utilisant le k-means comme procédure de recherche locale. L'algorithme commence par résoudre le problème d'une seule classe, le centre sera dans ce cas le barycentre des observations « G ». Pour le problème de deux classes, on lance n k-means en initialisant par le couple (G, z_i) à chaque fois (z_i est une observation parmi n), on retient la solution qui minimise l'erreur de classification. Le problème de K classes sera résolu donc en prenant les centres retenus à l'itération $K-1$ $(w_1, w_2, ..., w_{K-1})$. Cet algorithme ne pose pas le problème d'initialisation des centres. Il calcule d'une manière déterministe les centres des classes. Il fournit toutes les solutions possibles avec $1, \ldots, K$ clusters résolvant le problème de K classes.

Une autre variante de k-means a été proposée pour résoudre le problème des K classes qui ne sont pas linéairement séparables. Cette variante est le kernel-kmeans. Dans cette version de k-means, les observations sont projetés de l'espace d'entrée vers un autre espace de plus grande dimension qu'on appelle espace de fonctions, à travers une transformation non linéaire. La séparation des observations dans l'espace des fonctions est linéaire mais elle correspond à une séparation non linéaire dans l'espace d'entrée. Dans ce travail, nous nous sommes limités à la version classique des k-means.

1.2 Critère de validité d'une partition

En classification, les étiquettes des données sont inconnues a priori. Ainsi, il est important de mettre en place des procédures de validation de la classification. Plusieurs critères de validation ont été introduits [9] et [10] . Ces critères peuvent être regroupés en deux grandes familles :

- Les critères internes : ils utilisent seulement des informations internes aux données. Ces critères sont basés sur les notions de compacité et de séparabilité des classes.
- Les critères externes : ils permettent de comparer la partition fournie par une méthode de classification avec une autre partition de référence.

En classification, il est toujours possible de calculer des critères internes. Par contre, ce n'est pas le cas pour les critères externes car on ne dispose pas toujours d'une partition de référence. Un cas particulier où on dispose de cette partition est celui de la simulation. Il existe un grand nombre de critères tant internes qu'externes [9, 10]. Nous avons utilisé la silhouette comme critère interne et l'indice de Rand pour le critère externe.

1.2.1 Coefficient de silhouette

Le coefficient de silhouette est un critère interne qui permet d'évaluer la compacité et la séparabilité des classes. Il est défini pour toutes les observations, toutes les classes et pour la classification :

- pour une observation z_i : $S_z(z_i) = \frac{b_{z_i} - a_{z_i}}{max(b_{z_i}, a_{z_i})}$ où b_{z_i} est la moyenne des distances entre z_i et toutes les observations n'appartenant pas à sa classe alors que a_{z_i} est la distance moyenne entre z_i et les observations appartenant à sa classe.
- Pour une classe, on définit la moyenne des S_z : $S_c(c_k) = \frac{1}{n_k} S_z(z_i)$ où n_k est le nombre d'observations de la classe c_k.
- Enfin le coefficient de silhouette d'une classification est la moyenne des S_c :

$$S(C) = \frac{1}{K} \sum_{k=1}^{K} S_c(c_k) \tag{1.3}$$

La valeur de cet indice est comprise entre 1 et -1.

Pour une observation, si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc mal classé. A l'inverse, si cette différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc bien classé. Le coefficient de silhouette proprement dit est la moyenne des coefficients de silhouette pour tous les points. Nous notons que le calcul de ce coefficient nécessite la connaissance des distances entre points. Nous y reviendrons plus loin dans ce rapport.

1.2.2 L'indice de Rand

On souhaite comparer une partition $C = C_1 \dots C_K$ à une autre partition $P = P_1 \dots P_{K'}$ d'un ensemble d'observations Z. Cette dernière peut être soit la vraie partition des données (dans le cas des données étiquetées), soit construite par classification sur l'ensemble des données. Soit (z_v, z_u) une paire d'observations de Z on désigne :

- a : Le nombre de paires appartenant au même cluster dans C et au même groupe dans P.
- b : Le nombre de paires appartenant au même cluster dans C et à des groupes différents dans P.
- c : Le nombre de paires appartenant à des clusters différents dans C et au même groupe dans P.
- d : Le nombre de paires appartenant à des clusters différents dans C et à des groupes différents dans P.

Remarquons $a + b + c + d = M$ qui est le nombre maximum de toutes les paires dans l'ensemble des données. $M = \frac{N(N-1)}{2}$ (où N est le nombre total d'observations).

1.2.2.1 L'indice de Rand

L'indice de Rand est donné par :

$$R = \frac{a+d}{M} \tag{1.4}$$

Intuitivement $a + d$ peuvent être considérés comme le nombre d'accords entre C et P. La valeur de cet indice est entre 0 et 1. Une valeur égale à 0 indique que les deux partitions sont en désaccord pour toutes les paires des données alors qu'une valeur égale à 1 indique que les classes identifiées par les deux partitions sont identiques.

1.2.2.2 L'indice de Rand ajusté

En pratique, l'indice de Rand ne fournit jamais une valeur de 0. En effet, même une partition aléatoire a un indice positif. Une forme normalisée de cet indice a été proposée. Elle est définie par :

$$R_{adj} = \frac{R - E(R)}{max(R) - E(R)} \tag{1.5}$$

où $E(R)$ est l'espérance de l'indice de Rand et $max(R)$ la valeur maximale qu'il peut atteindre.

Chapitre 2

Consensus de partitions

Introduction

Ce chapitre a pour but de détailler la méthodologie de résolution du problème de combinaison de plusieurs partitions d'un ensemble d'objets (ou d'individus) en une seule partition. Ce problème est connu sous le nom de consensus de partitions ou agrégation de partitions. Il consiste à identifier une partition compromis d'un ensemble de partitions obtenues sur le même ensemble d'observations [11, 12] . L'idée principale des méthodes de consensus est d'agglomérer les partitions séparées (appelées partitions contributives), obtenues à partir de chaque tableau en une partition globale. Cette dernière partition doit être la plus similaire aux partitions contributives selon un certain indice, par exemple l'indice de Rand. D'autres méthodes relativement plus récentes dans l'approche de "ensemble cluster" [13] ont également abordé le problème de consensus de partitions. Parmi elles, nous avons la méthode CSPA (Cluster based Similarity Partitioning Algorithm) qui est basée sur une matrice dite d'association ou de connectivité dont la forme sera définie explicitement dans la suite.

Les auteurs de l'étude (L. Tao, C. Ding (2008)) ont souligné que la CSPA donne des résultats instables lorsque les partitions contributives sont significativement différentes. De plus, si certaines des partitions d'entrée sont fortement corrélées, cette redondance pourrait biaiser la partition finale vers ces partitions corrélées. Ils ont proposé alors une approche basée sur la factorisation de matrices non-négatives dite NMF [14].

Les techniques de consensus citées jusque là ont la particularité d'accorder le même poids à toutes partitions contributives, on parle alors de consensus partitions non-pondéré ou simple. En pratique, certaines partitions peuvent paraître particulières par rapport aux autres. Dans ce cas, accorder le même poids à toutes les partitions peut biaiser la partition compromis obtenue. C'est dans ce sens que des auteurs ont proposé des techniques de consensus pondéré de partitions. Autrement dit, les partitions seront pondérées en fonction de leur particularité. Parmi celles-ci, on peut citer l'extension de la méthode NMF dite "Weighted Nonnegative Matrix Factorization" (WNMF) [11]. Cette approche permet la recherche simultanée de la partition compromis et les poids associés aux partitions contributives. Nous avons également une méthode de consensus pondéré (dénommée RVCONS) basée sur le coefficient de corrélation RV [15] . Celle-ci a été proposée pour obtenir une matrice de connectivité agrégée qui est ensuite utilisée pour classifier les individus afin de trouver la partition de consensus [16]

Dans ce qui suit, nous décrivons la méthode NMF et sa version pondérée WNMF. Ces deux approches sont des techniques classiques pour déterminer le consensus de partitions. Ainsi, nous les utiliserons dans ce travail pour des raisons de comparabilité.

2.1 Formulation

Soit $P = \{P_1, P_1, P_2, ..., P_T\}$ un ensemble de T partitions différentes sur un ensemble I de n individus, tel que $I = \{O_1, O_1, O_2, ..., O_n\}$. Ces partitions sont les résultats des multiples partitionnements sur le même ensemble d'individus : plusieurs applications (initialisations) d'un même algorithme de classification, différents algorithmes sur le même jeu de données ou même algorithme sur différents jeux de données décrivant les individus à travers différents ensembles de variables.

Pour chaque partition P_t, on définit une matrice de connectivité ou d'adjacence $M(P_t)$ et une matrice d'indicatrices $H(P_t)$. La matrice de connectivité est une matrice $(n \times n)$ qui contient 1 si les deux individus i et j se trouvent dans la même classe, 0 sinon.
Exemple :

	O_1	O_2	O_3	O_4
O_1	1	0	0	1
O_2	0	1	1	0
O_3	0	1	1	0
O_4	1	0	0	1

TABLE 2.1 – Matrice de connectivité M

Dans cet exemple, l'ensemble des individus est donné par O= $\{O_1, O_2, O_3, O_4\}$. Les individus 1 et 4 sont regroupés ensemble, de même pour les individus 2 et 3.
La matrice d'indicatrices $H(P_t)$ est une matrice (n×K) telle que K est le nombre de classes $(c_1^t, c_2^t, \ldots, c_K^t)$ de la partition P_t. Chaque ligne i de H correspond à un individu et contient un seul 1 dans la colonne K tel que K est la classe de i, les autres cases de la ligne i contiennent 0.

Exemple :

	$cluster_1$	$cluster_2$
O_1	1	0
O_2	0	1
O_3	0	1
O_4	1	0

TABLE 2.2 – Matrice d'indicatrices H

La matrice de connectivité $M(P_t)$ est obtenue par $M(P_t) = H(P_t)H(P_t)'$ où $H(P_t)'$ désigne la transposée de $H(P_t)$. On remarque qu'il s'agit d'une matrice symétrique positive contenant que des 1 en diagonale.
Remarque :
Le nombre K de clusters peut être différent d'une partition à une autre.

On définit la matrice d'association $\tilde{M}$ qui est une simple moyenne des matrices de connectivité par :

$$\tilde{M}_{ij} = \sum_{t=1}^{T} w_t M_{ij}(P_t) \tag{2.1}$$

Elle représente l'association moyenne entre deux observations (i, j). Dans le cas d'un consensus simple, les poids sont donnés par $w_t = \frac{1}{T}$ pour toutes les partitions contributives. C'est le cas de la méthode NMF. Par contre, l'approche pondérée repose sur la détermination et la recherche des poids w_t, en fonction de la qualité et la particularité de chaque partition. La matrice d'association s'exprime donc comme une moyenne pondérée des matrices de connectivité en fonction de ces poids. C'est le cas des méthodes WNMF [11]

2.2 Consensus simple

Une méthode pour trouver le consensus a été décrite dans [14]. Cette méthode se base sur la programmation quadratique et utilise le principe de factorisation d'une matrice non négative NMF. Le consensus de partitions prend la forme du problème d'optimisation suivant :

$$\min_{U} \sum_{i,j}^{n} (\tilde{M}_{ij} - U_{ij})^2 = \min_{U} \|\tilde{M} - U\|^2 \tag{2.2}$$

Les contraintes imposées sur la matrice U (cf. [14]) emmène à devoir résoudre un problème d'optimisation qui satisfait des contraintes d'ordre n^3, chose qui est difficile.
Cependant, on peut imposer les contraintes d'une manière différente.
Supposons une matrice indicatrice $H = \{0, 1\}^{(nxk)}$ avec la contrainte que dans chaque ligne, il y a seulement un seul 1 et les autres éléments sont nuls.
Donc on peut remplacer la matrice U par $U = HH^{'}$. Le problème de consensus de partitions sera donc :

$$\min_{H} \|\tilde{M} - HH^{'}\|^2 \tag{2.3}$$

La contrainte qui impose que dans chaque ligne de H, il y a un seul élément non nul peut être exprimée de la manière suivante : $(H^{'}H)_{kl} = 0$ si $k \neq l$ et $(H^{'}H)_{kl} = n_k$ qui est le nombre d'observations dans la classe k.
Soit : $D = diag(H^{'}H) = (n_1, \ldots, n_k)$. Nous pouvons alors écrire le problème d'optimisation comme suit :

$$\min_{H^{'}H=D, H\geq 0} \|\tilde{M} - HH^{'}\|^2 \tag{2.4}$$

Cependant dans ce problème nous avons besoin de spécifier la taille de chaque classe. Nous avons donc besoin d'éliminer D du problème d'optimisation. Pour cela on suppose $\tilde{H} = H(H^{'}H)^{-1/2}$.

Donc $HH^{'} = \tilde{H}D\tilde{H}^2$, $\tilde{H}\tilde{H}' = H(H'H)^{-1/2}H = I$ et le consensus devient l'optimisation :

$$\min_{\tilde{H}'\tilde{H}=I,\tilde{H}\geq 0} \|\tilde{M} - \tilde{H}D\tilde{H}^{'}\|^2 \tag{2.5}$$

Ainsi, les deux matrices D et $\tilde{H}$ sont obtenues comme solution du problème d'optimisation et nous n'avons pas besoin de spécifier la taille de chaque classe. La solution à ce problème peut être donnée par tri-NMF, il s'agit d'un cas particulier où la matrice que nous voulons factoriser est symétrique. Dans ce cas la mise à jour de $\tilde{H}$ et D se font de la manière suivante :

$$\tilde{H}_{tk} \leftarrow \tilde{H}_{tk}\sqrt{\frac{(\tilde{M}'\tilde{H}D)_{tk}}{(\tilde{H}\tilde{H}'\tilde{M}'\tilde{H}D)_{tk}}}\tilde{H}_{tk} \leftarrow \tilde{H}_{tk}\sqrt{\frac{(\tilde{M}'\tilde{H}D)_{tk}}{(\tilde{H}\tilde{H}'\tilde{M}'\tilde{H}D)_{tk}}} \tag{2.6}$$

$$D_{tk} \leftarrow D_{tk}\sqrt{\frac{(\tilde{H}'\tilde{M}\tilde{H})_{ik}}{(\tilde{H}'\tilde{H}D\tilde{H}'\tilde{H})_{tk}}} \tag{2.7}$$

Cependant lorsque l'ensemble des partitions diffèrent largement, le consensus obtenu sur la base d'une moyenne simple des partitions n'est pas pertinent. D'un autre côté, s'il existe un sous-ensemble de partitions qui sont corrélées entre elles, cela peut pencher la solution finale vers ces partitions corrélées comme souligné dans 2. Ce problème peut être résolu par l'introduction d'un système de poids associés à chaque partition.

2.3 Consensus pondéré : Weighted NMF

T. Li et C. Ding ont proposé le « weighted » consensus [11]. L'idée fondamentale réside dans la définition de la matrice d'association $\tilde{M}$. Au lieu d'utiliser la moyenne simple, on introduit un poids propre à chaque partition tel que :

$$w = (w_1, w_2, \ldots, w_T)^{'}, \;\; w_t \geq 0, \;\; \|w\|_1 = \sum_{t=1}^{T} w_t = 1 \tag{2.8}$$

La définition de la matrice d'association sera donc :

$$\tilde{M} = \sum_{t=1}^{T} w_t M(P_t) \tag{2.9}$$

Le problème de consensus pondéré sera :

$$\min_{w,\tilde{H}} \|\tilde{M} - \tilde{H}\tilde{H}^{'}\|^2 \tag{2.10}$$

$$\|\tilde{M} - \tilde{H}\tilde{H}^{'}\|^2 = tr(\tilde{M}\tilde{M} - 2\tilde{H}'\tilde{M}\tilde{H} + \tilde{H}\tilde{H}^{'}\tilde{H}\tilde{H}^{'}) \tag{2.11}$$

Où tr est la trace de la matrice. Le problème de consensus pondéré devient (2.12), puisque

le premier et le troisième terme sont constants.

$$\max_{w,\tilde{H}} tr(\tilde{H}^{'}\tilde{M}\tilde{H}) \tag{2.12}$$

Le problème d'optimisation consiste à itérer les deux étapes suivantes :

1. optimise le critère sur $\tilde{H}$ à w en utilisant (2.6)
2. optimise w en fixant $\tilde{H}$. A noter que

$$J = tr[(\tilde{M} - \tilde{H}\tilde{H}^{'})(\tilde{M} - \tilde{H}\tilde{H}^{'})] = tr(\tilde{M}\tilde{M} - 2\tilde{H}^{'}\tilde{M}\tilde{H} + \tilde{H}\tilde{H}^{'}\tilde{H}\tilde{H}^{'})$$

que :

$$tr(2\tilde{H}^{'}\tilde{M}\tilde{H}) = b^{'}w \;\; où \;\; b_t = \tilde{H}'M(P_t)\tilde{H}$$

et :

$$\tilde{M}^2 = w_1^2M(P_1)^2 + ... + w_T^2M(P_T)^2 + 2w_1w_2M(P_1)M(P_2)... = w'Aw$$

tel que :

$$A_{tt'} = tr[M(P_t)M(P_{t'})]$$

Donc en fixant $\tilde{H}$, le problème d'optimisation devient :

$$\min w^{'}Aw - 2b^{'} + cte$$

Il s'agit d'un problème d'optimisation quadratique à contraintes linéaires à T variables. Le fait que A soit une matrice semi- définie positive, rend le problème d'optimisation convexe et nous obtenons la solution globale ce qui peut être résolu en utilisant la programmation quadratique standard.
Ce problème d'optimisation minimise automatiquement la redondance dans les partitions. En effet, parce qu'on minimise le terme $w^{'}Aw$, il est raisonnable de s'attendre à ce qu'un grand $A_{tt'}$ mènera à une faible valeur de $w_tw_{t'}$ ou l'un d'eux.
Donc si deux partitions sont similaires, un des poids correspondants dans la solution finale tend à être faible.

A notre connaissance, il n'existe pas une implémentation sous R de l'algorithme WNMF. L'une des tâches effectuées durant ce stage consistait alors à mettre en place une implémentation de cet algorithme sous R.

Remarques

En l'absence d'une vérité de terrain ou "partition théorique", le seul moyen d'évaluer la qualité de la partition obtenue reste les indices de qualité internes. Après un consensus de partitions sur plusieurs tableaux de données, il est impossible de calculer un indice de qualité interne. En effet, comme évoqué à la section 1.2.1 le calcul de ces indices nécessite

une matrice de distance obtenue à partir du tableau de données sur lequel la partition a été obtenue. Or, une partition consensus obtenue à partir de plusieurs tableaux et pas à partir d'un seul tableau n'a donc pas une unique matrice de distances. En effet, nous aurons autant de matrices de distances que de tableaux et aucune d'entre elles ne peut être choisie pour évaluer la partition consensus. Ce même problème sera rencontré lors de l'imputation multiple faisant appel au consensus de partitions à partir de tableaux imputés. Compte tenu de ces contraintes, évaluer la qualité d'une partition par des critères interne après imputation multiple devient problématique en pratique.

Par ailleurs, la stratégie classique basée sur ces indices de qualité internes pour choisir le nombre de classes optimal ne peut plus être utilisée. Pour palier à cette limite, nous avons proposé une méthode adaptée pour le calcul du coefficient de silhouette pour une partition consensus (cf. 5.3.3).

Chapitre 3

Imputation multiple

Introduction

Dans ce chapitre, nous présenterons la méthodologie de l'imputation multiple. Nous commencerons par présenter l'imputation multiple lorsque les données sont structurées en groupes. Ensuite, nous présenterons la règle de Rubin ainsi que son extension en classification.

Le principe de l'imputation consiste à « retrouver » approximativement la donnée manquante [2] [3]. Comme nous l'avons déjà cité plus haut, les méthodes d'imputation peuvent être regroupées en deux groupes : l'imputation simple et l'imputation multiple. Cependant, quelle que soit l'approche utilisée, une connaissance des propriétés de ces données manquantes notamment le mécanisme noté R qui pourrait être derrière est primordiale. En effet, il s'agit de comprendre la cause de la présence de ces données manquantes : si cette présence est en rapport avec les variables de la base, c'est-à-dire les autres questions dans le cadre d'une enquête par exemple, ou bien si elle est due tout simplement au hasard. Ainsi, ces mécanismes ou processus de génération des données manquantes peuvent être regroupés en trois groupes : MCAR, MAR et MNAR (Rubin, 1976).

- MCAR (Missing completly at random) : les données générées complètement au hasard. C'est lorsque la probabilité d'être manquante est la même pour toutes les cases pour une variable donnée. De surcroît, elle ne dépend pas des données elles-mêmes. Une erreur de saisie correspond à ce mécanisme car l'erreur ne dépend pas des données, elle est complètement indépendante.

- MAR (Missing at Random) : les données générées au hasard. C'est lorsque la probabilité d'être manquante dépend des données observées. Par exemple, l'appareil de mesure (baromètre par exemple) qui tombe en panne quand la température est élevée. Disposant de la température, on peut alors connaître la probabilité que la valeur de pression soit manquante.

- MNAR (Missing Not at Random) : les données non générées au hasard. Dans ce cadre de figure, la probabilité d'être manquante est liée à la partie non observée des données. Par exemple, lors d'un sondage, un individu fortuné aura plutôt tendance à ne pas répondre à une question portant sur ses revenus.

Ces mécanismes, notamment MAR et MCAR, sont souvent identifiés au moyen d'une analyse exploratoire en amont même si cette dernière ne permet pas de trancher entre les différents mécanismes. L'identification du mécanisme MNAR nécessitant une modélisation plus complexe que les deux autres ; on fait généralement l'hypothèse que le mécanisme est MAR par

défaut et on teste a posteriori la robustesse de l'analyse à la violation de cette hypothèse via une analyse dite de sensibilité.

Comme présenté plus haut, certains défauts de l'imputation simple, notamment du fait qu'on arrive pas à distinguer les valeurs imputées des valeurs observées, font qu'il est préférable de faire appel aux méthodes d'imputation multiple. Le principe de cette technique consiste à imputer plusieurs fois. Cela permet ainsi de disposer de plusieurs tableaux imputés permettant de refléter l'incertitude sur les données imputées. La finalité de cette approche est d'appliquer une méthode d'analyse sur un tableau de données avec des manquantes. Ainsi, sur chacun des tableaux complets, sera appliquée une méthode d'analyse souhaitée. La dernière phase consistera à agréger l'ensemble des résultats d'analyse des différents tableaux selon des règle bien spécifiques dites règles de Rubin [4].

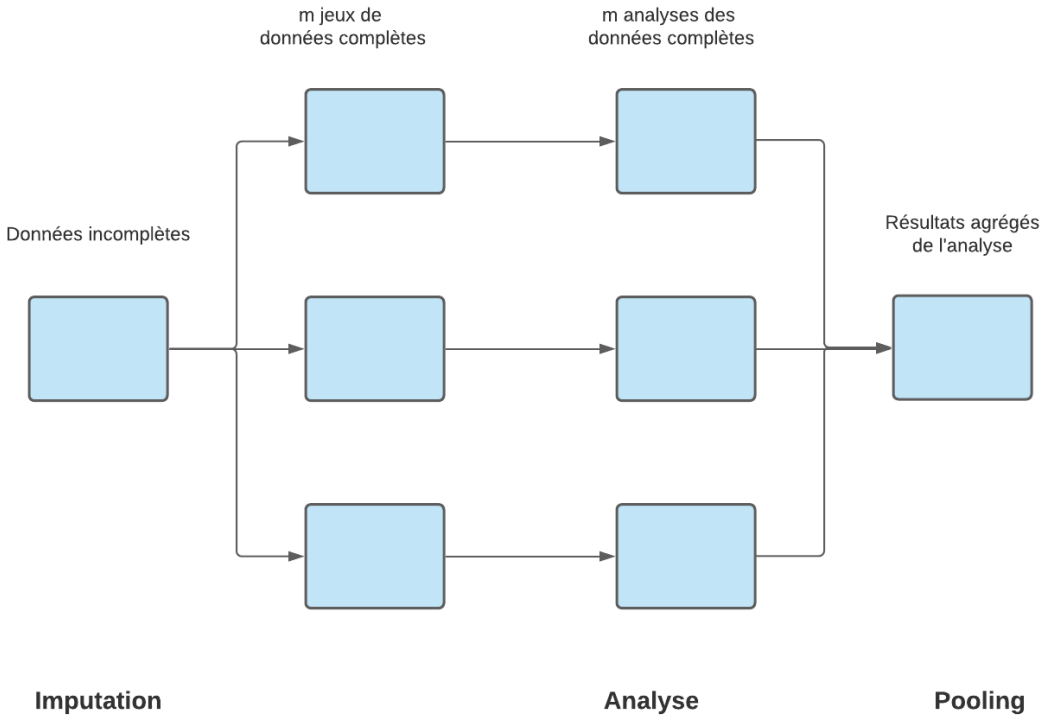


FIGURE 3.1 – Schéma d'une imputation multiple

Les règles de Rubin ont été étendues en clustering [5] ; c'est-à-dire lorsque l'objectif de l'analyse est de trouver une classification commune des m tableaux imputés (m correspond ici au nombre de tableaux imputés), mais aussi d'évaluer l'instabilité liée à l'imputation.

Plusieurs approches d'imputation multiple ont été développées. Pour toutes ces approches, le principe reste le même 3.2. Parmi celles-ci, nous avons la méthode JM-DP qui permet de tenir en compte la présence d'une structure en groupe.

3.1 Imputation multiple par la méthode JM-DP

JM-DP est une approche d'imputation multiple qui utilise une extension non-paramétrique du modèle de mélange, à savoir le mélange de processus de Dirichlet [17] de produits de distributions normales multivariées. Nous commençons par décrire le processus de mélange de Dirichlet, ensuite son extension dans le processus d'imputation multiple.

3.1.1 Modèle de mélanges de lois normales par processus de Dirichlet

Soit $Z = \{Z_1, ..., Z_n\}$ n observations complètes normalisées où Z_i est un vecteur de dimension p. On suppose que chaque individu i appartient à une et à une seule des K composantes latente du modèle de mélanges. On définit pour chaque individu $i \in \{1, ..., n\}$, $z_i \in \{1, ..., K\}$ la composante de l'individu i et $\pi_k = P(z_i = k)$. Deux configurations sont envisageables. Dans le premier cas, on suppose que les π_k sont les mêmes pour tous les individus. Pour une composante k, on considère que les p variables suivent une composante de la loi normale multivariée de moyenne μ_k et de variance Σ_k.
Considérons $\Theta(\mu, \Sigma, \pi)$ l'ensemble des paramètres, avec $\mu = \{\mu_1, ..., \mu_K\}$ et $\Sigma = \{\Sigma_1, ..., \Sigma_K\}$. Le modèle de mélange peut, alors être exprimé comme suit :

$$Z_i | z_i, \mu, \Sigma \sim N(Z_i | \mu_{z_i}, \Sigma_{z_i}) \tag{3.1}$$

$$z_i | \pi \sim Multinomial(\pi_1, ..., \pi_K) \tag{3.2}$$

Ainsi, la distribution marginale sur z_i est donnée par :

$$p(Z_i | \Theta) = \sum_{k=1}^{K} \pi_k N(Z_i | \mu_k \, \Sigma_k) \tag{3.3}$$

La spécification complète Bayésienne nécessite la connaissance de la distribution a priori de Θ. Selon [18], le couple de distribution (μ_k, Σ) est donné par :

$$\mu_k \mid \Sigma_k \sim N(\mu_0, h^{-1}\Sigma_k) \tag{3.4}$$

$$\Sigma \sim W^{-1}(f, \Phi) \tag{3.5}$$

$h > 0$ est un paramètre de précision, f étant le degré de liberté de $\Phi = diag(\phi_1, ..., \phi_p)$ et

$$\phi_j \sim \Gamma(a_\phi, b_\phi) \tag{3.6}$$

de moyenne $\frac{a_\phi}{b_\phi}$ pour $j = 1, ..., p$.

Le second cas consiste à définir π de la façon suivante :

$$\pi = v_k \prod_{g<k} (1 - v_g) \quad k = 1, ..., K \tag{3.7}$$

$$v_k \sim \beta(1, \alpha) \;\; k = 1, ..., K; \; v_K = 1 \tag{3.8}$$

$$\alpha \sim \Gamma(a_\alpha, b_\alpha) \tag{3.9}$$

Dans cette approche, on peut constater que les π_i ne sont plus uniformes. Ils sont plutôt décroissants permettant ainsi de gérer les problème lié à un grand nombre de composantes. En effet, selon cette configuration lorsque K est grand, les dernières composantes ont une probabilité d'appartenance très faible, voire nulle. (a_α, b_α) étant des constantes.

3.1.2 L'extension du modèle de Dirichlet dans le processus d'imputation multiple

Le modèle décrit plus haut est dit de Dirichlet. L'imputation multiple sous ce modèle nécessite de pouvoir en déterminer les paramètres, alors que les données sont incomplètes. Dans un cadre bayésien, on utilise classiquement un algorithme de data-augmentation. Le principe est d'alterner le tirage des paramètres dans leur distribution a posteriori et l'imputation des données selon les paramètres simulés. Ces étapes sont répétées jusqu'à convergence vers la distribution a posteriori des paramètres $\Theta(\mu, \Sigma, \pi)$. Pour effectuer l'imputation multiple selon cet algorithme, on répète à nouveau ces étapes après convergence de façon à récupérer m réalisations indépendantes des paramètres issus de la loi a posteriori. C'est à partir de ces m paramètres que l'on pourra imputer m fois le tableau incomplet.

De ce fait, l'imputation se fait de façon successive entre les tableaux de telle sorte que le tableau $t + 1$ est dépendant du tableau t (t correspond ici au nombre d'itérations).
Exemple :

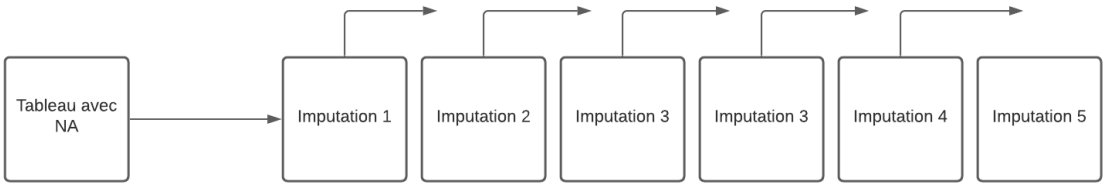


FIGURE 3.2 – Illustration du pas entre tableaux imputés

Dans cet exemple, 5 tableaux de données ont été obtenus successivement par imputation. Les paramètres qui ont servi à imputer le tableau 4 ont été obtenus à partir du tableau imputé 3 par exemple. Ainsi, le tableau 4 est dépendant du tableau 3. Ces deux tableaux sont séparés par un pas de 1. Le tableau imputé 5 et le tableau imputé 1 sont séparé par un pas de 5.

Dans la phase de l'imputation, il est important de tenir compte de ce pas de succession entre tableaux. Ce pas entre deux tableaux peut se comprendre comme le nombre d'imputations qui séparent deux tableaux dans la succession de l'imputation.

3.2 Règles de Rubin

On note Z l'ensemble des données : Z^{obs} les données observées, Z^{miss} celles manquantes et R le mécanisme lié au données manquantes. Ainsi, $Z = (Z^{obs}, Z^{miss})$. La distribution de la variable aléatoire est notée F.

In fine, l'imputation multiple (IM) vise à estimer la moyenne et la variance d'une statistique Q sur les réalisations de (Z^{obs}, R). Par exemple, Q pourrait être l'estimateur des moindres carrés des coefficients de régression dans un modèle linéaire. Les estimations associées sont obtenues en trois étapes :

1. m tableaux imputés à partir de Z sont imputés suivant la fonction de distribution $F_{Z^{miss}|Z^{obs}}$. Ce qui nous permet de disposer de m tableaux complet $(Z^{obs}, Z_t^{miss})_{1 \leq t \leq m}$.

2. La statistique Q est estimée sur chaque tableau complet; $\hat{Q}_m$ est l'estimation de Q obtenue à partir du tableau imputé m et (U_m) la variance associée.

3. Toutes les estimations sont agrégées selon les règles de Rubin (D. Rubin, 1976) conduisant à une seule estimation de Q notée $\bar{Q}$ et à une seule variance notée S

$$\bar{Q} = \frac{1}{m} \sum_{t=1}^{m} \hat{Q}_t \tag{3.10}$$

$$S = \underbrace{\frac{1}{m} \sum_{t=1}^{m} \hat{U}_t}_{\bar{U}} + \underbrace{\frac{1}{m-1} \sum_{t=1}^{m} (\hat{Q}_t - \bar{Q})^2}_{B} \tag{3.11}$$

Le terme $\bar{U}_t$ correspond à la variance d'un estimateur, ici $\bar{Q}_t$, obtenu à partir du tableau imputé t et B la variance estimée supplémentaire attribuable aux données non observées. Contrairement à la méthode d'imputation simple, l'imputation multiple (IM) permet le calcul de la variabilité liée à l'imputation à travers B.

Initialement, les règles de Rubin n'ont pas été développées pour la classification. En fait, après avoir imputé les données et effectuer la classification sur chaque tableau, on ne pourra pas effectuer une "moyenne" des partitions ou calculer "une variance" tel que proposée aux équations (3.10) et (3.11). Une extension de ces règles en classification a été proposé dans [5]. Cette extension, basée sur les techniques de consensus de partitions, fait l'objet de la section suivante.

3.2.1 Consensus de partitions après imputation multiple

Pour adapter les règles de Rubin en classification, V. Audigier, N. Niang [5] ont utilisé les techniques de consensus de partitions pour avoir l'équivalent "partition moyenne". Le consensus de partitions après IM vise à agréger plusieurs partitions variant uniquement par les valeurs imputées. Formellement, nous gardons la même formulation de la problématique

2.2 qui correspond à l'équivalent de la première règle de Rubin (3.10) en classification après IM. Ainsi, l'extension de ces règles nous permet de trouver la partition compromis après IM, mais aussi de calculer l'instabilité attribuable aux données manquantes.

3.2.2 Calcul de l'instabilité

En se basant sur la littérature existante, V. Audigier, N. Niang [5] proposent une équivalence de la règle 2 de Rubin dans le cadre non-supervisé. Comme préciser dans ce qui précède, ces extensions nous permettent également de calculer l'instabilité liée aux données imputées.

3.2.2.1 Calcul de l'instabilité dans le cadre de données complètes

L'évaluation de l'instabilité à l'étape 3 est importante dans la phase d'analyse. Pour atteindre cet objectif, les méthodes de rééchantillonnage sont intéressantes, surtout lorsque l'algorithme de classification appartient à la famille des méthodes géométriques comme les k-means, les k-medoids ou la classification hiérarchique qui se basent sur des calculs de distances. Des auteurs [19] et [20] ont proposé l'approche bootstrap pour mesurer l'instabilité à partir de n'importe quel algorithme de classification.

L'idée principale consiste à définir une distance δ entre les classes puis, à évaluer la distance δ à partir de plusieurs échantillons bootstrap et enfin à agréger toutes les distances en faisant la moyenne.
Plus précisément, la distance entre deux partitions P_t et $P_{t'}$

$$\delta_{F_Z}(P_t, P_{t'}) = \mathbb{P}_{F_Z}\{I(V_{Pt}(Z) = V_{Pt}(Z')) + I(V_{Pt'}(Z) = V_{Pt'}(Z')) = 1\} \tag{3.12}$$

où Z et Z' sont indépendamment tirés de la distribution F_Z et $V_{P_t}(Z)$ est la tesselation de Voronoï [21] pour Z selon la partition donnée par P_t . Cette distance mesure la probabilité de dissemblance entre les deux partitionnements. Sur la base de cette définition, l'instabilité de **P** est définie comme l'espérance des distances entre les partitions formées sur tous les échantillons aléatoires de taille n.

$$\mathbb{E}_{Z^n \sim F_Z^n, \hat{Z}^n \sim F_Z^n}[\delta_{F_Z}(P(Z^n), P(\hat{Z}^n)] \tag{3.13}$$

3.2.2.2 Dans le cadre de données incomplètes

Le calcul de l'instabilité en présence de données manquantes peut se faire en considérant l'espérance définie précédemment (3.13) sur les données observées. Selon la deuxième règle de Rubin, une telle instabilité peut être décomposée comme la somme d'une instabilité intra et d'une instabilité inter. Selon l'équation 3.11, l'instabilité intra peut être estimé à partir des

m tableaux imputés $(Z^{obs}, Z_t^{miss})_{1 \leq t \leq m}$. Elle est donnée par

$$\bar{U} = \frac{1}{m} \sum_{t=1}^{m} U_t^{boot} \tag{3.14}$$

où U_m^{boot} est l'instabilité calculée à partir de (Z^{obs}, Z_m^{miss}) [5] et l'instabilité intra est donnée par

$$B = \frac{1}{m^2} \sum_{t=1}^{m} \sum_{t'=1}^{m} \delta_{\hat{F}_{Z|Z^{obs}}}(P(Z^{obs}, Z_t^{miss}), P(Z^{obs}, Z_{t'}^{miss})) \tag{3.15}$$

Ainsi, l'instabilité totale est donnée par $S = \bar{U} + B$.

Moins la partition est stable, plus S sera grand. Une grande valeur de l'instabilité intermédiaire par rapport à l'instabilité totale indique une forte dépendance de la partition à l'imputation. Ainsi, pour le nombre de classes optimal, nous pouvons nous référer aux partitions fournissant la plus faible valeur de S.

Deuxième partie

Expérimentations et résultats

Chapitre 4

Apport du consensus pondéré en présence de données manquantes

Introduction

Dans ce chapitre, il s'agira d'illustrer dans un premier temps l'influence du pas en régression comme mentionné dans la littérature. Ensuite, notre première contribution consistera à évaluer dans le cadre de la classification non supervisée l'hypothèse émise en régression après imputation multiple. Plus précisément, on s'attend à ce que l'indépendance des m tableaux imputés (conditionnellement aux valeurs observées) conduise à m partitions des individus qui rendent compte de la diversité des partitions plausibles au vu des données observées. L'agrégation de ces partitions par le consensus doit ainsi fournir une estimation non-biaisée de la partition théorique. A l'inverse si les tableaux imputés ne sont pas indépendants, on s'attend à l'introduction d'un biais lié à la non-représentativité des m partitions issues des m tableaux imputés. Dans cette situation, on se demande si un consensus pondéré peut permettre de redresser l'échantillon des m partitions et ainsi améliorer les performances du consensus. L'objectif final est d'évaluer l'apport du consensus pondéré par rapport au consensus simple, après IM, compte tenu de l'indépendance entre les tableaux imputés.

4.1 Illustration sur l'influence du pas δ en régression

Dans cette illustration, nous montrons l'importance de la prise en compte du pas qui sépare les tableaux au cours du processus de l'IM lorsqu'on fait une régression après IM. A cet effet, nous présentons les expérimentions faites pour illustrer l'influence de ce pas en régression à travers une étude de simulation.

4.1.1 Génération des données

Au total, 500 tableaux de données ont été générés. Un tableau de données, initialement complet de p=8 variables et n=250 observations a été généré suivant un modèle Gaussien. Les paramètres de la loi μ et Σ ont été choisis de sorte à avoir des variables faiblement corrélées. Les paramètres considérés sont : μ=(0, 0, 0, 0, 1.5, 1.5, 0, 2.25) et la matrice

$$\Sigma = \begin{pmatrix} I_4 & 0 & & & \\ & 1 & 0.5 & 0.5 & 0.5 \\ 0 & 0.5 & 1 & 0.5 & 0.5 \\ & 0.5 & 0.5 & 1 & 0.5 \\ & 0.5 & 0.5 & 0.5 & 1 \\ & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}$$

Pour un tableau complet de données , nous avons introduit $\tau = 40\%$ de données manquantes suivant la distribution $Prob(r_{il} = 0) = \tau$ (mécanisme MCAR). Ce τ est jugé suffisant pour voir l'influence des données manquantes dans les simulations.

4.1.2 Plan d'expérimentation

Pour chaque tableau avec des données manquantes, on applique une imputation multiple. Le nombre de tableaux m varie et prend les valeurs {3, 5, 20} et le pas considéré δ={1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 25, 30, ..., 100}. Le processus est le suivant :

1. on considère un tableau parmi les 500 ;
2. phase d'imputation : on applique une imputation de m tableaux avec un pas δ_1 selon l'approche DA ;
3. phase d'analyse : on applique un modèle de régression de la variable 8 par les autres 7 variables de la base sur chacun de ces m tableaux ;
4. phase d'agrégation : on agrège les m modèles afin de disposer le modèle final.
5. On répète 2. 3. et 4. pour tous les pas δ_i de la liste δ.

Ce processus est répété sur tous les 500 tableaux générés. On calcule, ensuite, le taux de couverture, c'est à dire la probabilité d'avoir les coefficients théoriques [22] dans l'intervalle de confiance estimé, puis le biais entre le coefficient théorique et celui estimé.

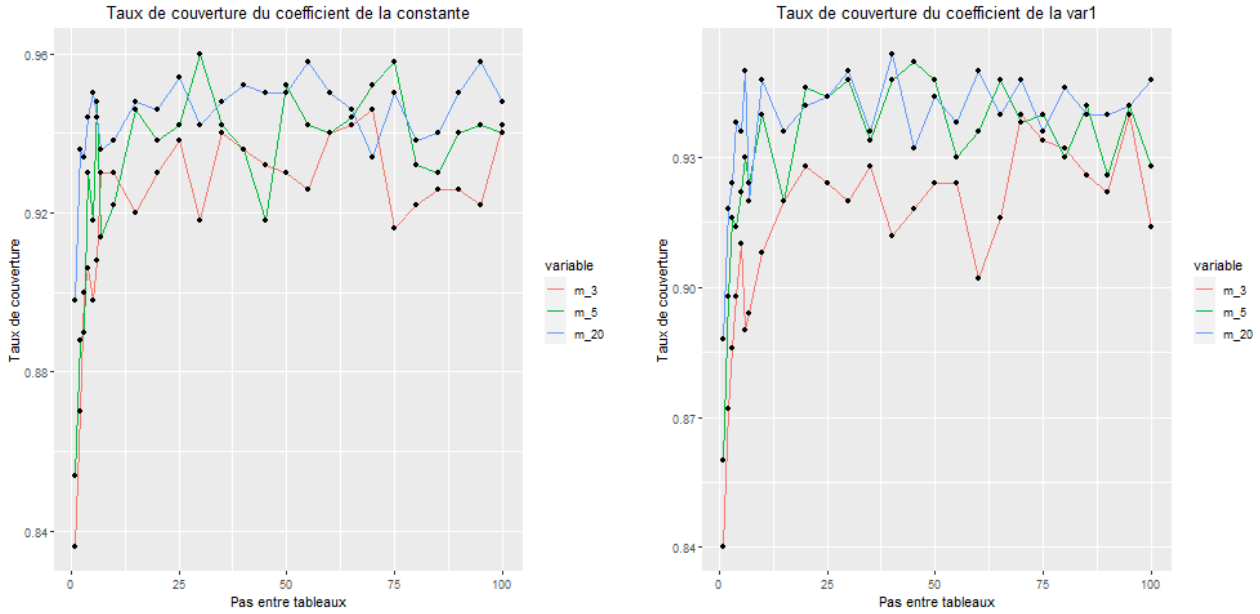
4.1.3 Critères d'évaluation

Dans cette analyse, nous examinons l'influence du pas, qui est un indicateur de l'indépendance entre tableaux. En effet, nous avons utilisé une approche d'imputation multiple DA où les tableaux sont imputés successivement pour tenir compte du pas entre tableaux. Les variables d'un même tableau sont imputées simultanément par un modèle Gaussien. Pour cela, nous examinons, dans un premier temps, le taux de couverture des coefficients de régression obtenus sur les 500 expériences. Cet examen sera fait en fonction du pas et du nombre de tableaux m. Dans un deuxième temps, nous examinerons le biais des coefficients de régression par rapport aux coefficients théoriques. Pour cela, nous allons comparer les erreurs quadratiques moyennes des coefficients de régression par rapport aux coefficients théoriques, en fonction du pas mais aussi du nombre de tableaux considérés.

4.1.4 Résultats

4.1.4.1 Taux de couverture

Les figures 4.1 montre l'évolution du taux de couverture en fonction du pas selon le nombre de tableaux. L'analyse des résultats représentés à travers ces figures montre que nous avons, en général, un plus fort taux de couverture lorsque les tableaux sont indépendants (pas grand). En effet, lorsque nous considérons les courbes d'évolution du taux de couverture en fonction du pas, on peut noter que quelque soit le nombre de tableaux m considéré, le taux de couverture augmente lorsqu'on passe d'un pas de 1 à un pas de 100. Autrement dit, lorsque les tableaux deviennent de plus en plus indépendants, le modèle de régression s'améliore. Cependant, cette évolution n'est pas tout à fait lisse. Cette fluctuation est attribuable à l'aléa lié aux expérimentations. Par ailleurs, nous constatons également que le taux de couverture s'améliore lorsque le nombre de tableaux considéré augmente. On peut voir, en effet, que la courbe des taux de couverture du nombre de tableaux m=20 est au dessus des autres courbes suivie de celle correspondant à un m=5 et enfin celle correspondant à un m=3 est toujours plus basse. Cela peut s'expliquer par le fait que peu de tableaux implique plus de dépendance au pas. Si le pas est assez grand, on peut prendre que 3 tableaux. Si le pas est petit (trop petit pour assurer l'indépendance), il faut plus de tableaux.



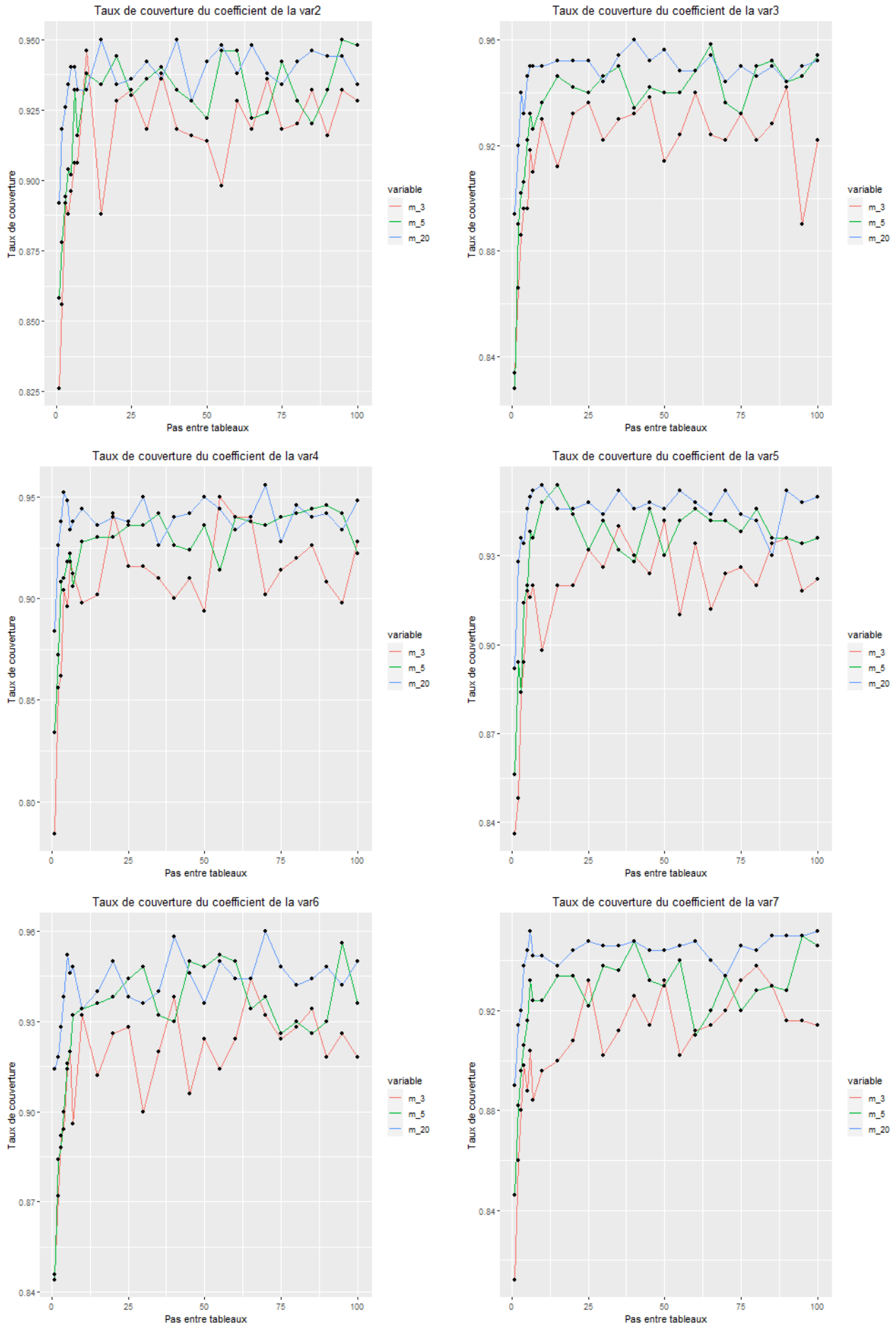
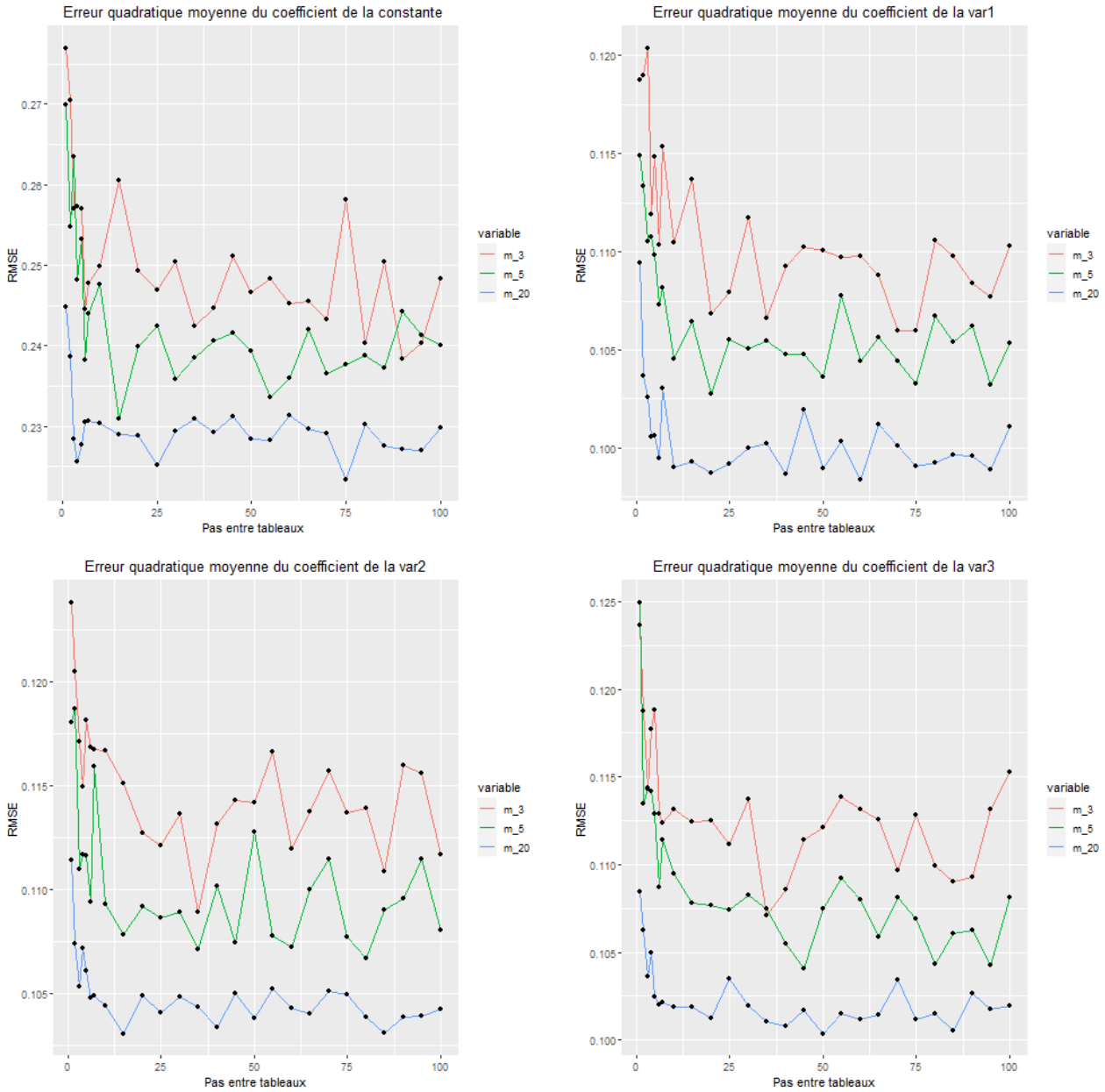


FIGURE 4.1 – Evolution de taux de couverture en fonction du pas et selon le nombre de tableaux m

4.1.4.2 Analyse du biais

Les figures 4.2 montrent l'évolution de l'erreur quadratique moyenne en fonction du pas et selon le nombre de tableaux m. A l'image des résultats obtenus sur l'analyse du taux de couverture, une analyse du biais à travers l'erreur quadratique moyenne montre que cette dernière est de plus en plus faible lorsque les tableaux deviennent de plus en plus indépendants (pas de plus en plus grand). Lorsqu'on observe les courbes de l'EQM obtenue en fonction du pas, on peut noter que lorsque le pas augmente, l'EQM diminue. Ainsi, le modèle estimé s'approche de plus en plus du modèle théorique. Par ailleurs, une augmentation du nombre de tableaux diminue le biais obtenu sur les coefficients de régression.



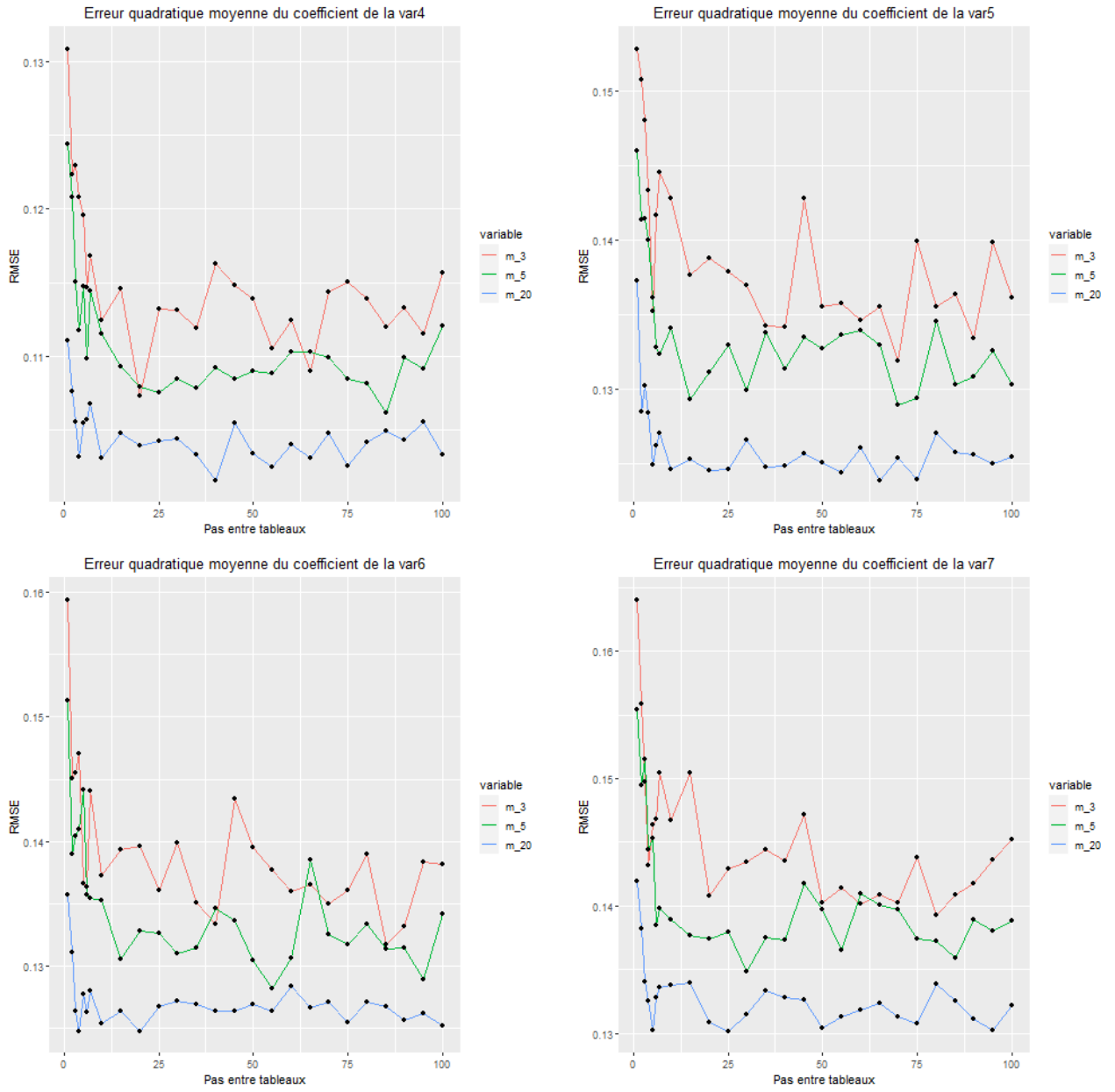


FIGURE 4.2 – Evolution de l'erreur quadratique moyenne en fonction du pas et selon le nombre de tableaux m

Synthèse

De l'analyse des résultats, on a pu noter l'influence du pas et du nombre de tableaux imputés sur la qualité du modèle de régression obtenu. Nous avons noté que lorsque les tableaux imputés sont indépendants, les coefficients de régression théoriques ont plus de chance de se retrouver dans l'intervalle de confiance estimé. De même, un grand nombre de tableaux imputés considérés augmente les chances que les coefficients de régression théoriques se trouvent dans l'intervalle de confiance estimé.

L'analyse du biais sur les coefficients de régression, à travers l'erreur quadratique moyenne, consolide les constats sur l'analyse du taux de couverture. En effet, l'EQM diminue lorsque le pas augmente, c'est à dire les tableaux deviennent de plus en plus indépendants. De même,

un grand nombre de tableaux considérés diminue le biais. Ces résultats s'expliquent par le fait qu'une imputation consiste en un tirage d'une valeur plausible de la valeur manquante. Lorsqu'on fait une imputation multiple en considérant des tableaux qui se ressemblent (tableaux dépendants), on fait des tirages mais on obtient presque la même valeur et par conséquent on reste proche d'un seul tirage qui est loin de la vraie valeur ; d'où un taux de couverture faible et une EQM importante. Par ailleurs, lorsqu'on considère un plus grand nombre de tableaux, on augmente le nombre de tirages et on se rapproche davantage de la vraie valeur ; ce qui entraîne une augmentation du taux de couverture et une diminution de l'erreur quadratique moyenne du biais.

Dans nos simulations, nous allons vérifier ces hypothèses dans le cadre non supervisé de la classification et donc des méthodes de consensus de partitions pour l'agrégation des résultats après IM. De façon plus explicite, nous supposons que lorsque les tableaux imputés sont séparés par un pas δ assez grand, le consensus de partitions sera plus performant au regard de l'indice de Rand ajusté (ARI) présenté plus haut.

4.2 Extension de l'influence du pas δ en classification

4.2.1 Plan de simulation et données

Les données ont été intégralement générées suivant des paramètres bien définis, lesquels paramètres nous détaillerons dans la suite de ce document. Dans nos travaux nous avons fait plusieurs simulations pour évaluer l'influences du pas en classification. Pour cela, nous avons regardé les performances du consensus en fonction du pas, du nombre de tableaux et de l'approche (simple, pondéré). Nous avons également regardé les relations entre tableaux à travers le coefficient RV qui correspond à la corrélation entre tableaux de données (cf. 5.3) et l'écart entre les tableaux imputés en fonction du pas.

Dans la section qui suit, nous présentons une partie de ces simulations et nous mettrons le reste en annexe. Concernant la simulation présentée sous cette section, elle porte sur les performances du consensus en fonction du pas, du nombre de tableaux et selon les deux approches NMF et WNMF.

Pour ce qui est de la présentation du plan de simulation, on décrit une seule expérience. Cette expérience est répétées 30 fois de suite.

4.2.1.1 Génération des données

Pour une expérience donnée, un tableau de données complet a été généré suivant un modèle de mélanges de deux gaussiennes. Le premier suit un p-multivarié gaussien $(p = 10)$ de moyenne nulle et de matrice de variance covariance

$$\Sigma(p) = \begin{pmatrix} I_5 & 0 & & & & \\ & 1 & \rho & \rho & \rho & \rho \\ 0 & \rho & 1 & \rho & \rho & \rho \\ & \rho & \rho & 1 & \rho & \rho \\ & \rho & \rho & \rho & 1 & \rho \\ & \rho & \rho & \rho & \rho & 1 \end{pmatrix}$$

La deuxième composante suit une loi gaussienne multivariée de moyenne μ = (0, 0, 0, 0, 0, 2, 2, 2, 2, 2) et de matrice de variance covariance $\Sigma(p)$. Au final, 100 observations (n=100) ont été générées et décrites par 10 variables. Les observations sont regroupées en deux classes. Étant donné qu'il s'agit d'observations avec une structure en groupe, nous avons redéfini les paramètres de distribution des variables. La corrélation entres ces variables est de $\rho = 0.3$. Pour un tableau de données complet, on introduit 30% de données manquantes suivant la distribution $Prob(r_{il} = 1) = \Phi(a_t + x_{i1})$ où Φ est la fonction de répartition de la loi normale et a_t une constante pour tenir compte du pourcentage des valeurs manquantes dans les estimations (mécanisme MAR).

4.2.1.2 Plan d'expérimentation

L'hypothèse de cette expérimentation est que lorsque les tableaux sont indépendants ou le nombre de tableaux imputés est important, on obtient de meilleures performances au regard de l'indice ajusté de Rand. Pour une expérience donnée, le plan d'expérimentation a été élaboré de la façon suivante, illustré à travers la figure 4.3.

1. On procède à une imputation multiple suivant la méthode JM-DP à l'aide du package clusterMI ; cette approche d'imputation permet de tenir en compte la présence d'une structure en groupe. Ce qui nous fournit 1000 tableaux complets. Dans cette imputation multiple, on considère que plus les tableaux sont proches (pas faible entre les tableaux), plus ils sont dépendants.

2. On considère un groupe de m tableaux imputés (m=20, m=5, m=3) en faisant varier le pas qui les sépare.
   Pour un m et un pas donnés, par exemple m=20 et un pas égal à 1 :

   (a) On choisit les 20 premiers des 1000 tableaux obtenus.
   On cherche la partition sur chacun de ces 20 tableaux et on fait le consensus par la méthode NMF puis par la méthode WNMF.
   On répète le processus en choisissant les 20 suivants et en commençant par le deuxième tableau. Ainsi, de suite jusqu'à parcourir tous les groupes possibles avec un pas de 1.

(b) On répète ce processus pour tous les pas δ_i de la liste :

$$\delta = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 20, 30\}$$

(c) Pour chaque δ_i, on considère la moyenne des indices de Rand ajustés obtenus sur le consensus de partitions obtenus sur les différents groupes de pas δ_i pour chaque approche (NMF et WNMF).

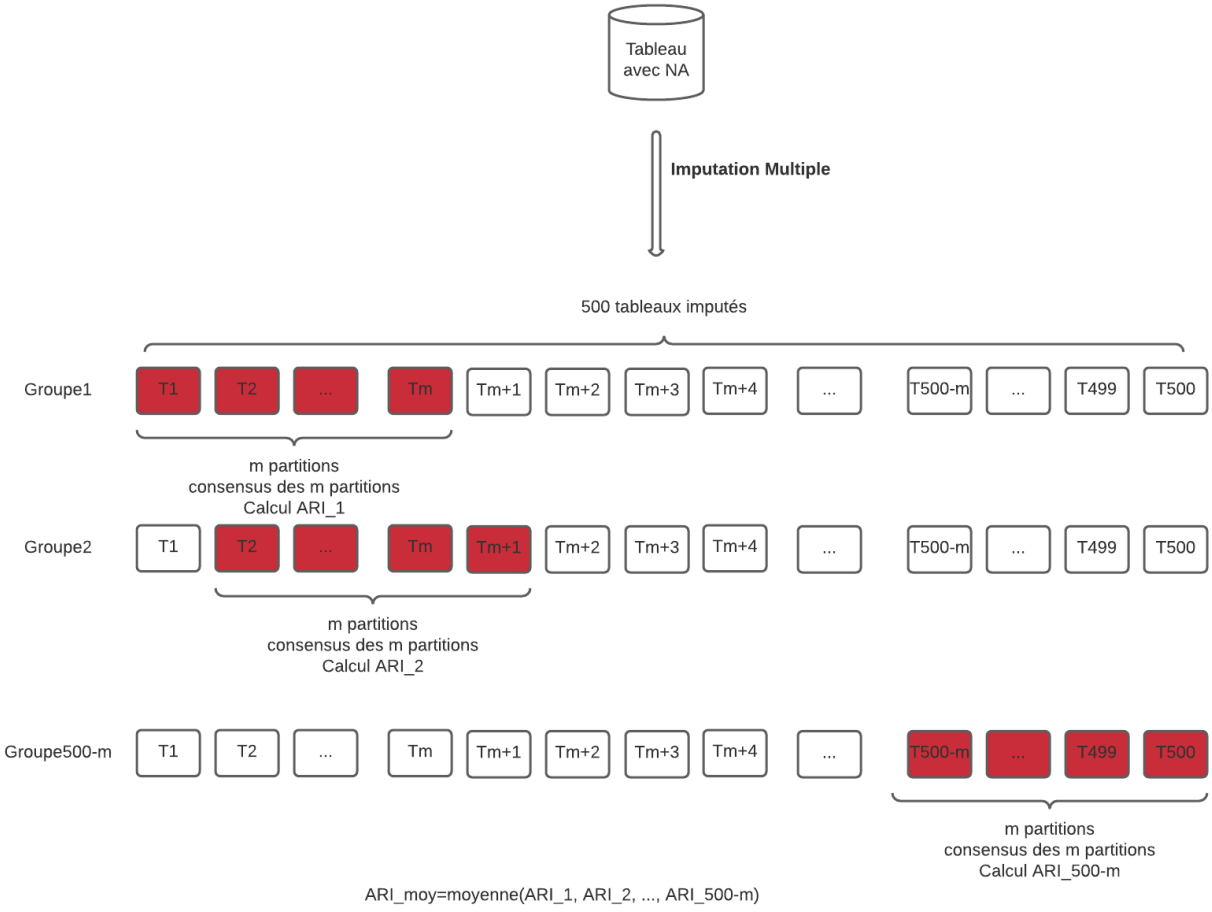


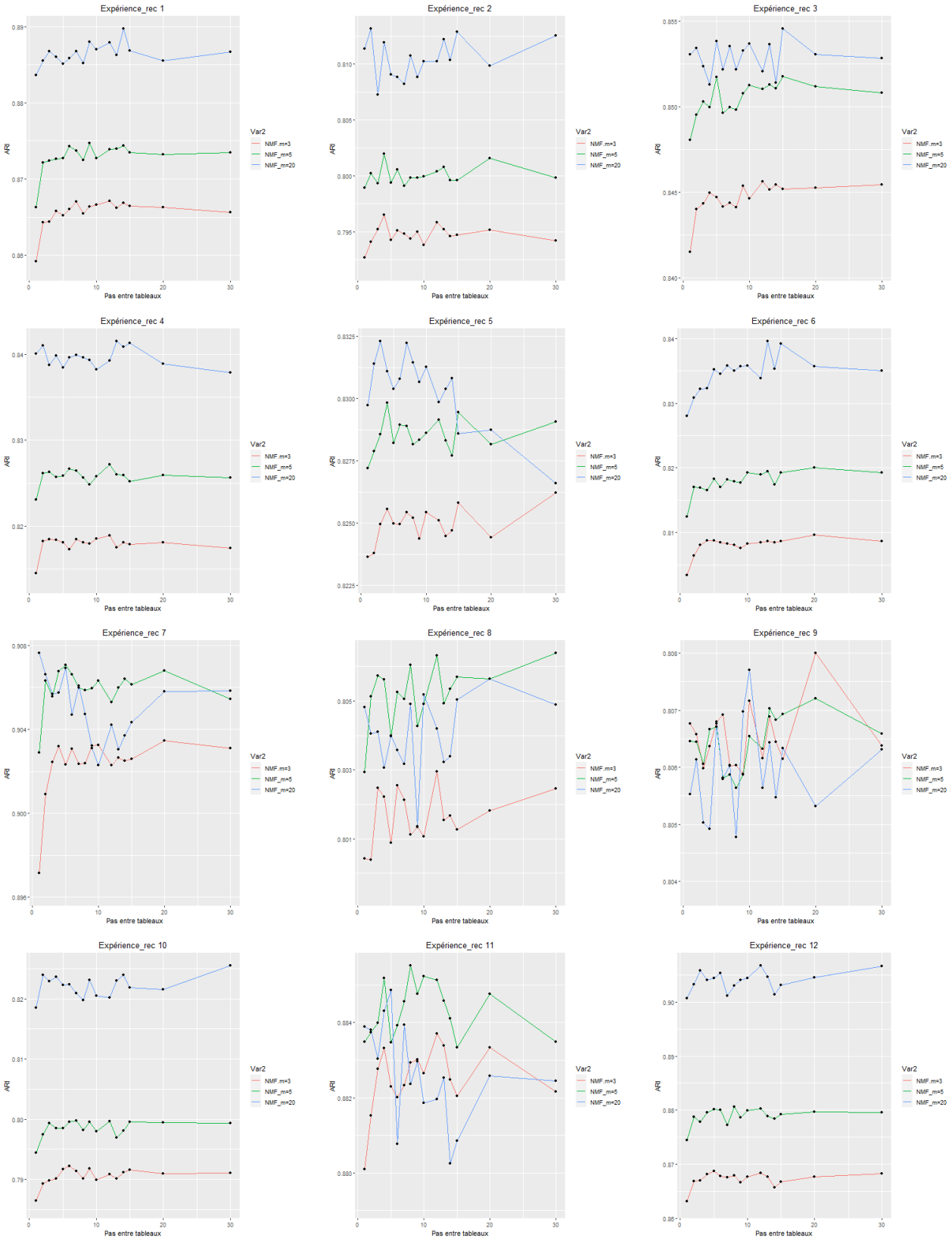
FIGURE 4.3 – Plan d'expérimentation pour $\delta = 1$

4.2.2 Résultats

Dans cette section, on montre les résultats obtenus sur les 30 expériences. On compare l'évolution de l'indice de Rand ajusté noté ARI selon la taille du groupe de tableaux imputés m et suivant le pas considéré δ.

4.2.2.1 Consensus simple NMF

Les figures 4.4 montrent l'évolution de l'indice de Rand (ARI) en fonction du pas et du nombre de tableaux m pour les 30 expériences. L'analyse de ces résultats montre que lorsqu'on considère un petit nombre de tableaux, m=3 et m=5 (respectivement en vert et en bleu), on note une augmentation de l'ARI en passant de $\delta = 1$ à $\delta = 30$. Ainsi, le pas influent sur la performance du consensus de partition lorsque le nombre de tableaux considérés est relativement faible. Par contre, lorsqu'on considère un nombre de tableaux relativement plus grand m=20, on observe une plus grande instabilité de l'évolution des ARI qui n'est pas liée aux pas δ.
Par ailleurs, la performance des consensus de partitions obtenus avec un nombre de tableaux relativement grand m=20 est meilleure pour toutes valeurs du pas. Cela peut être observé à travers les graphiques : 20 sur 30 des expériences montrent que la courbe en rouge, correspondant à m=20, est toujours au dessus des deux autres courbes. Autrement dit, l'ARI obtenu avec un plus grand nombre de tableaux considérés dans la phase d'imputation est en général plus élevé. Ainsi, il apparaît qu'il vaut privilégier un grand nombre de tableaux que de se soucier du pas lors de l'IM.


Expérience_rec 1
Expérience_rec 2
Expérience_rec 3
Expérience_rec 4
Expérience_rec 5
Expérience_rec 6
Expérience_rec 7
Expérience_rec 8
Expérience_rec 9
Expérience_rec 10
Expérience_rec 11
Expérience_rec 12
ARI
Pas entre tableaux
Var2
NMF.m=3
NMF_m=5
NMF_m=20


Expérience_rec 13
Expérience_rec 14
Expérience_rec 15
Expérience_rec 16
Expérience_rec 17
Expérience_rec 18
Expérience_rec 19
Expérience_rec 20
Expérience_rec 21
Expérience_rec 22
Expérience_rec 23
Expérience_rec 24
ARI
Pas entre tableaux
Var2
NMF.m=3
NMF_m=5
NMF_m=20

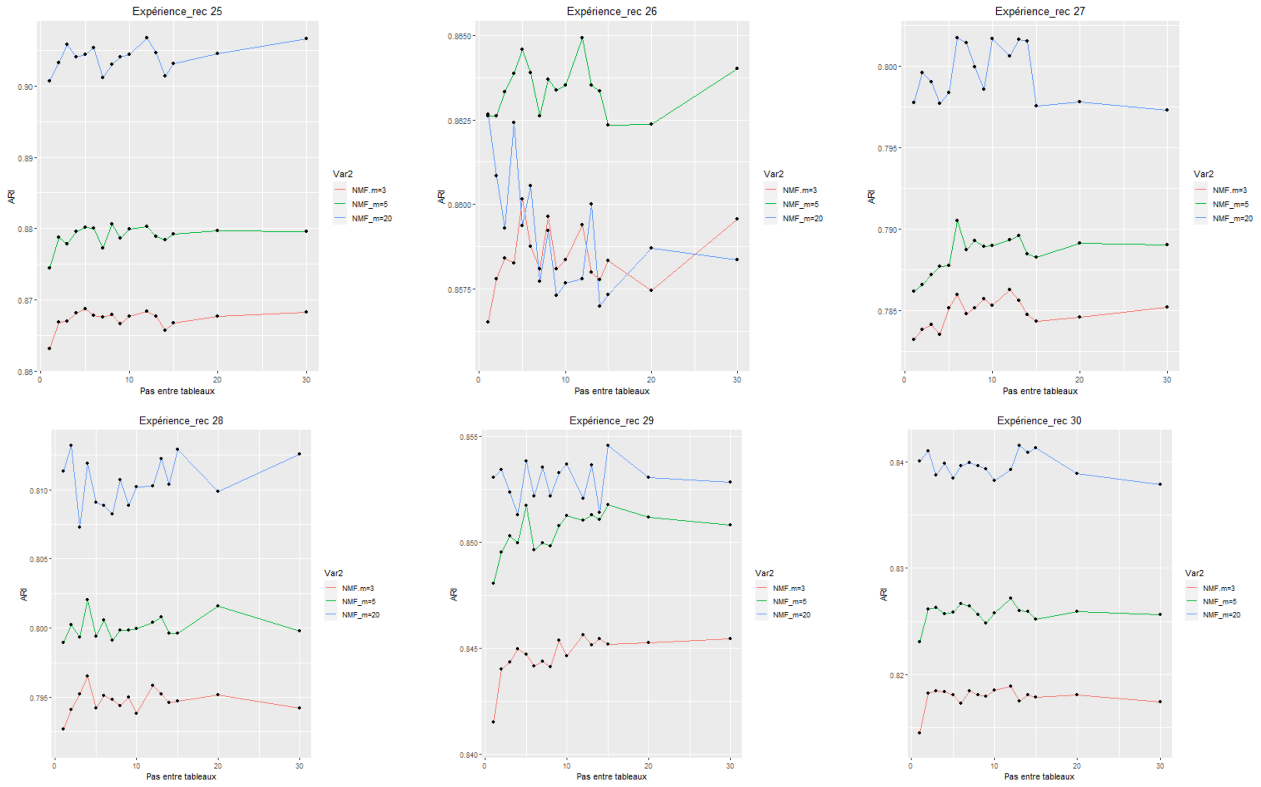


FIGURE 4.4 – L'évolution des indices de Rand ajustés moyens en fonction du pas et du nombre de tableaux considérés pendant l'IM ; approche NMF

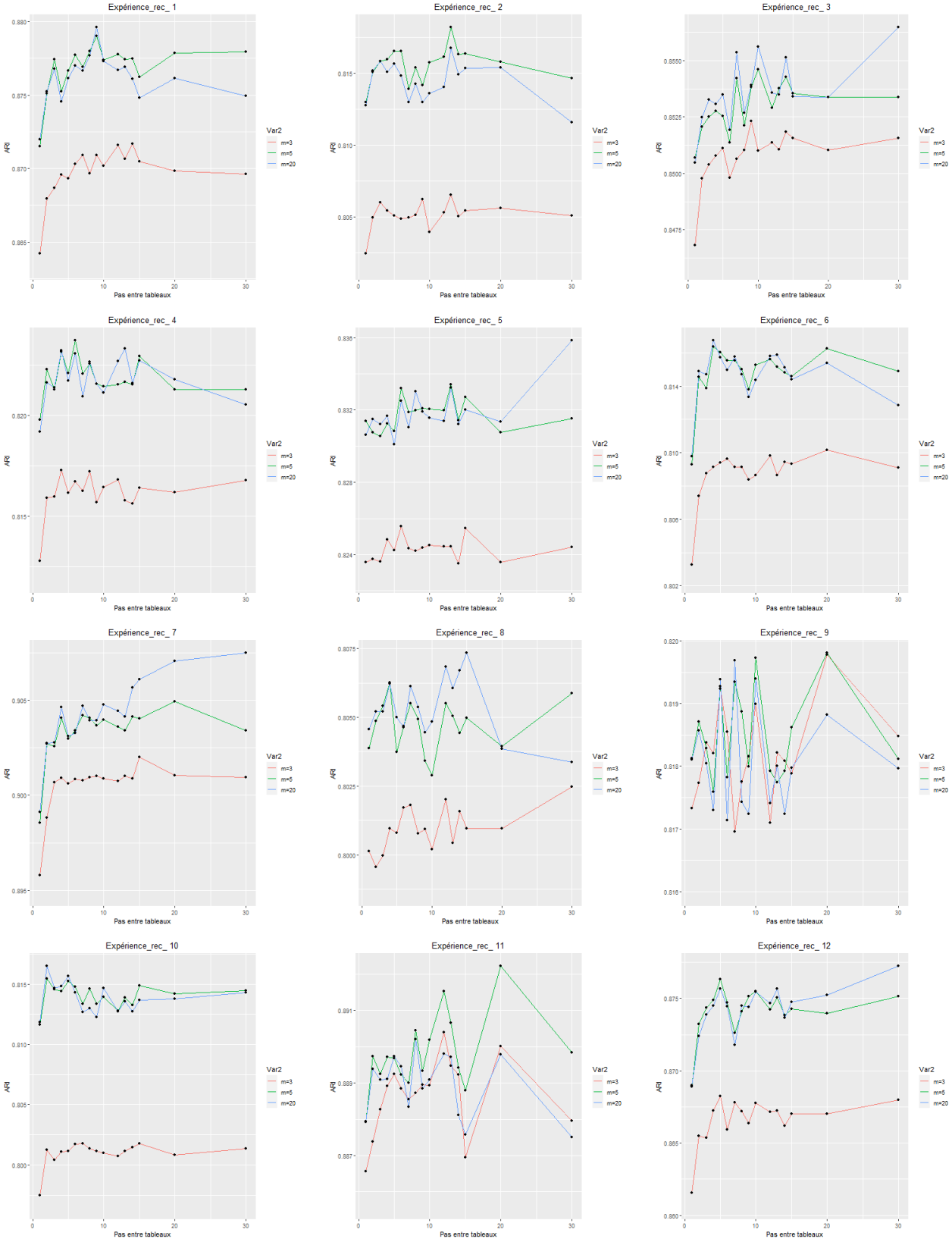
Cette conclusion peut être appuyée par une comparaison entre les ARI obtenus lorsque le nombre de tableaux varie. La figure 4.5 montre les box-plot des ARI obtenus par consensus simple NMF en fonction du nombre de tableaux m, lorsque le pas varie, pour les 30 expériences. L'affichage conjoint en box-plot des 30 expériences illustre mieux cette comparaison. Nous pouvons voir que la variation entre les ARI moyens est très faible pour un nombre de tableaux fixé (box-plots aplatis). Les box-plots, en rouge, correspondant à un nombre de tableaux m=20 se détache le plus vers une valeur supérieure.

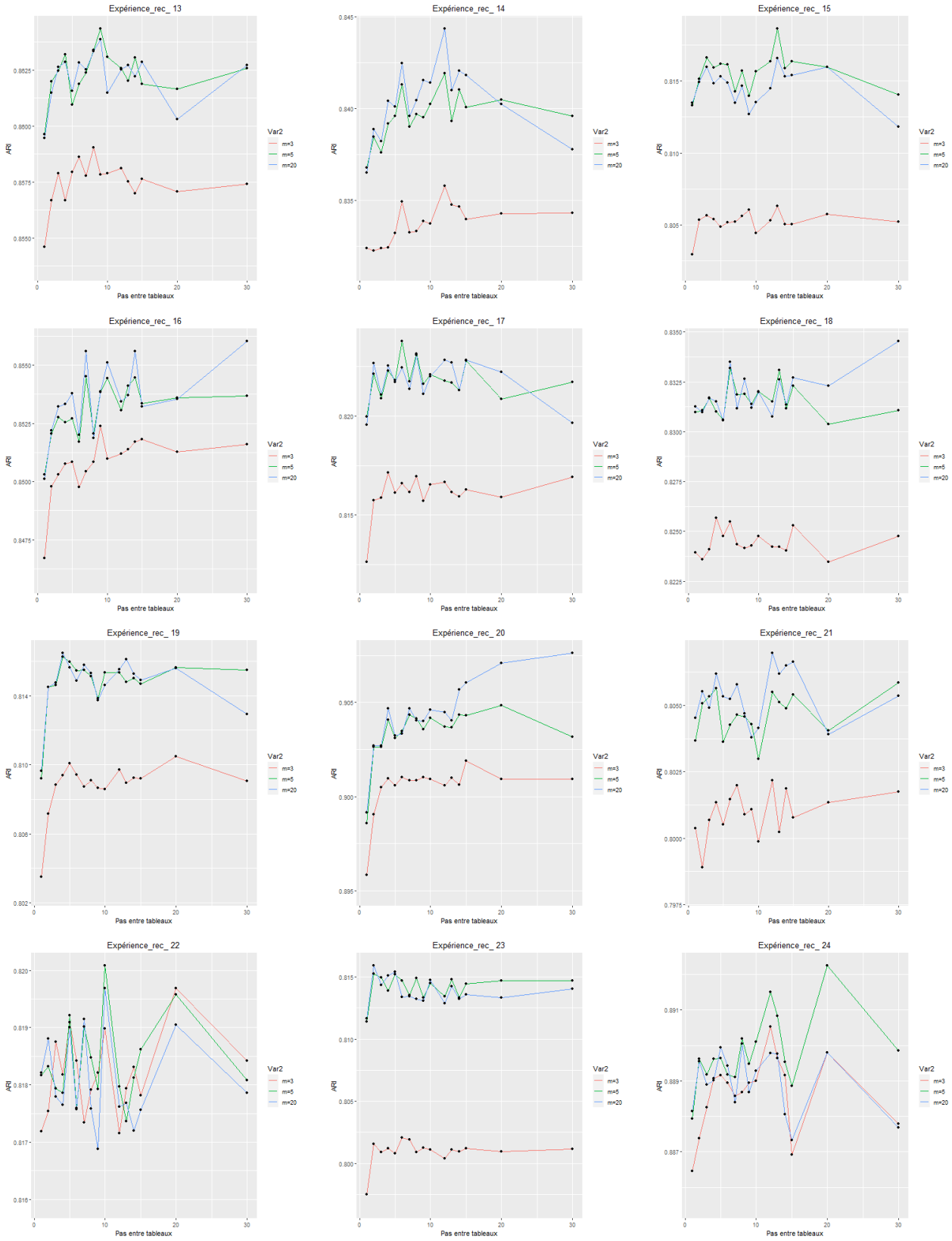
FIGURE 4.5 – Box-plot des indices de Rand ajustés des partitions obtenues par consensus simple NMF en fonction du nombre de tableaux m pour les 30 expériences

4.2.2.2 Apport du consensus pondéré par rapport au consensus simple

Pour évaluer l'apport du consensus pondéré par rapport au consensus simple après IM, nous avons analysé, dans un premier temps, l'évolution des ARI en fonction du pas comme précédemment.

Les figures 4.6 correspondent à l'affichage conjoint de l'évolution des ARI moyens en fonction du pas δ pour chaque m considéré pour les 30 expériences. Pour un m donné, la courbe représente l'évolution de l'ARI suivant le pas δ (en abscisse). Les résultats obtenus sont similaires à ceux observés avec la méthode NMF. Cependant, on note une meilleure stabilité des ARI en fonction du pas lorsqu'on considère m relativement grand m=20 Cela peut s'expliquer par le fait que dans l'approche WNMF, dans sa recherche de poids, si deux partitions se ressemblent, on attribue à l'une d'elles un faible poids. Ainsi, WNMF réduit le nombre de tableaux en attribuant à certains un poids quasi nul parce qu'ils ressemblent à d'autres. Ce qui permet de réduire l'aléatoire lié aux tableaux choisis.


Expérience_rec_ 1
Expérience_rec_ 2
Expérience_rec_ 3
Expérience_rec_ 4
Expérience_rec_ 5
Expérience_rec_ 6
Expérience_rec_ 7
Expérience_rec_ 8
Expérience_rec_ 9
Expérience_rec_ 10
Expérience_rec_ 11
Expérience_rec_ 12
ARI
Pas entre tableaux
Var2
m=3
m=5
m=20


Expérience_rec_ 13
Expérience_rec_ 14
Expérience_rec_ 15
Expérience_rec_ 16
Expérience_rec_ 17
Expérience_rec_ 18
Expérience_rec_ 19
Expérience_rec_ 20
Expérience_rec_ 21
Expérience_rec_ 22
Expérience_rec_ 23
Expérience_rec_ 24
ARI
Pas entre tableaux
Var2
m=3
m=5
m=20

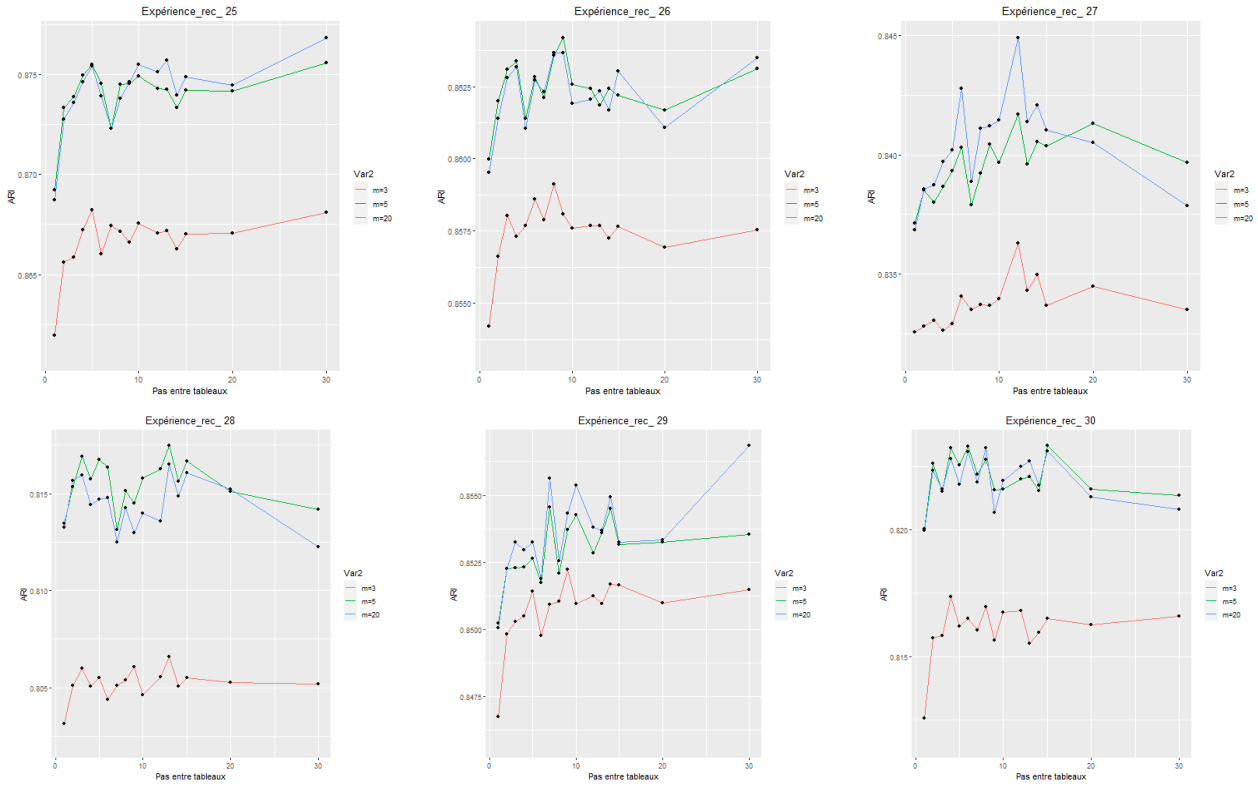


FIGURE 4.6 – L'évolution des indices de Rand ajustés moyens en fonction du pas et du nombre de tableaux considérés pendant l'IM ; approche WNMF

Pour évaluer l'apport du consensus pondéré par rapport au consensus simple, nous allons comparer les résultats issus des deux approches. Ici, pour une approche et un m considérés, on a vecteur de ARI correspondant aux ARI moyens par pas. Cette comparaison se fera en se limitant au nombre de tableaux m=3 et m=5 pour une meilleure lisibilité des graphiques. Les résultats sont données sous forme de box-plots pour chaque expérience (figure 4.7).

En comparant le consensus simple NMF et le consensus pondéré WNMF après imputation multiple, nous avons eu les résultats illustrés à travers la figure 4.7. La majeure partie des expériences montre une plus grande valeur de l'ARI avec une approche WNMF pour un m considéré, lorsque le pas varie. Ici, on considère pour un m donné, la liste des ARI où chaque ARI correspond à celui d'un pas.

Par exemple, en considérant l'expérience 1, en violet nous avons la méthode WNMF avec m=3 et en rouge la méthode NMF avec m=3 également. En comparant ces deux boites à moustaches, nous pouvons remarquer que la boite en violet tend à avoir des valeur plus grande comparer à celle en rouge. De même, pour la même expérience, nous constatons également que pour un m=5, l'approche WNMF fournit de meilleurs résultats de façon globale.

FIGURE 4.7 – Box-plots des indices de Rand ajustés des 30 expériences en fonction du nombre de tableaux considérés m et de la méthode de consensus NMF et WNMF

Synthèse

L'examen des résultats issus de nos simulations permettent de voir que le postula de départ, formulé sur la base des résultats en régression après imputation, doit être nuancé en classification. En effet, selon ce postulat, plus les tableaux imputés sont indépendants ou plus le nombre de tableaux imputés est grand, meilleurs sont résultats du consensus de partitions. Cependant, on constate qu'un tel postulat n'est pas toujours vérifié d'après nos résultats dans les simulations. En effet, les expérimentations montrent que le pas a une influence que lorsque le nombre de tableaux imputés est faible. Par contre, lorsqu'on considère un grand nombre de tableaux, nous avons une faible maîtrise sur les fluctuations de l'ARI en fonction du pas. Autrement dit, la performance du consensus dépend plus des tableaux considérés (l'aléa lié aux expérimentations) que du pas, contrairement à ce qui était observé en classification. Cependant, le nombre de tableaux imputés reste important. Plus on a tableaux imputés, plus on a de meilleures performances au regard de l'ARI.

L'analyse des résultats de ces simulations nous permet d'émettre une autre contribution en termes de méthodologie. Nous notons que dans le cadre du consensus de partitions après

IM, il n'est pas nécessaire d'imputer un très grand nombre de tableaux pour s'assurer de l'indépendance entre tableaux. En effet, cette indépendance entre tableaux imputés n'est pas déterminante dans la qualité de la partition consensus contrairement dans le cadre du modèle de régression agrégé après IM.

Chapitre 5

Application sur des données réelles

Introduction

Dans cette partie, nous présenterons une application des méthodes de consensus de partitions dans le cadre des données manquantes en comparant notamment l'approche de consensus pondéré et le consensus simple. Nous allons également voir le calcul de l'instabilité qui nous aidera dans le choix du nombre de classes.

5.1 Données et problématique

Les données ont été obtenues dans le cadre de la campagne nationale menée dans les logements par L'Observatoire de la Qualité de l'Air Intérieur (OQAI) entre 2003 et 2005. C'est une campagne qui avait comme objectif de dresser un état de la pollution de l'air dans l'habitat afin de donner les éléments utiles pour l'estimation de l'exposition des populations, la quantification et la hiérarchisation des risques sanitaires associés, ainsi que l'identification des facteurs prédictifs de la qualité de l'air intérieur. Elle est faite sur un échantillon de 567 logements représentatif du parc des 24 millions de résidences principales de la France continentale métropolitaine. Durant cette campagne, plusieurs informations relatives aux logements ont été recueillies. En outre, plus de 30 paramètres (chimiques, biologiques, physiques) de pollution ont été mesurés, sur une durée d'une semaine, à plusieurs emplacements à l'intérieur des logements, dans les garages attenants lorsqu'ils existent et à l'extérieur.

Dans cette étude, on s'intéresse particulièrement aux 34 polluants dont la description détaillée est donnée en annexe 5.3. Il s'agit des données dont la collecte n'était pas évidente engendrant ainsi un grand nombre de données manquantes. L'objectif de cette étude est de trouver la structure en classes des ménages au regard des données de polluant recueillies. Il s'agit, in fine, d'appliquer une classification tout en tenant compte de la présence de données manquantes. Pour ce faire, nous allons appliquer une imputation multiple afin de disposer de tableaux de données complets, puis faire une analyse de chaque tableau et agréger tous les résultats d'analyse. Ce processus sera fait compte tenu des résultats dans nos simulations.

5.1.1 Analyse descriptive

Nous avons réalisé une analyse descriptive des polluants. Le tableau 5.2 présente le résumé des résultats. On observe une présence systématique de valeurs manquantes dans des proportions allant de 2.65% à 48.85% du nombre d'observations initiales. On note par exemple une proportion de valeurs manquantes de 2.65% pour le formaldéhyde (ald21), 4.94% pour le 2-butoxy éthanol (cov51). Cette proportion atteint 18.52% pour le radon en chambre (radb.chb) et elle est maximale pour les PM (48.85% pour le PM2.5 et 47.62% pour les PM10).

Par la suite nous avons effectué une analyse des corrélations entre polluants. Pour cela, nous avons procédé à une imputation simple en utilisant les méthodes factorielles [23]. Les résultats sont résumés par la figure 5.1. On observe des regroupements de polluants corrélés entre eux. Par contre on voit une séparation entre les composés organique volatiles, les polluants chimiques (le radon, les PM) et polluants biologiques (allergènes de poussière).
Les variables PM10 et PM25 ainsi que les variables radb.sej et radb.chb ont une très forte corrélation (couleur bleu foncée, voir cf. échelle). Par contre les corrélations entre les COV et les variables PM10 et PM25 sont relativement faibles.

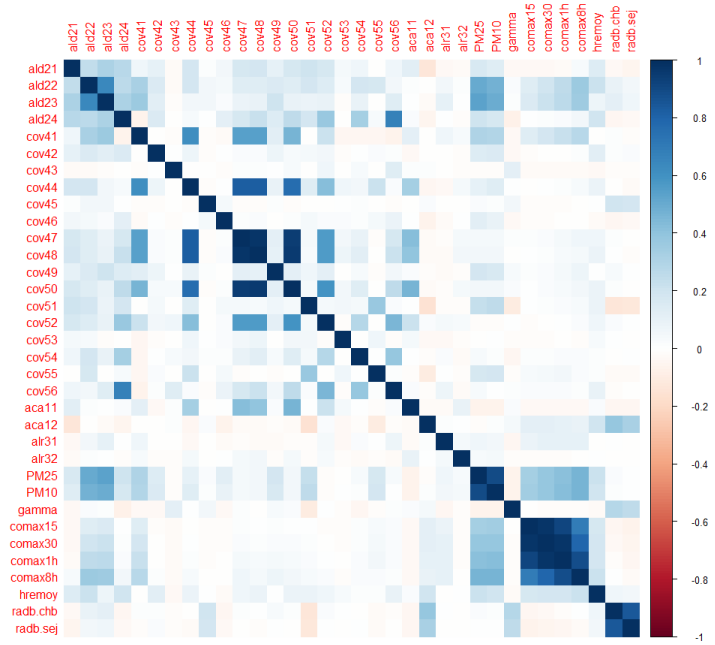


FIGURE 5.1 – Corrélation entre les polluants

Au regard de nos analyses descriptives, nous allons travailler sans les variables PM25 et PM10 à cause de leurs pourcentage de données manquantes qui atteignent presque la moitié des observations (près de 50%). Avant d'entamer l'analyse, nous allons procéder à une normalisation des variables de la base.

5.2 Pré-traitement des données

Plusieurs raisons motivent la normalisation des variables. Premièrement, c'est un moyen d'harmoniser les variables de la base et de corriger les valeurs aberrantes. Par ailleurs, il arrive fréquemment que les modèles statistiques, celles de classification non supervisée en particulier par modèle de mélanges, reposent sur une hypothèse de normalité des variables. De plus, dans notre processus d'imputation, on utilise une approche basé sur de modèles de mélange gaussien JM-DP : d'où l'intérêt de cette transformation. Quand cette hypothèse n'est pas vérifiée, on peut souhaiter effectuer une transformation des variables pour s'y ramener. Ainsi, pour se prémunir de tout ça, nous allons procéder à une normalisation des variables de la base. Pour la normalisation, il existe plusieurs approches, mais il est difficile de trouver la transformation optimale sans une méthode rigoureuse.

Ainsi, Box et Cox [24] ont proposé des transformations génériques qui sont paramétrables. Dans cette étude, nous utilisons la transformation la plus simple parmi celles-ci. Elle est définie comme suit :

$$z_i \mapsto f^{\lambda}(z_i) = \begin{cases} \frac{z_i^{\lambda}-1}{\lambda} & si \ \ \lambda \neq 0 \\ \ln(z_i) & sinon \end{cases} \tag{5.1}$$

En définissant correctement le paramètre λ, on arrive à rapprocher la distribution des variables de la normalité. Pour le choix de λ, nous avons évalué l'évolution des coefficients de corrélation linéaires entre les quantiles empiriques de la distribution transformée et ceux de la loi normale centrée réduite. Pour chaque variable, nous avons retenu le paramètre λ qui donne le plus grand coefficient de corrélation.

5.3 Résultats de l'application sur les données de logement

Dans cette section, nous allons présenter les résultats de la classification des ménages par les deux approches : NMF et WMNF après IM. Pour cela, nous allons commencer par chercher le nombre de classes optimal en utilisant l'instabilité liée aux données manquante (voir 3.2.2.2).

5.3.1 Choix du nombre de classes

Dans le cas de l'imputation multiple, on peut se baser sur la stabilité des partitions liées aux données imputées. L'explication de l'évaluation de cette stabilité sera donnée dans la suite de ce document. Pour la recherche du nombre de classes optimal, Niang N. et Audigier V. [5] ont proposé une approche basée sur l'instabilité S calculée à l'étape 3 des règles de Rubin. En effet, cette instabilité est un indicateur de qualité du consensus de partitions obtenu. Plus elle est faible, plus les résultats sont robustes et meilleure est la partition.

Dans notre approche, nous avons choisi les valeurs de K $\{2, 3, 4, 5, 6\}$. Pour chaque K, nous avons fait l'imputation multiple de 20 tableaux, puis mesuré l'instabilité S liées aux

données manquantes. Les résultats sont fournis par la figure 5.2.

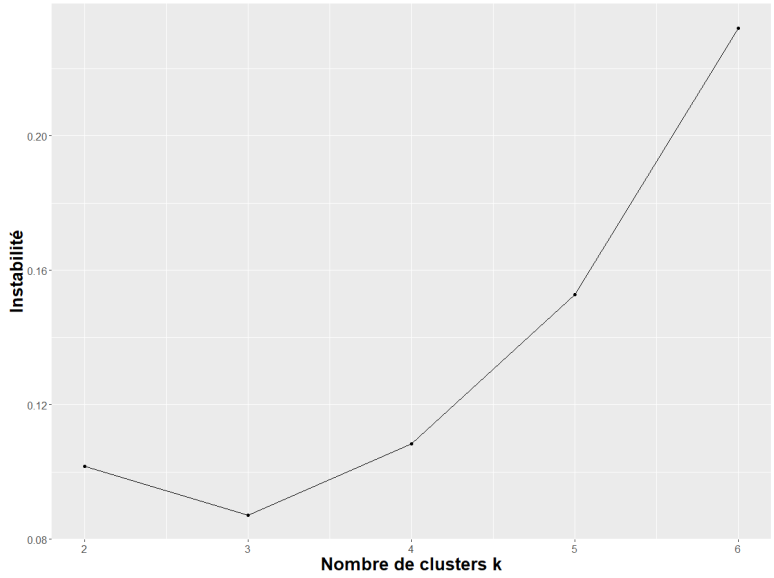


FIGURE 5.2 – Instabilité en fonction du nombre de classes K

Nous pouvons voir, à travers cette figure que l'instabilité est minimale lorsque le nombre de classes est de 3. Ainsi, on considère 3 classes dans la recherche de partitions. Dans la sous section suivante, nous présenterons les résultats de l'application NMF et WNMF sur les données de logement présentées dans ce qui précède.

5.3.2 Résultats de la classification

Dans le cadre de la recherche de partitionnement des logements selon la qualité de l'air intérieur, nous avons avons à notre disposition un tableaux de données contenant des manquantes. Nous avons adopté une approche de d'imputation multiple.

- Dans la phase d'imputation, 20 tableaux complets ont été imputés.
- Dans la phase d'analyse, nous avons appliqué sur chacun de ces tableaux une classification par la méthode des k-moyennes.
- L'agrégation a été faite selon les deux approches de consensus de partitions : l'approche NMF et celle WNMF.

Nous commençons par regarder la relation entre les partitions obtenues sur les 20 tableaux imputés.

5.3.2.1 Relation entre les partitions obtenues sur les 20 tableaux imputés

La figure 5.3 affiche les ARI entre les 20 partitions obtenues sur les 20 tableaux imputés. L'analyse de ces résultats montre que la plupart de ces partitions se ressemble. Par exemple, la partition 10 ressemble fortement à la partition issue du tableau 11 avec un ARI de 0.97.

Le consensus est fait sur ces partitions. Nous allons ainsi comparer les résultats issus des deux approches de consensus : à savoir NMF et WNMF.

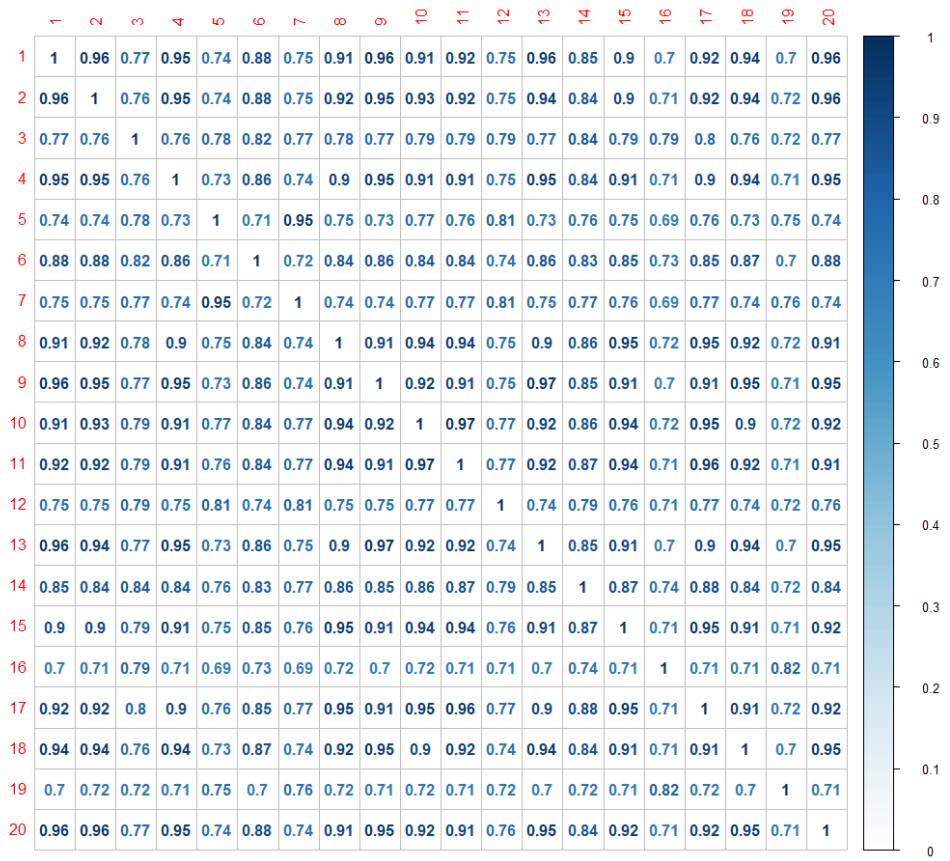


FIGURE 5.3 – L'indice de Rand entre les partitions obtenues sur les 20 tableaux

5.3.2.2 Présentation des résultats

Après l'application des approches NMF et WNM, nous avons obtenus les résultats suivants. L'approche NMF considère que toutes les partitions sont équivalentes : elles ont toutes le même poids à savoir $\frac{1}{20} = 0.05$. Par contre, le WNMF permet de rechercher un système de poids selon la particularité des partitions et leur ressemblance. Ainsi, l'approche WNMF fournit les poids suivants (voir tableau 5.1).

	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5	Partition 6	Partition 7	Partition 8	Partition 9	Partition 10
W	0,0000	0,0359	0,0000	0,0000	0,0247	0,0000	0,0000	0,2313	0,0000	0,0000
	Partition 11	**Partition 12**	**Partition 13**	**Partition 14**	**Partition 15**	**Partition 16**	**Partition 17**	**Partition 18**	**Partition 19**	**Partition 20**
W	0,0000	0,0000	0,0000	0,0000	0,2492	0,0000	0,0827	0,0733	0,3028	0,0001

TABLE 5.1 – Les poids obtenus avec l'approche WNMF

Dans la présentation de l'approche WNMF, nous avions vu que lorsque deux partitions se ressemblent, on attribue, a priori, a une des deux partitions, à un poids assez faible. Ainsi, les résultats obtenus sur les données réelles illustre ce fait. En effet, nous constatons que la plupart des partitions ont un poids très faible, voire nul. Par exemple, la partition 1 ressemble

fortement à la partition 2, au regard de l'ARI, avec un ARI de 0.96. De ce fait, on considère que la partition 2 et attribue un poids quasi nul à la partition 1. Ainsi, certaines partitions ont eu des poids nuls étant donné qu'elles ressemblent fortement à d'autres partitions. Cela est avantageux dans la mesure où il permet d'éviter la redondance de certaines partitions ; chose qui pourrait biaiser le consensus.

Caractérisation des classes obtenues

Une exécution des deux approches de consensus de partitions fournit les partitions visualisées dans les figures 5.4. On peut voir que les approches donnent des partitions un peu similaires.

FIGURE 5.4 – Visualisation des classes obtenues avec les deux approches de consensus de partitions sur le premier plan de l'ACP sur le tableau après imputation par la méthode factorielle

On peut caractériser les classes obtenues en regardant les variables les plus discriminantes pour les deux partitions. On peut visualiser à travers la figure 5.5 les variables les plus discriminantes par rapport aux deux partitions obtenues. Dans les deux approches, nous avons les mêmes variables : trois variables relatives à la concentration maximale en monoxyde de carbone (comax1h : Maximum des moyennes sur 1 heure, comax30 : Maximum des moyennes sur 30 min , comax15 : Maximum des moyennes sur 15 min) sur la semaine.

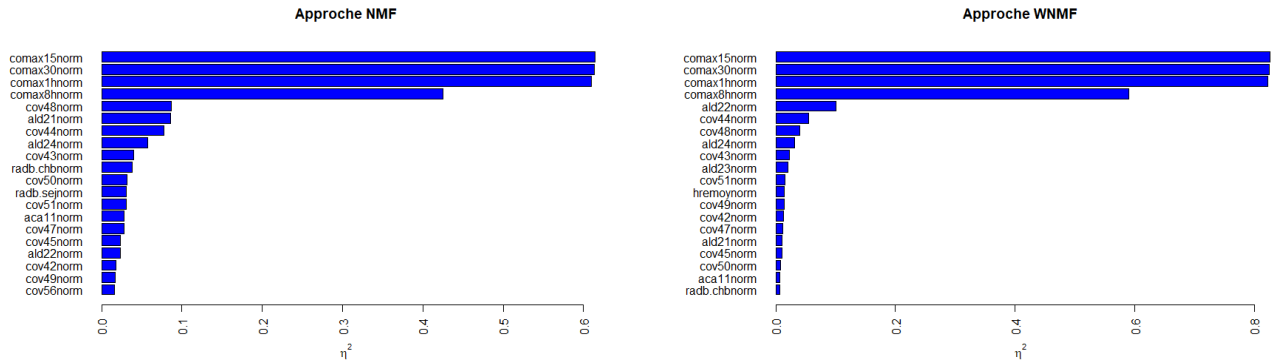


FIGURE 5.5 – Les variables les plus discriminantes par rapport aux deux partitions obtenues

Les figures 5.6 montrent la distribution, selon les 3 classes, du maximum des moyennes pendant 30 minutes en termes de concentration en monoxyde carbone à l'intérieur des logements selon les deux approches. L'analyse de ces résultats montre que :

- avec l'approche NMF, les classes 2 et 3 regroupent les logements à l'intérieur desquels une forte concentration maximale en monoxyde de carbone a été notée contrairement à la classe 1 ;
- avec l'approche WNMF, c'est la classe 3 qui regroupe les logements au sein desquels une faible concentration maximale en monoxyde de carbone a été notée contrairement aux deux classes.

Cette distribution reste la même pour les deux autres variables dont l'illustration est donnée en annexe (cf. figure 5.7).

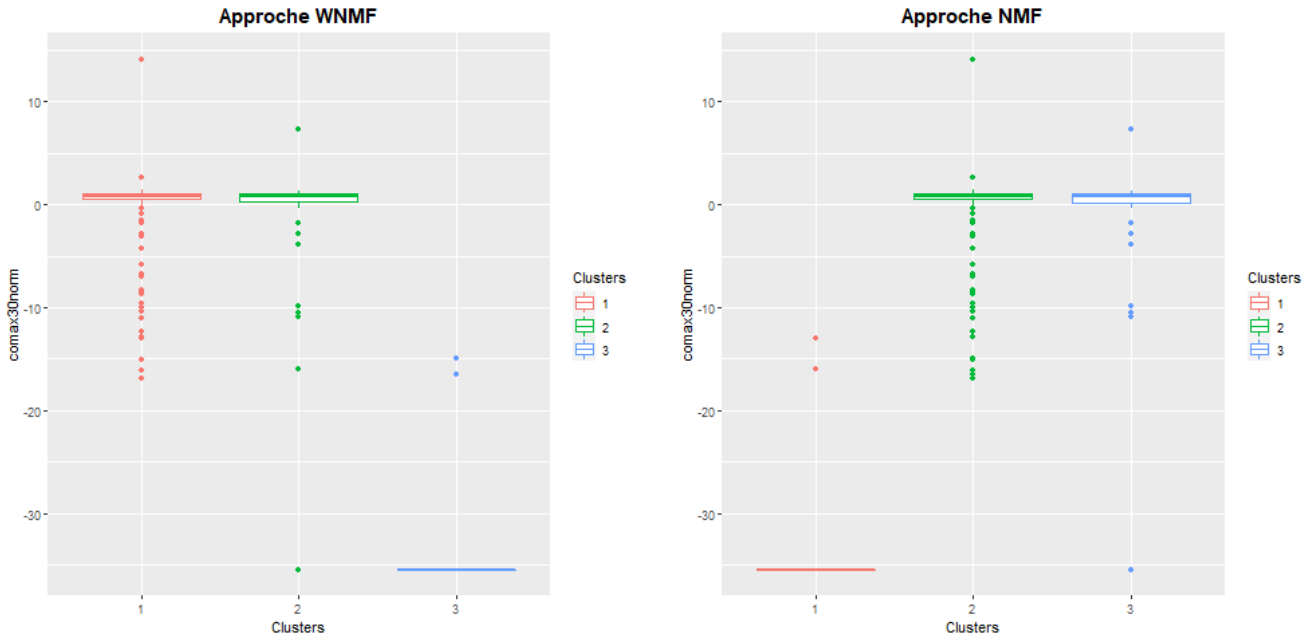


FIGURE 5.6 – Distribution du maximum des moyennes pendant 30 minutes en termes de concentration en monoxyde carbone

5.3.3 Performances des partitions obtenues

Nous avons également analysé la performance pour les deux approches. N'ayant pas ici une vérité terrain, seuls des indices de qualité interne tels que le coefficient de silhouette sont disponibles. C'est ce dernier qui a été utilisé dans cette analyse.

Cependant, comme évoqué précédemment, le calcul de cet indice nécessite une matrice de distance obtenue à partir du tableau de départ. En clustering après IM celle-ci n'est pas disponible directement.

Une contribution méthodologique de notre travail a consiste à proposer une méthode de calcul d'indice de qualité interne, plus précisément l'indice de silhouette, après imputation multiple. Nous proposons de calculer cet indice sur la base du tableau moyen des tableaux

imputés. Pour l'approche NMF, on considère la moyenne simple des m tableaux imputés. Pour l'approche WNMF, on considère la moyenne pondérée par les poids fournis par l'algorithme WNMF. C'est cette matrice moyenne qui sera considérée comme la matrice de distance.

Une comparaison des performances des deux approches nous donne des indices de silhouette similaires : 0.293 pour NMF et 0.286 pour WNMF. Il est possible aussi de comparer la similarité entre les deux partitions obtenues selon les deux approches grâce à l'ARI. Cet indice entre les deux partitions est de 0.97 stipulant une forte ressemblance entre les deux partitions.

Ainsi, l'application des approches NMF et WNMF après IM montre qu'il n'y a pas une grande différence entre les deux approches en termes de performances. Cependant, on note qu'avec l'approche WNMF, la partition consensus est obtenue sur 7 tableaux puisque les autres ont eu des poids nuls alors que l'approche NMF a utilisé tous les 20 tableaux imputés.

CONCLUSION

L'objectif principal du stage était d'évaluer l'apport du consensus pondéré lorsque les données manquantes sont gérées par imputation multiple. En particulier, il s'agissait d'évaluer l'apport de la méthode WNMF par rapport à celle dite NMF en tenant compte de l'indépendance théorique entre les tableaux imputés mais aussi du nombre de tableaux considérés. Après avoir fait une revue sur les méthodes de classification, les méthodes de consensus de partitions et sur les méthodes d'imputation multiple, nous avons commencé par vérifier l'importance de la prise en compte de l'indépendance entre les tableaux lors de la phase d'analyse dans le cadre de la régression. Nous avons, ensuite, essayé d'étendre ces résultats en clustering, à travers notamment les méthodes NMF et WNMF. L'idée était de vérifier l'hypothèse selon laquelle plus les tableaux imputés sont indépendants meilleures seront les performances du consensus de partitions, comme en régression. Il s'est agit enfin, d'évaluer l'apport du consensus de partitions pondéré WNMF par rapport au consensus simple NMF.

Des expérimentations sur des données simulées ont permis de voir que l'indépendance entre les tableaux imputés avait une influence sur les résultats d'analyse que lorsque le nombre de tableaux imputés était relativement faible.
Ainsi, une première contribution est de noter que lorsqu'on travaille en clustering après imputation multiple, en prenant un nombre relativement élevé de tableaux, il n'est pas nécessaire de s'assurer de l'indépendance entre ces derniers car celle-ci n'améliore pas la qualité de la partition finale.

Par ailleurs, nous avons noté que les deux approches de consensus de partitions (pondéré et simple) après imputation multiple n'ont pas une grande différence en termes de performance. Ce constat est également noté au niveau de l'application sur les données réelles portant sur l'air intérieur des logements en France.

Une autre contribution est celle liée au calcul de l'indice de silhouette après imputation

multiple. En effet, le calcul de cet indice nécessite une matrice de distance entre les observations obtenue à partir du tableau sur lequel la partition a été obtenue dont on ne dispose pas en imputation multiple. Nous avons ainsi proposé une approche pour calculer cet indice en clustering après imputation multiple.

Une contribution informatique de ce travail a consisté à implémenter sur R l'application de l'approche WNMF qui n'existait pas encore sur R à notre connaissance. Cette tâche nous a permis d'approfondir nos compétences en développement informatique avec le logiciel R.

Le sujet traité dans ce rapport présente plusieurs perspectives dans un contexte marqué par les gros volumes de données hétérogènes contenant des données manquantes. L'hétérogénéité et les données manquantes deviennent des caractéristiques incontournables des données auxquelles le scientifique doit désormais faire face. Or, cela pose de grandes difficultés pour l'analyse car les méthodes classiquement envisagées ne permettent pas de gérer ces deux caractéristiques simultanément. Dans les futurs travaux, on peut envisager d'investiguer une nouvelle classe de modèles, par exemple les modèles clusterwise pour le traitement de données complexes par la présence de données manquantes et une structure des individus en groupes inconnus.

Bibliographie

[1] M. A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82 :528–540, 1987.

[2] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London, 1997.

[3] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2002.

[4] DONALD B. RUBIN. Inference and missing data. *Biometrika*, 63(3) :581–592, 12 1976.

[5] Vincent Audigier and Ndèye Niang. Clustering with missing data : which equivalent for Rubin's rules ? 39 pages, November 2020.

[6] Ludmila Kuncheva and Lakhmi C. Jain. Nearest neighbor classifier : Simultaneous editing and feature selection. *Pattern Recognit. Lett.*, 20(11-13) :1149–1156, 1999.

[7] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[8] A. Likas, N. Vlassis, Aristidis Likas, Nikos Vlassis, and J.J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36 :451–461, 2001.

[9] Prem P. Jain and Brian Pridemore. Case study : Net-centric mission threads modeling and analysis using BPMN. In William K. McQuay and Waleed W. Smari, editors, *2008 International Symposium on Collaborative Technologies and Systems, CTS 2008, Irvine, California, USA, May 19-23, 2008*, pages 563–564. IEEE, 2008.

[10] Richard C. Dubes and Anil K. Jain. Clustering techniques : The user's dilemma. *Pattern Recognit.*, 8(4) :247–260, 1976.

[11] Chris H. Q. Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8) :3913–3927, 2008.

[12] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3 :583–617, 2002.

[13] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3 :583–617, March 2003.

[14] Tao Li, Chris H. Q. Ding, and Michael I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA*, pages 577–582. IEEE Computer Society, 2007.

[15] C. Lavit, Y. Escoufier, R. Sabatier, and Pierre Traissac. The ACT (STATIS method). *Computational Statistics and Data Analysis*, pages 97–119, 1994.

[16] Ndèye Niang and Mory Ouattara. Weighted consensus clustering for multiblock data. In *SFC 2019*, Paris, France, September 2019.

[17] Hang J. Kim, Jerome P. Reiter, Quanli Wang, Lawrence H. Cox, and Alan F. Karr. Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3) :375–386, 2014.

[18] Michael Lavine and Mike West. A Bayesian method for classification and discrimination. *The Canadian Journal of Statistics*, 20(4) :451–461, 1992.

[19] Junhui Wang. Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4) :893–904, 2010.

[20] Yixin Fang and Junhui Wang. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3) :468–477, 2012.

[21] Maxime Guillon, François Leitner, and Laurence Nigay. VTE : une technique de pointage à distance. In *Actes informels de la 26e conférence francophone sur l'Interaction Homme-Machine : démonstration*, pages 8–9, Lille, France, 2014.

[22] Wikipédia. Loi normale multidimensionnelle — wikipédia, l'encyclopédie libre, 2021. [En ligne ; Page disponible le 11-avril-2021].

[23] Julie Josse and François Husson. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, 56(6) :1869–1879, 2012.

[24] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, (2) :pp. 211–252.

ANNEXES

Annexe A : Description univariée des données à mettre en annexe

	ald21	ald22	ald23	ald24	cov41	cov42	cov43	cov44	cov45
Min.	1.29	1.77	0.00	1.62	0.00	0.00	0.00	1.51	0.00
1st Qu.	14.59	7.86	0.70	9.09	1.33	0.90	0.05	7.32	0.60
Median	19.67	11.46	1.07	13.58	2.01	0.90	1.01	12.02	1.41
Mean	22.73	14.12	1.35	19.74	2.69	4.69	10.77	22.12	3.92
3rd Qu.	28.99	17.21	1.62	22.94	3.18	4.65	1.64	21.05	2.62
Max.	86.33	94.62	12.93	368.54	22.78	170.14	4087.21	414.20	684.33
NA's	15.00	15.00	15.00	15.00	28.00	28.00	28.00	28.00	28.00
NA's (%)	2.65	2.65	2.65	2.65	4.94	4.94	4.94	4.94	4.94
	cov49	cov50	cov51	cov52	cov53	cov54	cov55	cov56	aca11
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
1st Qu.	0.69	1.49	0.75	2.44	2.26	3.00	0.00	3.72	0.24
Median	0.94	2.25	1.55	4.05	4.14	5.38	0.00	6.24	1.60
Mean	1.26	4.10	2.79	6.64	51.76	18.25	0.04	17.33	16.75
3rd Qu.	1.42	3.99	3.07	6.78	12.61	11.95	0.00	12.56	12.86
Max.	35.12	112.32	60.59	111.69	4809.76	1774.09	12.21	502.10	608.00
NA's	28.00	28.00	28.00	28.00	28.00	28.00	28.00	28.00	129.00
NA's (%)	4.94	4.94	4.94	4.94	4.94	4.94	4.94	4.94	22.75
	PM25	PM10	gamma	comax15	comax30	comax1h	comax8h	hremoy	radb.chb
Min.	1.20	1.60	0.01	0.00	0.00	0.00	0.00	25.50	5.00
1st Qu.	14.00	22.00	0.05	0.00	0.00	0.00	0.00	43.10	20.00
Median	44,488.00	31.30	0.06	3.00	2.80	2.10	0.50	48.70	34.00
Mean	37.08	53.59	0.07	5.14	4.53	3.85	1.69	48.75	63.58
3rd Qu.	35.40	56.80	0.09	6.00	5.45	4.90	2.30	54.30	65.00
Max.	567.70	522.60	0.26	130.70	90.50	52.70	33.00	72.80	1,115.00
NA's	277.00	270.00	29.00	33.00	33.00	33.00	33.00	66.00	105.00
NA's (%)	48.85	47.62	5.11	5.82	5.82	5.82	5.82	11.64	18.52
	cov46	cov47	cov48	aca12	alr31	alr32	radb.sej		
Min.	0.00	0.00	0.75	0.01	0.09	0.51	4.00		
1st Qu.	0.00	1.49	3.55	0.20	0.09	0.51	19.00		
Median	0.00	2.24	5.49	1.61	0.09	0.51	36.00		
Mean	0.45	3.81	10.63	7.15	0.61	0.74	68.84		
3rd Qu.	0.00	3.72	10.02	7.65	0.09	0.51	69.25		
Max.	39.46	85.30	232.82	129.38	27.40	12.09	1,983.00		
NA's	28.00	28.00	28.00	129.00	22.00	22.00	103.00		
NA's (%)	4.94	4.94	4.94	22.75	3.88	3.88	18.17		

TABLE 5.2 – Description univariée des données à mettre en annexe

Annexe B : Caractérisation des classes de logements selon les variables les plus discriminantes

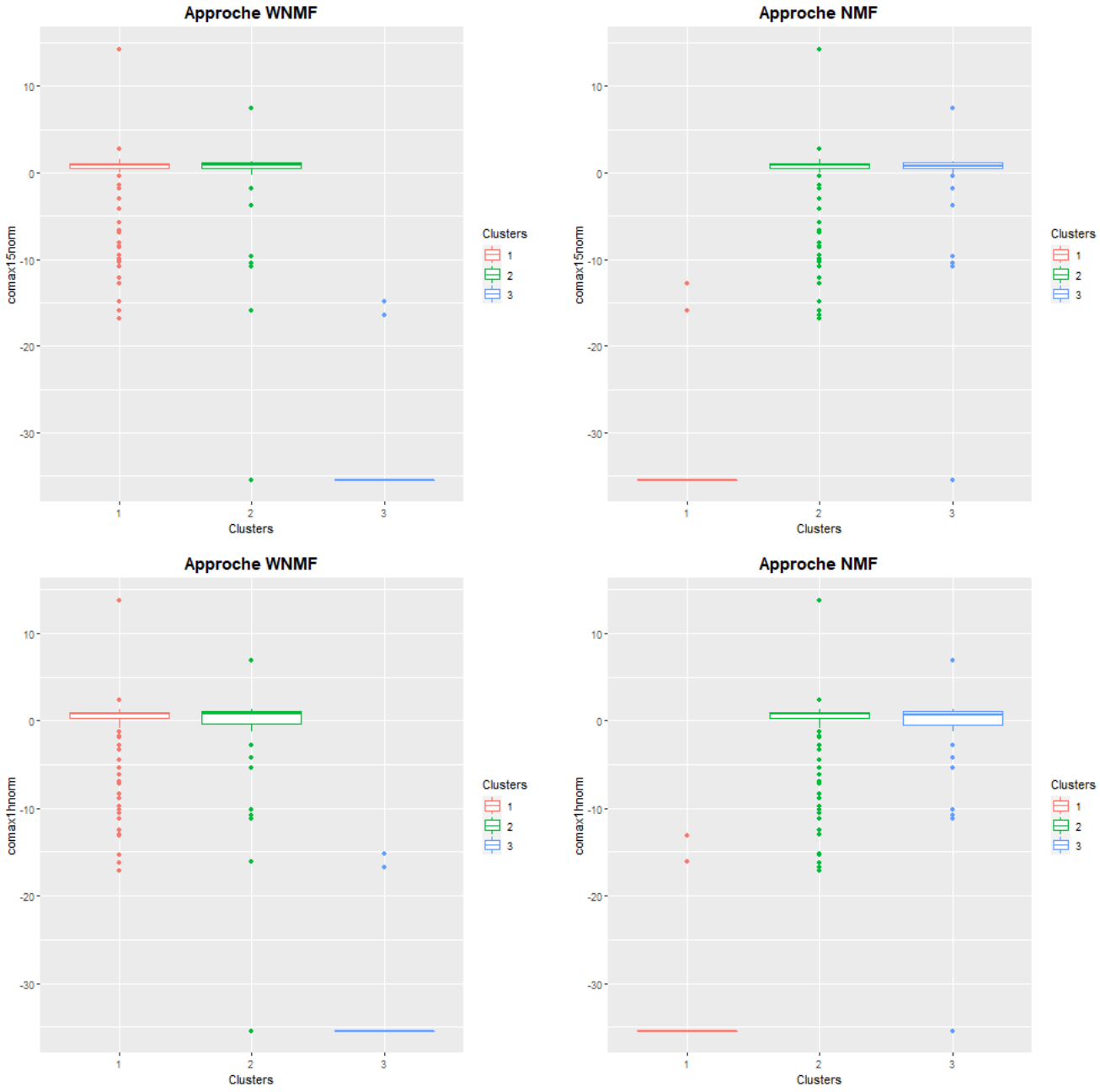


FIGURE 5.7 – Distribution du maximum des moyennes pendant 30 minutes et 1 heure en termes de concentration en monoxyde carbone

Annexe C : Le coefficient de corrélation RV

Soient les tableaux de données Z_t et $Z_{t'}$. On commence par considérer l'objet représentatif W_t de chaque table Z_t. W_t est une matrice (n×n) qui contient tous les liens inter-individus de Z_t.

$$W_t = Z_t Z_t^{'} \tag{5.2}$$

Pour comparer deux tableaux Z_t et Z_t', on utilise le coefficient de corrélation vectorielle RV donnée.

$$RV_{tt'} = \frac{trace(W_t, W_{t'})}{\sqrt{trace(W_t)^2 trace(W_{t'})^2)}} \tag{5.3}$$

Annexe D : Description des variables de polluants

Variable	Signification	Nom	Unité
aca11	Concentration en Derf1	Allergènes dans la poussière	µg/g
aca12	Concentration en Derp1	Allergènes dans la poussière	µg/g
ald21	Concentration en formaldéhyde	Aldéhydes	µg/m3
ald22	Concentration en acétaldéhyde	Aldéhydes	µg/m3
ald23	Concentration en acroléine	Aldéhydes	µg/m3
ald24	Concentration en hexaldéhyde	Aldéhydes	µg/m3
alr31	Concentration en chat (Fel d1)	Allergènes dans l'air	ng/m3
alr32	Concentration en chien (Can f1)	Allergènes dans l'air	ng/m3
co15	Maximum des moyennes sur 15 min	Monoxyde de carbone	ppm
co30	Maximum des moyennes sur 30 min	Monoxyde de carbone	ppm
co1h	Maximum des moyennes sur 1 heure	Monoxyde de carbone	ppm
co8h	Maximum des moyennes sur 8 heures	Monoxyde de carbone	ppm
comin	Minimum sur la semaine	Monoxyde de carbone	ppm
comoy	Moyenne sur la semaine	Monoxyde de carbone	ppm
comax	Maximum sur la semaine	Monoxyde de carbone	ppm
cov41	Concentration en benzène	Composés organiques volatils	µg/m3
cov42	Concentration en 1-méthoxy-2-propanol	Composés organiques volatils	µg/m3
cov43	Concentration en trichloroéthylène	Composés organiques volatils	µg/m3
cov44	Concentration en toluène	Composés organiques volatils	µg/m3
cov45	Concentration en tétrachloroéthylène	Composés organiques volatils	µg/m3
cov46	Concentration en 1-méthoxy-2-propyl acétate	Composés organiques volatils	µg/m3
cov47	Concentration en ethylbenzène	Composés organiques volatils	µg/m3
cov48	Concentration en m+p-xylène	Composés organiques volatils	µg/m3
cov49	Concentration en styrène	Composés organiques volatils	µg/m3
cov50	Concentration en o-xylène	Composés organiques volatils	µg/m3
cov51	Concentration en 2-butoxy éthanol	Composés organiques volatils	µg/m3
cov52	Concentration en 124-trimethylbenzène	Composés organiques volatils	µg/m3
cov53	Concentration en 1,4-dichlorobenzène	Composés organiques volatils	µg/m3
cov54	Concentration en n-décane	Composés organiques volatils	µg/m3
cov55	Concentration en 2-butoxy éthyl acétate	Composés organiques volatils	µg/m3
cov56	Concentration en n-undécane	Composés organiques volatils	µg/m3
hremoy	Humidité moyenne	Humidité relative	%

TABLE 5.3 – Description variables de polluant

Annexe E : Matrice des indices de Rand entre les 1000 partitions deux à deux issues es 1000 tableaux générées lors des 4 première

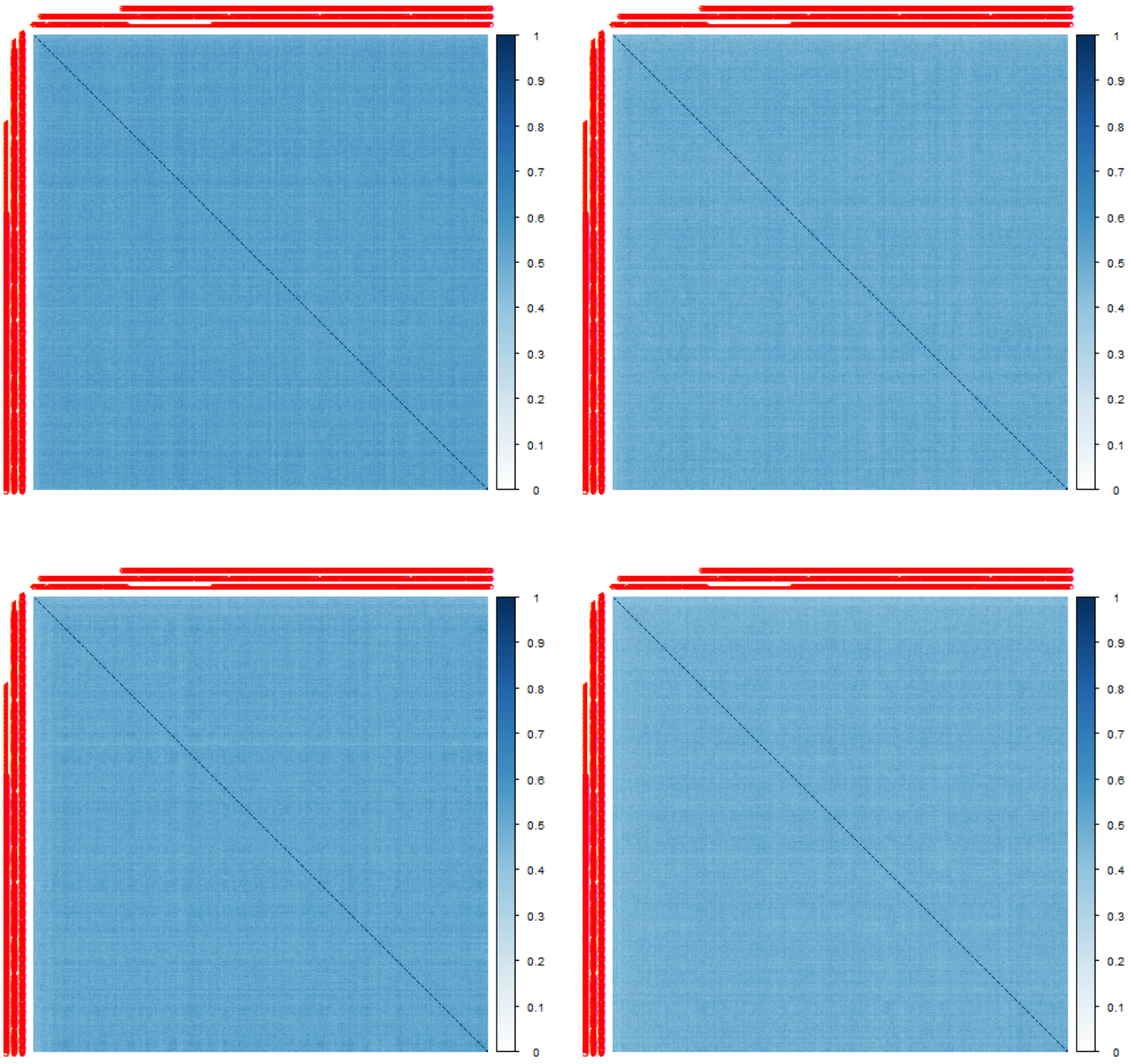


FIGURE 5.8 – Matrice des indices de Rand entre les 1000 partitions deux à deux issues des 1000 tableaux générées lors des 4 première

Annexe F : Distribution des indices de Rand selon le pas pour m=20 tableaux des 4 premières expériences

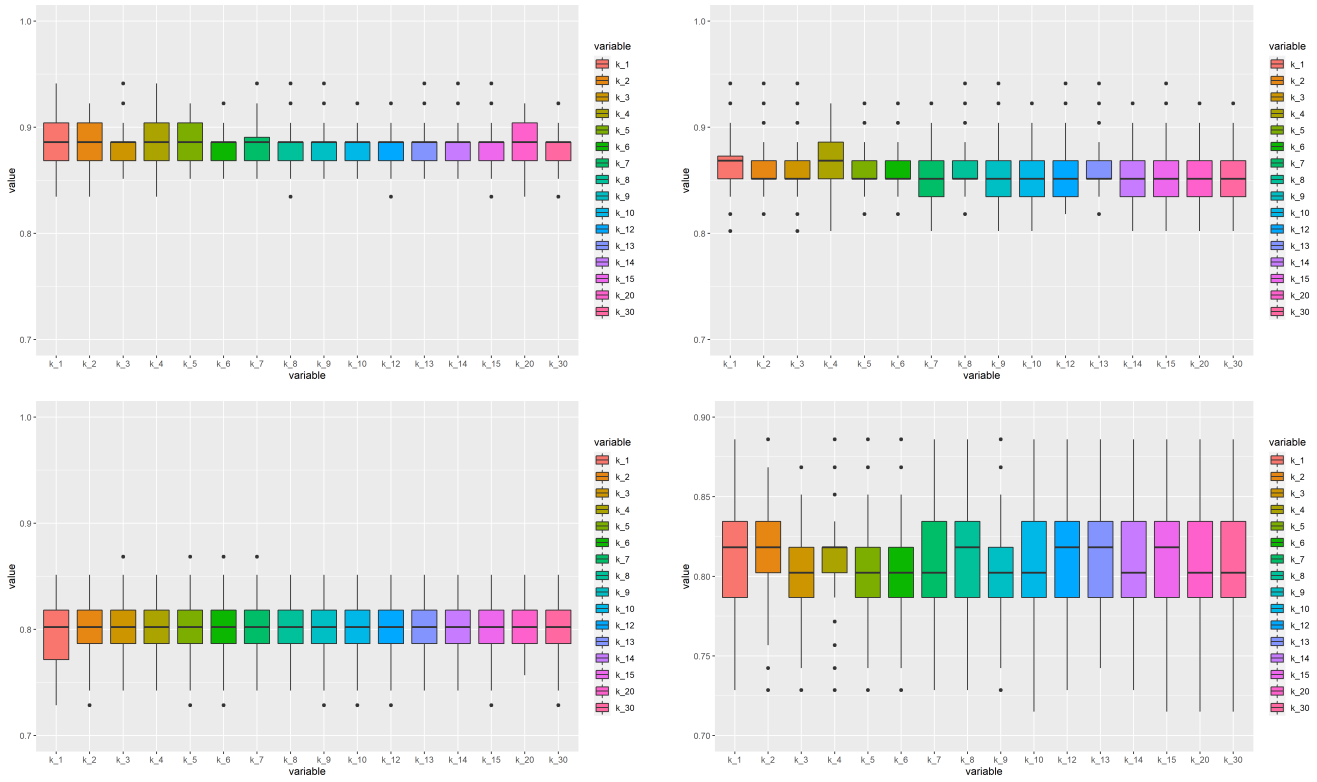


FIGURE 5.9 – Distribution des indices de Rand selon le pas pour m=20 tableaux des 4 premières expériences

Table des matières

REMERCIEMENTS ii

Liste des figures iii

Liste des tables iv

INTRODUCTION GENERALE 1

I Cadre méthodologique 3

1 Méthodes de classification 4
- 1.1 Algorithmes de classification . . . 4
- 1.2 Critère de validité d'une partition . . . 6
  - 1.2.1 Coefficient de silhouette . . . 6
  - 1.2.2 L'indice de Rand . . . 7
    - 1.2.2.1 L'indice de Rand . . . 7
    - 1.2.2.2 L'indice de Rand ajusté . . . 7

2 Consensus de partitions 8
- 2.1 Formulation . . . 9
- 2.2 Consensus simple . . . 10
- 2.3 Consensus pondéré : Weighted NMF . . . 11

3 Imputation multiple 14
- 3.1 Imputation multiple par la méthode JM-DP . . . 16
  - 3.1.1 Modèle de mélanges de lois normales par processus de Dirichlet . . . 16
  - 3.1.2 L'extension du modèle de Dirichlet dans le processus d'imputation multiple . . . 17
- 3.2 Règles de Rubin . . . 18
  - 3.2.1 Consensus de partitions après imputation multiple . . . 18
  - 3.2.2 Calcul de l'instabilité . . . 19
    - 3.2.2.1 Calcul de l'instabilité dans le cadre de données complètes . . . 19
    - 3.2.2.2 Dans le cadre de données incomplètes . . . 19

II Expérimentations et résultats 21

4 Apport du consensus pondéré en présence de données manquantes 22

4.1 Illustration sur l'influence du pas δ en régression 22
4.1.1 Génération des données 22
4.1.2 Plan d'expérimentation 23
4.1.3 Critères d'évaluation 23
4.1.4 Résultats 24
4.1.4.1 Taux de couverture 24
4.1.4.2 Analyse du biais 26
4.2 Extension de l'influence du pas δ en classification 28
4.2.1 Plan de simulation et données 28
4.2.1.1 Génération des données 28
4.2.1.2 Plan d'expérimentation 29
4.2.2 Résultats 30
4.2.2.1 Consensus simple NMF 31
4.2.2.2 Apport du consensus pondéré par rapport au consensus simple 35

5 Application sur des données réelles 41
5.1 Données et problématique 41
5.1.1 Analyse descriptive 42
5.2 Pré-traitement des données 43
5.3 Résultats de l'application sur les données de logement 43
5.3.1 Choix du nombre de classes 43
5.3.2 Résultats de la classification 44
5.3.2.1 Relation entre les partitions obtenues sur les 20 tableaux imputés 44
5.3.2.2 Présentation des résultats 45
5.3.3 Performances des partitions obtenues 47

CONCLUSION 48

BIBLIOGRAPHIES A

ANNEXES I
Annexe A I
Annexe B II
Annexe C III
Annexe D IV
Annexe E V
Annexe E VI

Table des matières VI