UNIVERSITÉ DE MONTRÉAL

TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNEES À LA MAINTENANCE CONDITIONNELLE

ABDERRAZAK BENNANE
DÉPARTEMENT DE MATHÉMATIQUE ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRESENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ES SCIENCES APPLIQUÉES
(GÉNIE INDUSTRIEL)
AOÛT 2010

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNÉES À LA MAINTENANCE CONDITIONNELLE

Présenté par : BENNANE ABDERRAZAK
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. ADJENGULE Luc, Ph. D., président
M^{me} YACOUT Soumaya, D. Sc., membre et directrice de recherche
M. CHINNIAH Yuvin, Ph. D., membre

DÉDICACE

À mes parents.

À mon épouse Bouchra.

À mes enfants Fatima et Yassine.

Je dédie ce modeste travail et je leur dis merci.

« Il me paraît que le seul moyen de faire une méthode instructive et naturelle est de mettre ensemble les choses qui se ressemblent, et de séparer celles qui différent les unes des autres. »

Georges Louis Leclerc de Buffon – 1749

Histoire naturelle, générale et particulière, avec la description du Cabinet du Roy

Premier discours; page21

REMERCIEMENTS

Ce travail n'aurait pas vu le jour sans l'aide et le constant encouragement de mon directeur de recherche Madame Soumaya Yacout. Je la remercie profondément de sa patience, son soutien et ses conseils.

Mes remerciements vont aussi aux professeurs du département de mathématiques et de génie industriel, grâce à qui j'ai acquis de précieuses connaissances lors de mon passage à l'École Polytechnique de Montréal, c'est une expérience des plus enrichissantes de ma vie.

Pendant toutes ces années d'études et pendant toute la durée de mes travaux, il y avait la femme de ma vie qui a cru en moi jusqu'au but, je lui dis : « Merci beaucoup Bouchra! »

RÉSUMÉ

La qualité des données d'apprentissage est une problématique dans de nombreuses applications de classification supervisée, car la qualité de classification de n'importe qu'elle méthode est définie par la qualité des données utilisées dans le processus de traitement c.à.d. à l'entrée du système de classification.

Lors de l'utilisation du classificateur basé sur la méthode d'analyse logique des données (LAD) dans le domaine de la maintenance conditionnelle, il est très fréquent de confronter le problème de données manquantes. Ce phénomène se manifeste lorsque les valeurs n'ont pas pu être observées, elles ont été perdues ou elles n'ont pas été enregistrées. La présence de ces dernières entraîne un dysfonctionnement du processus de traitement logique des données, puisque le classificateur LAD, ne peut pas apprendre à partir des bases de données incomplètes. Si l'on veut l'utiliser, il faut donc adopter une méthode d'imputation de ces données.

En l'absence d'une méthode de traitement des données numériques manquantes pour le classificateur LAD, l'élaboration d'une nouvelle méthode statistique s'avère une alternative très intéressante pour substituer les données manquantes et, par la suite, générer des modèles de classification par LAD.

Dans cette optique, nous proposons dans ce mémoire une méthode statistique de substitution des valeurs manquantes. L'objectif de cette méthode est de remplacer la valeur manquante par les deux possibilités extrêmes que peut prendre cette valeur suivant les valeurs disponibles de la variable en question, et suivant l'information des classes dont on dispose.

Nous avons également mis l'accent sur la validation de notre approche, qui a bénéficié des techniques du test statistique non paramétrique. Cela nous a permis de confirmer les résultats de différents tests de la nouvelle méthode sur des données réelles dans le cadre de trois applications concernant la classification supervisée.

Les travaux présentés dans ce mémoire s'inscrivent dans une thématique de recherche de longue haleine poursuivie au sein de l'équipe de recherche de Mme Soumaya Yacout. Ils font notamment suite aux travaux de thèse de David S. (David S. 2007) axés en particulier sur l'implantation de l'analyse logique des données pour la maintenance conditionnelle.
Le nombre réduit des travaux sur ce sujet nous laisse envisager une suite prometteuse de ce type d'approche.

Organisation du mémoire

Ce mémoire est organisé en deux parties. La première est consacrée aux aspects méthodologiques de la substitution des valeurs manquantes et propose une nouvelle méthode de substitution portant sur les données disponibles et l'information sur les classes des observations dont on dispose. La seconde partie développe les différents tests réalisés pour la validation de la nouvelle méthode de substitution des valeurs manquantes dans le contexte de l'analyse logique des données. Chacune de ces deux parties est divisée en deux chapitres.

Première partie

Le premier chapitre de ce document présente une introduction à la problématique des valeurs manquantes et un état de l'art des méthodes de substitution proposées dans la revue de littérature. Dans le chapitre 2, on aborde la théorie de substitution des valeurs manquantes par MIN-MAX, et on présente la méthodologie de traitement logique des données dans le contexte de la maintenance conditionnelle.

Seconde partie

Au chapitre 3, nous développons l'étude expérimentale et nous présentons les résultats de l'évaluation des performances de la méthode de substitution des données manquantes proposée.
Enfin, dans le chapitre 4, nous présentons une synthèse des travaux réalisés et en dégageons quelques perspectives.

ABSTRACT

The quality of learning data is an issue in a number of applications of monitored classification, as the classifying quality in any method is defined by the quality of the data used in the processing phase; *i.e.* at the entry of classification system.

It is very common to face the issue of missing data when using the classifier based on the logical analysis of data method (LAD). This phenomenon is noticed when values can not be noted, are lost or have not been saved. One of these cases causes a dysfunction of the logical processing of data, as the classifier LAD cannot get its information from incomplete databases. Should we use it, we should adopt a method that removes this data.

In the absence of a method of processing missing digital data for classifier LAD, setting up a new statistical method would appear as a very beneficial alternative to catch up for missing data and then to create classification patterns using LAD.

In this perspective, we propose in this work a statistical method to substitute missing values. The aim of this thesis is to search how to replace the missing value with the two extreme possibilities following the available values of the variable in question, and following the information on the classes available.

We also focus on the validation of our approach that took advantage from techniques of the non-parametrical statistical test. This allowed us to reassert the results of the various tests of the new method on true data as per three applications concerning the monitored classification.

The works presented in this thesis are the outcome of the ambitious research project led by the team of Dr. Soumaya Yacout. They are also the continuation of the works included in the thesis of David S. (David S. 2007) focused on the introduction of the logical analysis of data for conditional maintenance.

The small number of research on this subject makes this kind of approach promising.

Thesis Outline

This thesis consists of two parts. The first part deals with the methodological aspects of the substitution of missing values and proposes a new method of substitution, that is based on available data and the information on the classes available. The second part develops the various tests done to validate the new method of missing values substitution in the context of logical analysis of data. Each of these two parts comprises two chapters.

First part

The first chapter of this thesis introduces the issue of missing values and a literature review of the substitution methods. Chapter two deals with the theory of substitution of missing values by MIN-MAX, and presents a methodology of data logical processing in the context of conditional maintenance.

Second part

Chapter three presents an elaboration of the experimental study and provides the results of the performance evaluation of the proposed method of data substitution.

Finally, Chapter four provides a summary of the works done and draws some conclusions.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

LAD	Logical Analysis of Data – analyse logique des données
LCC	Life Cycle Cost
CGP	Coût global de Possession
CM	Maintenance Corrective
PM	Maintenance Préventive
CH_4	Méthane
H_2	Hydrogène
MCAR	Valeur manquant entièrement au hasard
MAR	Valeur manquante au hasard
NMAR	Valeur ne manquante pas au hasard
CD	Case Deletion ou suppression de cas
MCI	Imputation par la moyenne de la classe
kNNI	La technique de k-plus proche voisin (k-nearest-neighbors imputation),
MMI	Min-Max Imputation
S +	Classe des observations positives
S -	Classe des observations négatives
AIQ	Amplitude interquartile PM
TIC	Technologies de l'Information et des Communications
cbmLAD	Logiciel pour application de LAD a la maintenance conditionnelle
VP	Le taux de vrais positifs
FP	Faux positifs
VN	Vrais négatifs
FN	Faux négatifs
ε	L'ensemble des observations
ε^o	L'ensemble des observations sans valeurs manquantes

ε^m	L'ensemble des observations avec des valeurs manquantes
MS	Marge de séparation
D^+_{min}	Valeur minimale de la fonction discriminante des observations positives
D^-_{max}	Valeur maximale de la fonction discriminante des observations négatives.
ACC	Précision (Accuracy)
C2H6	Ethane
C2H6	Ethylene
C2H2	Acétylène

INTRODUCTION

Dans le domaine de la maintenance conditionnelle, le problème ne réside plus dans l'accès à l'information, mais plutôt dans la qualité de l'information, son filtrage et son prétraitement, puisque la qualité du diagnostic de l'état de l'équipement est définie par la qualité des données utilisées à l'entrée du système du diagnostic [1].

Dans le domaine de la maintenance conditionnelle, il est très fréquent de confronter le problème de données manquantes. Ce phénomène se manifeste lorsque les valeurs n'ont pas pu être observées, elles ont été perdues ou elles n'étaient pas enregistrées. La présence de ces dernières entraîne un dysfonctionnement du processus de traitement logique des données.

Pour la technique de classification appelée analyse logique des données (Logical Analysis of Data, LAD), en l'absence de méthode de traitement des données numériques manquantes, l'élaboration d'une nouvelle méthode de prétraitement des données, apparaît comme une alternative de grand intérêt pour substituer les données manquantes et par la suite, générer des patrons qui mènent à la classification de l'état de l'équipement.

C'est dans ce contexte que s'inscrit notre travail. Nous proposons dans ce mémoire une méthode de substitution des valeurs manquantes. Le but de cette méthode est d'utiliser la structure spéciale de LAD pour remplacer les valeurs manquantes.

Nous considérons ici une méthode de remplacement des données, qui génère des patrons robustes qui peuvent classifier les observations même en l'absence de certaines données.

La validation de notre approche est basée sur des techniques de test statistique non paramétrique, cela nous a permis de confirmer les résultats des différents tests. Ces notions seront développées à travers des applications proposées dans ce document.

Le travail présenté ici s'inscrit dans une thématique de recherche de longue haleine poursuivie au sein de l'équipe de recherche de la Professeure Soumaya Yacout. Ils font notamment suite aux travaux de thèse de David S. (David S. 2007) [2], axés en

particulier sur l'implantation de l'analyse logique des données pour la maintenance conditionnelle.

Le nombre réduit de travaux sur ce sujet nous laisse envisager une suite prometteuse de ce type d'approche.

CHAPITRE 1. GÉNÉRALITÉS ET PROBLÉMATIQUES DE L'ÉTUDE

Ce chapitre aborde le cadre théorique de notre étude après avoir situé le contexte et la problématique.

1.1 Contexte et problématique de l'étude

La présente étude porte sur le traitement des données manquantes en vue de la préparation des données et intervient dans le cadre de l'application de la technique de classification appelée analyse logique des données LAD pour la maintenance conditionnelle. Elle répond d'une part à la question liée au traitement des données manquantes, et d'autre part au besoin de mise en pratique de la théorie de l'analyse logique des données LAD avec des valeurs manquantes.

Pour l'application de la méthode de l'analyse logique des données (LAD), dans le domaine de la maintenance conditionnelle, le service en charge des activités de la maintenance doit recueillir des observations précises et exhaustives pour produire des modèles de classification d'une plus grande précision.

Le classificateur LAD ne peut analyser des bases de données incomplètes. Si l'on veut l'utiliser, il faut adopter une méthode de substitution de ces données qui permet d'obtenir d'excellents résultats de classification.

L'objectif premier de notre étude est donc de proposer une méthodologie adaptée de substitution, afin de réduire le biais introduit par la présence de données problématiques (manquantes/ aberrantes) pour certaines observations. Mais avant de proposer notre méthode pour réaliser cet objectif, nous allons voir une brève présentation de la fonction maintenance et une introduction à la problématique de valeurs manquantes et aberrantes.

1.2 Les différents concepts de la maintenance

Actuellement, la maintenance conditionnelle s'impose comme la meilleure solution permettant d'accroître les performances et d'améliorer le niveau de sûreté de fonctionnement de tout système industriel. Elle permet d'assurer la pérennité des

équipements et de veiller à ce que le système ne tombe pas en panne, ce qui permet de maintenir le fonctionnement de l'appareil de production et de garder le seuil de productivité à un niveau stable.

Depuis les premiers travaux de Barlow et Hunter, (1960) [3], un grand nombre de politiques de maintenance ont été développées et mises en application, pour améliorer la disponibilité des systèmes, pour empêcher leurs défaillances et pour réduire leurs coûts de cycle de vie (communément connu LCC pour Life Cycle Cost ou CGP pour Coût global de Possession). Ces politiques peuvent être regroupées en deux grandes familles : la Maintenance Corrective (CM) et la Maintenance Préventive (PM).

Une synthèse des politiques de maintenance préventive est donnée par Simmons et Pollock, (2005), Kobbacy et Jeon, (2002) [4-5].

Dans la figure 1.1, nous présentons les différents concepts suivant le type de maintenance étudié. Alors que la mise en place d'opérations correctives ne dépend que de l'occurrence d'une panne, les maintenances préventives peuvent être programmées en fonction de différents paramètres.

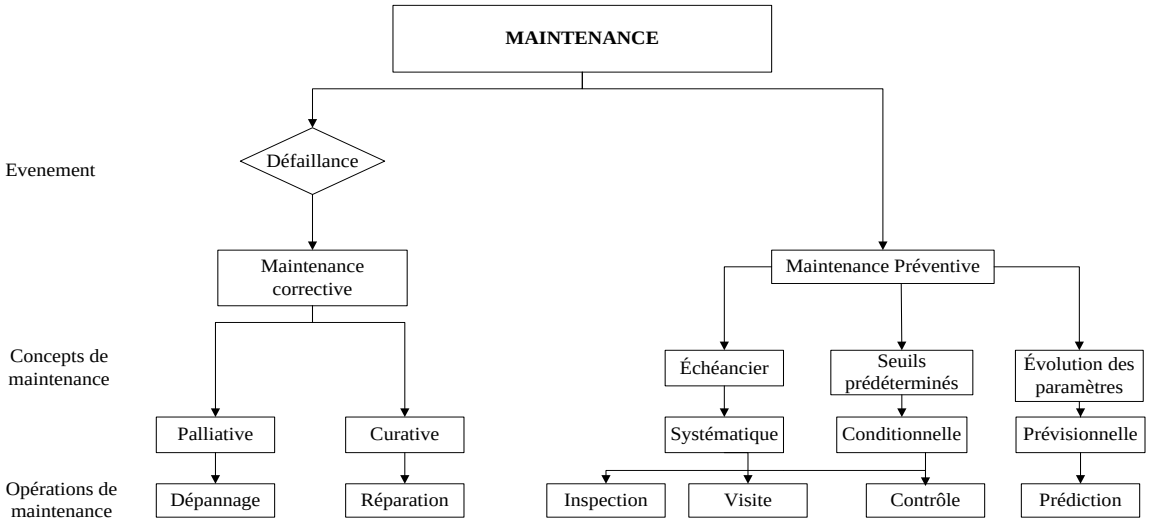


Figure 1.1- Différents concepts de maintenance

Il existe deux façons complémentaires d'organiser les actions de maintenance (Figure 1.1) :

A- Maintenance préventive : maintenance exécutée à des échéanciers prédéterminés ou selon des critères prescrits (seuils prédéterminés), afin de réduire la probabilité de défaillance ou la dégradation du fonctionnement d'un bien. Cette maintenance préventive peut être décomposée à son tour en :

   a. **Maintenance systématique :** maintenance préventive effectuée systématiquement, soit selon un calendrier (à périodicité temporelle fixe), soit selon une périodicité d'usage (heures de fonctionnement, nombre d'unités produites, nombre de mouvements effectués, etc.).
   b. **Maintenance conditionnelle :** maintenance préventive réalisée à la suite de relevés, de mesures, de contrôles révélateurs de l'état de dégradation de l'équipement.
   c. **Maintenance prévisionnelle :** maintenance réalisée à la suite d'une analyse des prévisions extrapolées de l'évolution de paramètres significatifs de la dégradation du bien.

B- Maintenance corrective : Elle consiste à intervenir sur un équipement une fois que celui-ci est défaillant. Cette maintenance corrective peut être décomposée encore en :

   a. **Maintenance palliative :** Dépannage (donc provisoire) de l'équipement, afin de permettre la continuité de l'exploitation du bien sans pour autant traiter les causes ; elle doit toutefois être suivie d'une action curative dans les plus brefs délais.
   b. **Maintenance curative :** Il s'agit là d'une maintenance qui s'attaque réellement au fond du problème en essayant de faire une réparation (donc durable) consistant en une remise à l'état initial.

1.3 Impact économique de la maintenance sur la performance globale de l'entreprise

Présentant pour certains secteurs d'industrie jusqu'à 60 % des coûts de transformation, la maintenance est une fonction dont les coûts doivent être non seulement suivis, mais aussi analysés et optimisés. En effet, les coûts de la maintenance constituent la majeure partie des coûts opérationnels, que ce soit dans le secteur manufacturier ou dans l'industrie de procédés.

Selon R. Keith Mobley (1989) [6], aux États-Unis seulement, plus de 200 milliards, de dollars par année va à l'entretien des installations et des équipements industriels, et le résultat de la mauvaise gestion de la maintenance constitue une perte de plus de 60 milliards de dollars par année. On y indique également que les coûts de la maintenance peuvent représenter environ 15 % du coût du produit fini pour les industries agro-alimentaires, tandis que dans les pâtes et papier, le fer et l'acier, et d'autres industries lourdes, ces coûts peuvent frôler 60 % du coût total de la production.

La maintenance conditionnelle est utilisée pour réduire les coûts d'entretien; ainsi, la réduction du nombre d'actions de maintenance corrective (diminutions de pièces de rechange et les coûts de main-d'œuvre), et la bonne planification de l'entretien préventif (augmente la disponibilité des articles et des avoirs) se traduiront par une diminution du budget d'entretien et par l'augmentation de la sécurité industrielle (Certains risques d'accident peuvent être provoqués par un équipement non contrôlé et non entretenu, comme les risques d'explosion, d'incendie, de déversement et fuites de produits dangereux).

Plusieurs recherches sur les coûts d'entretien, les budgets, et les économies potentielles ont été effectuées. Alsyouf, (2004) et Fararooy et Allan [7-8].

1.4 L'utilisation des techniques de surveillance et de diagnostique pour la fonction maintenance

Dans un grand nombre d'applications industrielles, on constate une demande croissante en matière de remplacement des politiques de maintenance curative par des stratégies de

maintenance préventive. Cette mutation d'une situation où on « subit les pannes » à une situation où on « maîtrise les pannes », nécessite quelques moyens technologiques ainsi que la connaissance des techniques d'analyse appropriées. La fonction surveillance en continu de l'évolution de l'équipement à travers des données quantifiables et qualifiables permet ainsi de prévenir tout dysfonctionnement et d'éviter les fausses alarmes qui peuvent ralentir la production. Basseville et *al*,(1996) [9].

Nous présentons à la figure 1.2 les techniques les plus courantes en surveillance d´équipements industriels définies par Nicolas PALLUAT (2006) [10].

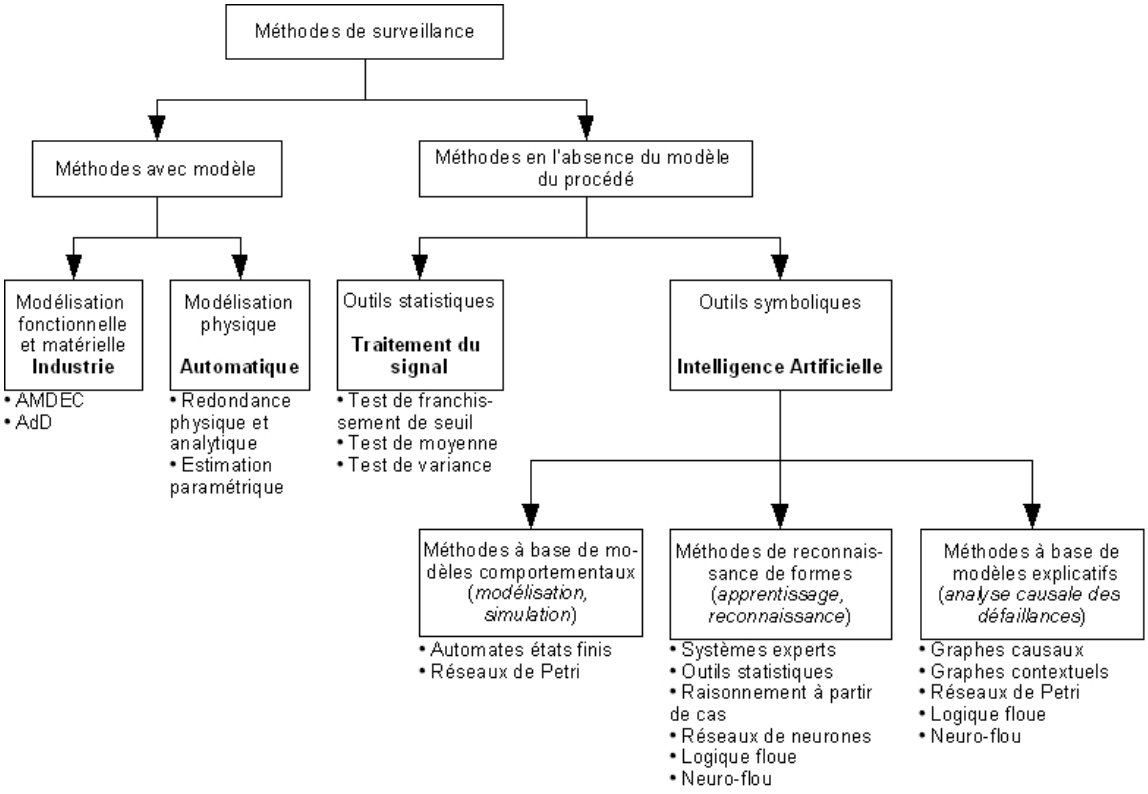


Figure 1.2- Différentes méthodes de surveillance industrielle Nicolas PALLUAT (2006).

Selon la Figure 1.2, il existe deux méthodes de surveillance industrielle :

A- **Méthode avec modèle** : Elle se base sur un modèle formel ou mathématique du fonctionnement de l'équipement et elle se subdivise à son tour en :

a. **Méthodes par modélisation fonctionnelle et matérielle :** Le principe consiste à établir à priori et de la manière la plus complète possible, les liens entre les causes initiales des défaillances et leurs effets mesurables.
b. **Méthodes par modélisation physique :** elles se basent sur la redondance physique et analytique ainsi que l'estimation paramétrique, elles ont pour principe de comparer les mesures effectuées sur le système aux informations fournies par le modèle.

B- **Méthode en l'absence d'un modèle** : Utilisée pour les applications industrielles dont le modèle est difficile, voire impossible à obtenir et elle se subdivise à son tour en :

a. Méthodes utilisant des outils statistiques,
b. Méthodes symboliques de l'Intelligence Artificielle.

1.5 L'utilité de traitement des données

D'après Elizabeth Vannan (2001) [11], les données ne sont pas supposées être parfaites, mais précises, complètes, consistantes, cohérentes, opportunes, et flexibles et ceci dans le but de satisfaire des besoins**.**

- ***Précises*** : les données ne doivent pas contenir des erreurs ;
- ***Complètes*** : toutes les valeurs doivent être présentes ;
- ***Cohérentes*** : les données doivent satisfaire un ensemble de contraintes et doivent représenter le phénomène qu'elles mesurent ;
- ***Opportunes*** : les données doivent être disponibles quand elles sont réclamées ;
- ***Flexibles*** : les données doivent être décrites de façon à ce qu'elles peuvent être analysées de différentes méthodes.

Selon Paul Jermyn (1999) [12], le traitement des données représente environ 60 à 80 % du temps impliqué dans le processus d'extraction de l'information; la difficulté dépend de la nature des problèmes existants avec les données à savoir:

- Les données erronées : par exemple une valeur quantitative à la place d'une valeur qualitative.

- Les données manquantes : absence de données / valeurs.
- Les données aberrantes : les données qui s'écartant de la norme, du modèle normal
- Les données doublons : données faussement dupliquées.

Donc le traitement des données doit être réalisé pour les raisons suivantes :

- La résolution des problèmes qui peuvent nous empêcher d'appliquer une technique de diagnostic.
- La compréhension de la nature des données pour pouvoir les analyser.
- L'extraction plus efficace de l'information.

A partir de ce qui a été mentionné ci-dessus, notre problématique dans ce chapitre est de déterminer les démarches et les techniques de traitement des données manquantes et aberrantes afin de bien préparer les données pour une exploitation et une extraction efficace de l'information, puisque la qualité des résultats finaux est largement conditionnée par le soin porté à cette première étape.

1.6 Cadre théorique

Dans cette section, il sera question pour nous de présenter la littérature sur les différentes théories concernant le traitement de données manquantes.

Peu importe la rigueur que l'on se fixe, il y aura toujours des données problématiques, comme le soulignaient BRION P. et CLAIRIN Rémy (1997) [13], et il faut faire avec en trouvant une méthode robuste pour leur traitement. A cet effet, il convient alors de connaître les méthodes qui réduisent et affaiblissent l'effet des valeurs manquantes sur le résultat.

Avant de définir les méthodes appropriées aux traitements de ce phénomène, il nous paraît nécessaire de donner la définition des données aberrantes et de faire une distinction entre les différentes formes de données manquantes. Quand parle-t- on d'une valeur manquante totale ou partielle ? Cette distinction sera suivie par une description des types de mécanismes des valeurs manquantes.

1.6.1 Données aberrantes

Les données aberrantes sont définies comme des données qui ne sont pas en accord avec la majorité des données, selon Chiang et *al*, (2003) [14]**,** toutefois on peut dire que les données aberrantes se trouveront à la périphérique du nuage formé par l'ensemble des données, suivant la définition de Grubbs (1969) [15], « *an outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs* » .

Les données aberrantes pouvant être causées soit par une raison physique connue, exemple le mauvais étalonnage de l`appareil de mesure, erreur d'écriture de données, soit par une raison non connue.

Pour repérer les données aberrantes, un contrôle de cohérence sur les données est nécessaire, l'usage du bon sens et de l'expérience est le plus sûr. Par exemple, des données de température d'un roulement, très élevé ou très faible, en tout cas très différent de la grande majorité des valeurs, doivent conduire à s'interroger sur la validité de ces données, c'est pour cette raison que la réflexion sur la cohérence des données est essentielle.

L'exemple suivant (tableau 1.1 et 1.2) semble un peu extrême, mais cette situation peut être rencontrée au moment du traitement des données.

Tableau 1.1- Exemple sans valeur aberrante

Observation	Attribut a_1
w_1	1
w_2	3
w_3	5
w_4	9
w_5	12

Tableau 1.2- Exemple avec valeur aberrante

Observation	Attribut a_1*
w_1	1
w_2	3
w_3	5
w_4	9
w_5	**120**

Les quatre premières valeurs dans chaque colonne contiennent les mêmes mesures.
Toutefois, dans le tableau 1.2, la cinquième entrée de l'attribut (a_1*) a un grand écart par rapport à la valeur dans le tableau 1.1.
À noter que dans la présence d'une mesure aberrante, la médiane des données ne change pas.
La médiane est robuste (généralement, il ne varie pas beaucoup) en présence d'un petit nombre de valeurs aberrantes, par contre la moyenne change rapidement.

1.6.2 Données manquantes

Une donnée incomplète est une donnée pour laquelle la valeur de certain attribut est inconnue, ces valeurs sont dites manquantes.
Soit l'observation est un vecteur des valeurs de certains indicateurs ou attributs, les valeurs manquantes peuvent être de deux natures :

- Valeur manquante totale, c'est- à-dire que toute l'observation manque.
- Valeur manquante partielle, c'est-à-dire que l'observation est présente mais il manque certaines valeurs de cette observation.

Exemple 1 : Dans le tableau 1.3, la valeur de l'attribut (a_4) pour l'observation (w_2) est manquante, la valeur n'est pas présentée et l'observation (w_2) est dite incomplète.

Tableau 1.3- Valeur manquante et donnée incomplète

	Attributs			
Observations	a_1	a_2	a_3	a_4
w_1	56	98	10	5
w_2	40	87	21	?

Des valeurs manquent parce qu'elles n'ont pas pu être observées; elles ont été perdues ou elles n'étaient pas enregistrées.

1.6.3 Mécanisme des valeurs manquantes

Avant de commencer le traitement de cette problématique, il faut évaluer le mécanisme des valeurs manquantes et ensuite faire le choix de la méthode de traitement.

Selon Little, R.J.A., and Rubin, D.B. (2002) [16]**,** il y a trois mécanismes distincts de valeurs manquantes :

- Valeur manquante entièrement au hasard (MCAR, Missing Completely at random) : le fait de ne pas avoir la valeur pour une variable, est indépendant des autres variables.
- Valeur manquante au hasard (MAR, Missing at random) : le fait de ne pas avoir la valeur pour une variable, est dépendant seulement des valeurs observées.
- Valeur ne manquant pas au hasard (NMAR, Non missing at random) : le fait de ne pas avoir la valeur pour une variable ne dépendant que des valeurs manquantes.

Exemple :

Supposons que le niveau de Méthane **CH_4** et d'Hydrogène **H_2**,dans l'huile de refroidissement de plusieurs transformateurs est mesuré en janvier **H_2**(janvier), d'autres prélèvements ont été faits sur certains transformateurs une deuxième fois en février **H_2**(février)**,** mais pas sur la totalité.

Le tableau 1.4 montre les données simulées pour 15 transformateurs. Les deux premières colonnes du tableau indiquent les données complètes pour **H_2** et **CH_4**, les autres colonnes indiquent les valeurs **H_2** (février)**,** qui restent suivant les trois mécanismes des valeurs manquantes.

Dans le premier mécanisme, les 15 mesures ont été choisies au hasard parmi celles qui sont mesurées en janvier ; ce mécanisme est **MCAR**. Dans le deuxième mécanisme,

ceux qui sont mesurées en février ont été sélectionnées parce que la mesure de **CH_4** de même observation en janvier dépasse la valeur 300 ppm (300 < **CH_4**) ce mécanisme est **MAR**.

Dans le troisième mécanisme, les mesures réalisées en février et qui dépassent la valeur 300 ppm (**H_2** (février) > 300) ne sont pas indiquées sur le tableau. Cela pourrait se produire, par exemple, si on a fait les mesures sur tous les transformateurs, mais la personne responsable a décidé d'enregistrer la valeur de février comme si elle est dans une fourchette de valeur prédéterminée. Ce troisième mécanisme est un exemple de **MNAR**.

Un autre cas possible de **MNAR**, par exemple, la mesure de février ne doit être enregistrée que si elle est sensiblement différente de la mesure de janvier.)

Tableau 1.4- Simulation des trois mécanismes des valeurs manquantes

	Attributs					
Observations	**CH4**	**H2 (janvier)**	**H2 (février)**			
			Complete	**MCAR**	**MAR**	**MNAR**
1	13	24	33	33	---	---
2	24	127	127	---	---	---
3	4066	9474	100	100	100	---
4	1053	507	233	233	233	---
5	695	416	411	411	411	411
6	207	441	566	---	---	566
7	61	65	100	100	---	---
8	87	16	78	---	---	200
9	38	212	855	---	---	855
10	1393	800	50	50	50	---
11	770	199	200	---	200	---
12	17424	425	588	---	588	588
13	95	1076	800	---	---	800
14	754	244	52	---	52	---
15	107	127	30	---	---	---

Selon Simon et Smonoff (1986) et Little (1988) [17-18], il est difficile de traiter des données de type MNAR et MAR, ce qui nous incite à faire l'hypothèse que le manque de données est de nature MCAR, dans la pratique, et comme l'expliquent Schafer et

Graham (2002) [19], il est quasiment impossible de déterminer lequel des trois mécanismes est à l'œuvre à partir des données.

1.6.4 Revue de littérature des méthodes de traitement des données manquantes

De nombreuses techniques de traitement des données manquantes ont été développées dans les années 90, Hu et *al.* (2000) [20], sans prétendre être exhaustifs, en identifiaient déjà plus d'une vingtaine, pour la plupart issues des recherches en statistique. Depuis, les chercheurs en intelligence artificielle et fouille de données (Data mining), se sont mis à étudier la question et à développer de nouvelles techniques. Recenser l'ensemble de ces techniques serait fastidieux. Aussi avons-nous opté pour une mise en évidence des principales caractéristiques des différentes méthodes. Nous pouvons alors présenter les techniques les plus usitées et avoir ainsi une vue d'ensemble du domaine; ces méthodes s'appliquent selon la nature du processus et parfois compte tenu du nombre d'observations.

Selon Kline (1998) [21], Song et Shepperd, (2007) [22], il y a trois stratégies possibles pour traiter des données manquantes :

(a) Utilisation des procédures de suppression, présentée à la section 1.6.4.1.

(b) Utilisation des procédures de remplacement, (substitution) les données manquantes par les valeurs présentes, 1.6.4.2.

(c) Utilisation des procédures de modélisation de la distribution des données manquantes et les estimés par certains paramètres, 1.6.4.3.

Cette première introduction conduit à la typologie de la figure 1.3. Nous allons maintenant nous focaliser sur les techniques de substitution correspondant à la stratégie (b), en essayant de dégager les caractéristiques qui permettent de les différencier. La technique étiquetée CD, pour Case Deletion ou suppression de cas, correspond à la stratégie (a) dans laquelle on se ramène à une base de données complète par suppression de toutes les observations contenant au moins une valeur manquante.

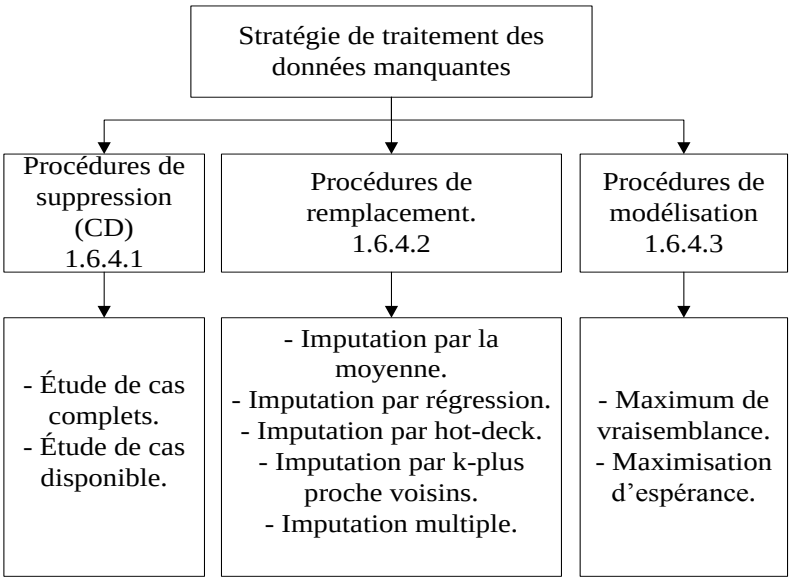


Figure 1.3- Les grandes catégories des méthodes pour le traitement des données manquantes

Chacune de ces familles des techniques est discutée ci-dessous.

1.6.4.1 Procédures de suppression

A. Étude des cas complets (Listwise deletion)

Cette méthode permet de se ramener à une base de données complète par réduction de la dimension du problème. Pour cela, tous les exemples de la base contenant des valeurs manquantes sont supprimés (On peut également choisir de supprimer toutes les variables dont certaines observations manquent, mais il faut être prudent car certaines peuvent être essentielles pour l'analyse). Par conséquent, cette méthode sacrifie un grand nombre de données (Malhotra, 1987) [23]. Selon Kim et Curry (1977) [24], la suppression de 10 % des données de chaque variable dans une matrice de cinq variables peut facilement provoquer l'élimination de 59 % des observations de l'analyse. Kaufman (1988) [25], rapporte qu'il a vu une baisse de dimension de l'échantillon de 624 à 201 avec l'utilisation de la méthode de suppression listwise.

Les techniques statistiques d'analyse des données ayant besoin d'un nombre suffisant d'observations pour que leurs résultats soient valides. Dans des cas qui ne sont pas rares,

où la quasi-totalité des exemples possède des valeurs manquantes, elle devient même inutilisable.

D'autre part, les statistiques, telles que la moyenne ou la variance, seront fortement biaisées, à moins que le mécanisme de génération des données ne soit complètement aléatoire (MCAR) (Magnani, 2003) [26].

Malgré le fait que la grande perte de données réduit la puissance et l'exactitude statistiques (Little et Rubin, 2002) [16], cette technique, du fait de sa simplicité est fréquemment l'option implicite pour l'analyse dans la plupart des progiciels statistiques.

Cette méthode est aussi appelée par certains auteurs (Schafer, 1997; Little et Rubin 2002) [27-16], la méthode d'analyse des données disponibles (*available-case analysis*).

B. Étude des cas disponibles (Pairwise deletion)

Dans cette méthode, on ne considère que les cas où ces variables sont complètement observées. Par exemple, si la valeur de l'attribut A est absente pour une observation, les autres valeurs pour le restes des attributs de la même observation pourraient encore être employées pour calculer des corrélations, telles que celle entre les attributs B et C.

Comparée à la premier méthode (Étude des cas complets), selon Roth, (1994) [28], cette méthode conserve beaucoup plus de données qui auraient été perdues si on employait la méthode d'étude des cas complets .

Il s'agit d'une autre méthode proposée par les logiciels statistiques, mais généralement problématique : le nombre d'observations (n) varie pour le calcul de chaque valeur de la nouvelle base de données, le risque d'obtenir une base de données réduite est grand et, encore une fois, la représentativité sera biaisée si les données manquantes ne sont pas distribuées de façon complètement aléatoire.

Les études de Monte Carlo ont montré que la suppression par la méthode **Listwise deletion** donne des évaluations moins précises des paramètres d'estimation (Gleason et Staelin, 1975, Kim et Curry, 1977, Malhotra, 1987, Raymond, 1986 et Raymond et Roberts, 1987) [29-21-23-30-31]. La méthode **Pairwise deletion** est uniformément plus précise (Gleason et Staelin, 1975, Kim et Curry, 1977 et Raymond, 1986) [29-21-30], bien que les différences puissent parfois être minimes (Raymond, 1986) [30].

Toutefois, il existe certaines raisons (Allison 2001) [32], (Little et Rubin. 2002) [16], pour la considérer une bonne méthode CD doit être appliquée uniquement dans les cas où le nombre de valeurs manquantes est relativement faible (Rubin 1987) [33].

De plus, il paraît que, même si les données sont manquantes selon un mécanisme complètement aléatoire, les méthodes qui utilisent l'ensemble de l'information contenue dans la matrice de données sont plus efficaces que les méthodes basées sur les données complètes (Little et Rubin, 2002) [16]. .

1.6.4.2 Procédures de remplacement

Ces procédures visent à se ramener à une base complète en trouvant un moyen adéquat de remplacer les valeurs manquantes. On nomme ce procédé imputation, complétion ou substitution.

De façon générale, des procédures de remplacement peuvent être employées dans certains cas, tant qu'on a une bonne raison pour remplacer.

Généralement il est facile d'exécuter les procédures de remplacement, et certaines sont incluses comme options avec les logiciels statistiques. Les avantages les plus importants de ces procédures sont la conservation de la dimension de la base de données, par conséquent, de la puissance statistique d'analyses. Dans une plus ou moins large mesure, toutes les procédures de remplacement sont décentrées s'il y a une distribution non-aléatoire des valeurs manquantes. Cependant, le remplacement des données manquantes est approprié quand les corrélations entre les variables sont faibles (Little et Rubin, 2002) [16], (Quinten et Raaijmakers, 1999) [34]. .

Différentes procédures de remplacement de données manquantes ont été élaborées au cours des années. Généralement on constate que les différences entre les diverses méthodes diminuent avec:

(a) une plus grande dimension de la base de données,

(b) un plus petit pourcentage des valeurs manquantes et

(c) une diminution au niveau des corrélations entre les attributs (Raymond, 1986) [30].

Cependant, Kromrey et Heines (1994) [35], ont rapporté qu'il y a une différence entre les méthodes de remplacement des données manquantes si les effets des traitements sur les statistiques analytiques sont pris en considération. Avec de plus grandes dimensions

de la base de données, en fait, les différences entre les diverses procédures de remplacement augmentent ; ceci fournit davantage d'évidence qu'en évaluant l'efficacité des traitements des données manquantes, l'exactitude d'estimer la valeur des données manquantes et l'exactitude d'estimer les effets statistiques doivent être considérées.
Trois types de procédures de remplacement peuvent être distingués : imputation basée sur la moyenne, basée sur la régression et le hot-deck imputation.

A. Imputation par la moyenne

Les valeurs manquantes de chaque attribut sont remplacées par la moyenne de l'attribut considéré. Il y a deux variantes de l'imputation par la moyenne : Imputation par la moyenne totale, imputation par la moyenne de sous-groupe. Pour l'imputation par la moyenne totale, la valeur absente d'un attribut est remplacée par la moyenne des valeurs de cet attribut de toutes les observations. Pour l'imputation par la moyenne de sous-groupe (classe), la valeur manquante est remplacée par la moyenne du sous-groupe (classe) de l'attribut en question.
L'inconvénient de cette méthode est la sous-estimation de la variance et de biaiser la corrélation entre les attributs, cela veut dire que la distribution des données est loin d'être préservée.
Selon Pigott (2001) [36], cette manière de procéder serait encore moins recommandable que l'utilisation de la méthode d'étude des cas complets. Il est attendu que, même si les données sont manquantes selon un mécanisme complètement aléatoire, l'estimé de la moyenne de la distribution sera valide, mais, par contre, l'estimé de l'écart type s'avère automatiquement biaisé (Little et Rubin, 2002) [16]. En remplaçant les valeurs manquantes par une valeur constante, la variance de l'attribut s'avère inévitablement réduite. Il s'en dégage que la valeur de l'erreur type, diminuée par la réduction de la variance et par l'augmentation de la taille de l'échantillon, sera artificiellement plus petite que ce qui aurait dû être observé (Pigott, 2001) [36].
Les études ont été quelque peu concluantes concernant l'efficacité de la substitution par la moyenne. Selon Kim et Curry (1977) [24], la substitution par la moyenne est moins précise que la méthode **Listwise deletion**, alors que d'autres, ont prouvé que la substitution par la moyenne est plus précise que le **Listwise deletion** et le **Pairwise**

deletion (Chan et Dunn, 1972, Chan et al, 1976 et Raymond et Roberts, 1987) [37-38-31].

Dans le contexte de la classification, comme celui que nous étudierons en section au chapitre deux, il est très intéressant d'utiliser la moyenne relative à chaque classe et non pas la moyenne de l'ensemble des observations, les classes sont connues à l'avance (classification supervisée).

Dans notre cas, la méthode d'imputation par la moyenne de la classe est appelée **MIC**, le C indique que l'on tient compte de l'information de la classe.

B. Imputation par régression

C'est une approche en deux étapes : d'abord, on estime les rapports entre les attributs, et puis on emploie les coefficients de régression pour estimer la valeur manquante (Frane, 1976) [39]. La condition fondamentale de l'utilisation de l'imputation par régression est l'existence d'une corrélation linéaire entre les attributs. La technique suppose également que les valeurs sont manquantes au hasard.

Dans le contexte des valeurs manquantes, deux modèles de régression sont en général employés : la régression linéaire et la régression logistique. Cette dernière est plutôt utilisée pour traiter les variables discrètes, alors que la régression linéaire est appliquée sur des variables continues (Little et Rubin, 2002) [16].

Pour chacune de ces méthodes, il est possible de tenir compte de l'information de classe en n'utilisant que les observations d'une même classe pour estimer les paramètres de régression.

L'inconvénient de cette méthode, c'est les hypothèses qui sont faites sur la distribution des données. Supposer une relation linéaire entre les variables, revient à faire des hypothèses qui sont rarement vérifiées, dans cette situation, le remplacement des valeurs manquantes par des valeurs prédites basées sur un modèle biaisé ne constitue pas un traitement approprié.

Ces méthodes, seraient beaucoup plus efficaces, exclusivement dans le cas où le modèle de régression est adéquat (Sinharay, Stern, et Russel, 2001) [40].

C. Imputation par hot-deck

L'imputation hot-deck est une procédure qui consiste à remplacer les valeurs manquantes d'une observation par des valeurs empruntées à d'autres observations similaires, définies comme étant celles pour lesquelles les valeurs sont les plus identiques à celles de l'observation présentant une donnée manquante, l'hypothèse sur laquelle elle s'appuie est que les probabilités de présence des valeurs sont égales dans les cas d'imputation.
Même si ce type de méthodes préserve les distributions des variables, elles risquent d'altérer les relations entre les variables (Sinharay, Stern et Russel, 2001) [40].

D. Imputation par k-plus proches voisins

La technique de k-plus proche voisin (kNNI - k-nearest-neighbors imputation), (Chen and Shao 2000; Engels and Diehr 2003; Zhang 2008; Zhang et *al*. 2008) [41-42-43-44], est une technique utiliser pour la substitution des valeurs manquante, avec la valeur du plus proche voisin dans l'ensemble de données.
Pour chaque observation contenant des valeurs manquantes, on recherche ses k plus proches voisines. Dans le cas de variables continues, la valeur de remplacement correspond simplement à une moyenne pondérée des valeurs prises par ces k voisins pour la variable en question.
La difficulté réside dans le choix du paramètre k et de la métrique utilisée, les distances les plus utilisées étant l'euclidienne, celle de Mahalanobis ou encore celle de Pearson.
Dans notre cas particulier, cette technique se place dans un contexte d'apprentissage supervisé et dispose donc d'une variable classe, pour calculer la distance entre chaque observation contenant une valeur manquante et chacune des classes. Les k plus proches voisins de l'observation considérée, parmi ceux qui appartiennent à la même classe, sont alors utilisés pour déterminer la valeur de remplacement. La version de cette méthode, qui tient compte de l'information de la variable classe, est utilisée par Song et Shepperd (2007) [22], sous le nom de MINI.
Pour parvenir à l'application de cette méthode dans notre cas, nous proposons d'utiliser une version adaptée, proposée par Zhang, S. (2008) [43], que nous notons NNI avec k=1, qui tient compte de la valeur à gauche et la valeur à droite, les plus proches voisins d'une valeur manquante, alors que la méthode kNNI choisit k plus proches voisins.

Chaque valeur manquante est remplacée par la moyenne arithmétique des plus proches voisins existants ci-dessus et / ou en dessous de la valeur manquante dans la même classe (S + ou S-).

Les trois cas possibles sont les suivants:

• Si la valeur est située entre deux valeurs existantes, la valeur manquante est remplacée par la moyenne de ces deux valeurs.

• Si la valeur est située au début d'une classe, alors elle sera remplacée par la plus proche valeur inférieure.

• Si la valeur est située à l'extrémité d'une classe, alors elle sera remplacée par la plus proche valeur supérieure.

L'avantage de cette méthode est de ne faire aucune supposition quant à la distribution des données, et de prendre en considération la corrélation entre variables.

E. Imputation multiple

Cette méthode est suggérée par Rubin (1978) [45], cela fait plus de 30 ans, et décrite en détail par Rubin (1987) [33] et Schafer (1997) [27]. Pour des références sur l'imputation multiple, il y a celles de Rubin (1996) [46], ainsi que Rubin et Schenker (1986) [47]. Dans le but de prédire une valeur pour toute donnée manquante, l'imputation multiple, au lieu de procurer une seule matrice à analyser, va à la place produire *m* matrices de données plausibles. Ces *m* matrices (souvent cinq suffisent) contiennent les mêmes données observées, mais les valeurs pour les données prédites peuvent être différentes. Cette variabilité entre les valeurs prédites des *m* matrices reflète l'incertitude face à l'imputation (Fichman, et Cummings, 2003) [48]. Ces matrices de données sont ensuite analysées comme si elles étaient des bases de données complètes, et elles sont combinées dans une unique base de données récapitulative. Des exemples typiques d'imputation par cette méthode sont traités par Rubin (1981) [49] et Freund (1995) [50]. Selon certains chercheurs, il apparaît que l'imputation multiple serait robuste même lorsque les données sont manquantes selon un mécanisme non aléatoire. Puisqu'elle représente la façon de faire la plus prometteuse (Schafer et Graham, 2002, Little et Rubin, 2002) [19-16] et qu'elle est disponible sur certains logiciels, elle semble

constituer une méthode préférable à celles présentées précédemment. Dans tous les cas, Fichman et Cummings (2003) [48], soulignent qu'elle est préférable à la méthode listwise deletion, mais son inconvénient c'est la complexité de calcul des *m* matrices (contraintes d'espace mémoire et de temps de traitement) ainsi que l'obligation de réaliser une analyse statistique pour chaque matrice.

1.6.4.3 Procédures basées sur un modèle

A. Maximum de vraisemblance

Sous sa forme plus simple, l'approche de maximum de vraisemblance pour analyser des données manquantes, suppose que les données observées sont tirées d'une distribution normale multivariée (DeSarbo et *al*, 1986) [51], Lee et Chiu (1990) [52]. Au lieu d'imputer des valeurs aux données manquantes, ces méthodes définissent un modèle à partir des données disponibles et basent les inférences de ce modèle sur la vraisemblance de la distribution des données sous ce modèle.

B. Maximisation d'espérance

Une dernière approche assez fréquente consiste à utiliser l'algorithme de maximisation d'espérance EM (Expectation-Maximization) pour estimer les valeurs manquantes (Dempster et *al*, 1977; Ghahramani et Jordan, 1994; Little et Rubin, 2002) [53-54-16], qui est un processus itératif (Laird, 1988 et Ruud, 1991) [55-56]. Il est généralement utilisé pour estimer les paramètres d'une densité de probabilité. Il peut être appliqué sur des bases de données incomplètes, et présente l'avantage de procéder à l'estimation des valeurs manquantes en parallèle de l'estimation des paramètres.

On suppose l'existence d'un modèle de génération des données, par exemple un mélange de gaussiennes pour les variables continues. Les paramètres du modèle sont calculés suivant la méthode du maximum de vraisemblance, de manière itérative.Zou et *al*. (2005) [57] ont présenté une nouvelle méthode de substitution basée sur une version simplifiée d'EM.

À partir d'une estimation par défaut des valeurs manquantes, les paramètres du modèle sont ré-estimés, à chaque itération, à partir de la matrice complète, de manière à

accroître la vraisemblance des données. Le modèle avec ses nouveaux paramètres est alors utilisé pour ré-estimer les valeurs manquantes. Puis on recommence jusqu'à ce que la convergence soit atteinte (ou considérée comme telle). À la fin de l'exécution de l'algorithme, on dispose non seulement des paramètres de notre modèle, mais également d'une matrice de données complétée.

Cette technique est très coûteuse en temps de calcul comme beaucoup d'approches itératives (Hu et *al*. 2000; Magnani, 2003) [20-26]. De plus elle demande la spécification d'un modèle de génération des données. Cette tâche implique de faire un certain nombre d'hypothèses, ce qui est toujours délicat. Pour ces raisons, l'application d'EM pour remplacer les données manquantes n'est pas toujours envisageable.

Un grand nombre de méthodes de traitement des valeurs manquantes sont explorées dans les diverses études relevées dans la littérature. Il est impensable de tenter d'inclure l'ensemble de ces diverses méthodes dans une seule recherche; les méthodes de traitement les plus souvent comparées ont donc été considérées, le tableau 1.5 présente un résumé des différentes méthodes discutées dans ce mémoire.

Tableau 1.5- Taxinomie des techniques de substitution des valeurs manquantes.

Technique	Description	Champ d'application	Avantage	Inconvénient	Référence
Procédures de suppression					
Étude des cas complets (Listwise deletion)	Supprime toutes les observations dont certaines valeurs sont manquantes.	Il convient d'éviter	Facile à utiliser (par défaut dans la plupart des logiciels statistiques).	Sacrifie une grande quantité de données et a un impact négatif sur les paramètres d'estimation (corrélation – régression) et sur la puissance statistique.	Kim and Curry (1977) [24], Raymond (1986) [30], Malhotra (1987) [23], Little and Rubin (2002) [16].
Étude des cas disponibles (Pairwise deletion)	Crée une matrice de corrélation avec les valeurs disponible (chaque couple de variables est pris deux à deux)	Lorsque les données sont relativement faible (moins de 10 %).	Préserve davantage les données et est plus précise que la suppression listwise	Corrélations ou covariances biaisées	Gleason and Staelin (1975) [29], Kim and Curry (1977) [24], Raymond (1986) [30], Roth (1994) [28].
Procédures de remplacement					
Imputation par la moyenne totale (Total mean substitution)	Remplacer par la moyenne des valeurs disponibles de la variable, toutes les valeurs manquantes pour la même variable.	Lorsque les corrélations entre les variables sont faibles ($r < |.20|$) et le taux des leurs manquantes moins que 10 %.	Préserve la taille de la base de données et la rend facile à utiliser.	La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservée).	Ford (1976) [88], Raymond (1986) [30], Little and Rubin (1987) [16], Kaufman (1988) [25], Quinten and Raaijmakers (1999) [34].
Imputation par la moyenne de chaque classe (Subgroup mean substitution)	Remplacer par la moyenne des valeurs disponibles de la variable de la même classe, toutes les valeurs manquantes pour la même variable et dans la même classe.	Quand il est facile de définir les classes (classification supervisée).	Donne de meilleurs résultats, par rapport à Imputation par la moyenne totale.	La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservée).	Ford (1976) [88].

Technique	Description	Champ d'application	Avantage	Inconvénient	Référence
Imputation multiple	D'abord, estimation de m > 1 ensembles de valeurs plausibles pour les données manquantes sont créés. Chacun de ces ensembles est utilisé pour remplir les données manquantes et ainsi créer m ensembles complets de données, et elles sont combinées dans une unique base de données récapitulative.	Sous l'hypothèse que les valeurs manquantes sont aléatoires.	L'induction statistique (écart-type, p-values, etc.) qui découle de l'IM est généralement valide car elle incorpore l'incertitude engendrée par les données manquantes.	La complexité de calcul des m matrices (contraintes d'espace mémoire et de temps de traitement). Ne permet pas de seulement compléter une base de données… mais oblige à réaliser une analyse statistique.	Rubin (1978) [45], (Schafer et Graham, (2002) [19], Little et Rubin, 2002) [16], Fichman et Cummings (2003) [48].
Imputation par régression (Régression imputation)	On utilise les valeurs disponibles pour estimer les paramètres d'un modèle de régression, puis utiliser ces paramètres pour estimer la valeur manquante.	Lorsque plus de 20 % des données sont manquantes et les variables sont fortement corrélées.	Préserver l'écart des données estimées par rapport à la moyenne et la forme de la distribution	Distorsions des degrés de liberté et augmentation artificielle des relations entre les variables	Frane (1976) [32], Raymond and Roberts (1987) [31], Little and Rubin (2002) [16].
L'imputation hot deck (Hot-deck imputation)	Remplacer une valeur manquante par la valeur de la même variable à partir d'un cas similaire dans l'ensemble de données	Lorsque la similarité entre les cas est facile à déduire.	Préserve les distributions des variables	Risquent d'altérer les relations entre les variables	Ford (1983) [88], Sinharay, Stern et Russel (2001)[40].
Imputation par k-plus proche voisins (k-nearest-neighbors imputation)	Remplacer les valeurs manquantes par la valeur du k plus proche voisin dans l'ensemble de données.	Lorsque la mesure de distance entre les k plus proche voisins est facile à déduire, et les données sont chronologique.	Il ne fait aucune supposition quant à la distribution des données, et de prendre en considération la corrélation entre variables.	La difficulté réside dans le choix du paramètre k.	Chen and Shao (2000) [41],Engels and Diehr (2003) [42], Zhang (2008) [44], Zhang et al. (2008); Song et Shepperd (2007) [22].
Procédures de modélisation					
Maximum de	Les paramètres sont estimés par les	Lorsque les données	Augmentation de la précision si	Les hypothèses de la distribution	DeSarbo et al. (1986) [51],

Technique	Description	Champ d'application	Avantage	Inconvénient	Référence
vraisemblance (Maximum likelihood)	données disponibles et les valeurs manquantes sont estimées en fonction des paramètres.	observées sont tiré d'une distribution normale multi variée.	le modèle est correct.	exigée par la technique sont relativement strictes.	Lee and Chiu (1990) [52].
Maximisation d'espérance (Expected maximization)	An iterative process that continues until there is convergence in the parameter estimates.	Lorsque les hypothèses de répartition sont remplies.	Augmentation de la précision si le modèle est correct.	L'algorithme prend du temps à converger, et est trop complexe.	Laird (1988) [55], Little and Rubin (2002) [16], Malhotra (1987) [23], Ruud (1991) [56].

1.7 Choix méthodologique

Avant de réfléchir sur la méthode de traitement à appliquer aux données manquantes, un chercheur doit tout mettre en place pour ne pas avoir des valeurs manquantes (Fichman et Cummings, 2003) [48], puisqu'il n'existe pas de méthode totalement efficace pour traiter le problème des données manquantes. Comme l'affirme Allison (2001) [32], la meilleure méthode de traitement c'est de ne pas avoir de valeurs manquantes! Les méthodes permettent, au mieux, de réduire les biais induits par la présence de ces données manquantes.

La théorie statistique prévoit plusieurs méthodes pour l'estimation des valeurs manquantes qui font défaut à l'observation. Ainsi, l'on rencontre plusieurs pratiques dans les études statistiques. Ces pratiques sont plus ou moins basées sur l'intuition et le bon sens, plutôt que sur une théorie proprement dite. Surtout quand il s'agit de différents domaines d'application.

Dans ce sous-chapitre, l'accent sera mis sur les méthodes que nous utilisons dans notre étude, nous ne prétendons pas couvrir l'ensemble des méthodes, mais nous évoquons les plus courantes, celles que nous avons incluses dans notre travail.

Toutes n'ont pas les mêmes propriétés, aussi est-il important de bien spécifier les objectifs que l'on s'assigne avant de choisir une méthode de substitution afin de pouvoir vérifier l'adéquation entre objectifs et propriétés de chaque méthode. Les principaux objectifs que l'on peut vouloir poursuivre sont les suivants :

1. **Précision de la substitution** : la valeur de remplacement doit être aussi proche que possible de la vraie valeur.
2. **Précision de l'étape d'analyse** : dans notre contexte, la phase d'analyse correspond à la construction d'un modèle de classification supervisée. Un des objectifs est alors de maximiser la performance du classificateur.
3. **Complexité minimale.**

La présence de données aberrantes et manquantes est un problème important dans le contexte de notre travail, il faut les gérer avec précaution afin d'éviter de détériorer la performance de LAD.

Certaines méthodes de traitements spécifiques pour les données manquantes selon un mécanisme non aléatoire sont proposées, mais elles demeurent pour le moment expérimentales (Sinharay, Stern, et Russel, 2001) [40].
Un grand nombre de méthodes de traitement des valeurs manquantes sont explorées dans les diverses études relevées dans la littérature. Il est impensable de tenter d'inclure l'ensemble de ces diverses méthodes dans une seule recherche; les méthodes de traitement les plus souvent comparées ont donc été considérées.

1.7.1 Détection des valeurs aberrantes

Selon Daniel T. Larose (2005) [58], il y a deux méthodes numériques pour identifier une valeur aberrante.

A. Normalisation par le test Z :

Pour identifier les valeurs aberrantes, on peut utiliser la normalisation par le test Z, souvent une valeur aberrante peut être identifiée parce qu'elle est à trois écarts types supérieure à la moyenne, c'est-à-dire qu'elle a un test Z de normalisation qu'est, soit : $X^* < -3$ ou $X^* > 3$.
Avec X^*= (x-moyenne (x)) / écart type (x)
Exemple : Exemples des tableaux 1.1 et 1.2

- La moyenne (a_1) = 6
- L'écart type (a_1) = 4,472

Pour la valeur de 120 on a $X^* = (120-6)/ 4,472 = 25,5 > 3$, donc la valeur 120 c'est une valeur aberrante.

Inconvénient : la moyenne et la variance doivent être connues, toutes deux incluses dans la formule de la normalisation par le test Z, sont assez sensibles à la présence de valeurs aberrantes.

B. Utilisation de l'amplitude interquartile (méthode utilisée dans le cadre de notre travail).

Les quartiles de l'ensemble de données divisent les données en quatre ensembles, chacun contenant 25 % des données :

Le 1er quartile (Q1) est le 25e centile.

Le 2eme quartile (Q2) est le 50e centile (la médiane).

Le 3eme quartile (Q3) est le 75e centile.

L'amplitude interquartile (AIQ) est une mesure de la variabilité beaucoup plus robuste que l'écart-type, il est calculer comme suit : AIQ = Q3-Q1.

La détection des valeurs aberrantes est définie comme suit :

Xi est une valeur aberrante si Xi < Q1-1.5 AIQ ou si Xi> Q3+1.5 AIQ.

Le calcul à effectuer pour déterminer la position des quartiles est le suivant :

$$PQ_j = j(n+1)/4$$

Avec j=1, 2 ou 3, selon le quartile que nous désirons calculer, et n= le nombre d'observations.

Exemple : pour le tableau 1.1

- $PQ_1 = 1*(5+1) / 4$

  $PQ_1 = 1.5$
- $PQ_3 = 3*(5+1) / 4$

  $PQ_3 = 3.75$

Donc le premier quartile se situera entre la 1^{ere} et la $2^{ème}$ observation, et le troisième quartile entre la $3^{ème}$ et la $4^{ème}$ observation, c.à.d. $Q_1= (1+3)/2 = 2$

et $Q_3 = (5+9)/2 = 7$.

L'écart interquartile vaut : $AIQ = Q_3-Q_1 = 7-2 = 5$

Une valeur de X est considérée aberrante si X< 2-1.5 x 5 ou si X> 7+1.5 x 5.

c.à.d. X < -5.5 ou si X > 14.5.

Pour la valeur de 120 on a 120 > 14.5, donc la valeur 120 c'est une valeur aberrante.

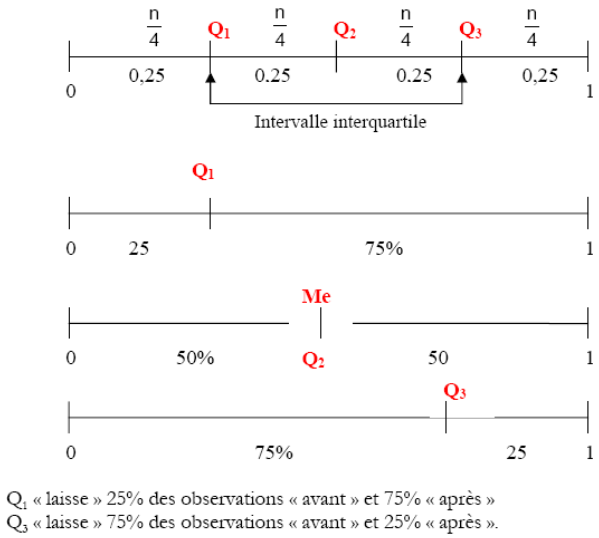


Figure 1.4- Intervalle interquartile

Une fois identifiées, les valeurs aberrantes seront discutées et traitées par la suite comme des valeurs manquantes (la valeur aberrante sera éliminée c'est-à-dire convertie en valeur manquante)

1.8 Conclusion

Dans ce chapitre, nous avons pu voir dans un premier temps les différents concepts de la fonction maintenance, et l'impact économique de cette fonction sur le rendement global d'une entreprise. Nous avons ensuite exposé différents systèmes de surveillances et de diagnostic utilisés pour optimiser la fonction maintenance.

Ces systèmes de surveillances doivent traiter un nombre important de données, et doivent pouvoir exploiter le savoir-faire des automaticiens.

Après avoir présenté les différentes méthodes de substitution des valeurs manquantes, on a présenté la méthodologie de choix de la technique de substitution.

Dans le chapitre suivant, la théorie de substitution des valeurs manquantes par MIN-MAX qu'on propose et qui est propre à LAD est abordée, ainsi qu'une présentation de la méthodologie d'analyse logique des données dans le contexte de la maintenance conditionnelle.

CHAPITRE 2. LE TRAITEMENT DES DONNÉES MANQUANTES ET ABERRANTES POUR CBM-LAD

2.1 Introduction

Le domaine de la maintenance industrielle d'aujourd'hui est de plus en plus équipé de système d'acquisition numérique de mesure et de suivi de l'état des équipements qui génère un volume de données important. Ces données peuvent être utilisées pour déduire de futures décisions affectant l'état de santé des équipements.

Pour pouvoir analyser et extraire une information pertinente de ces bases de données, il est essentiel de travailler sur des données fiables, ce qui nécessite l'évaluation de la qualité de ces dernières avec des techniques permettant de traiter les données manquantes, aberrantes ou erronées. Après cette phase d'évaluation, une bonne préparation de données devient une étape préalable clef dans le processus de traitement de données avant de les exploiter par un logiciel.

L'objectif premier de ce chapitre est donc de proposer une méthodologie adaptée de remplacement des données manquantes, cette méthodologie exploite la structure spécifique de LAD. Pour cette raison, avant de présenter la méthodologie proposée pour la substitution des valeurs manquantes, nous introduisons l'approche de LAD "Logical Analysis of Data" dans le contexte de la maintenance conditionnelle.

2.2 Analyse logique des données (LAD)

Avec l'avènement des Technologies de l'Information et des Communications (TIC), les volumes de données brassés par les entreprises sont devenus énormes. Les décideurs croulent désormais sous l'information à tel point qu'il est extrêmement difficile pour eux

d'avoir une bonne vision de leurs données afin d'en tirer profit. Cette situation paraît relativement paradoxale dans la mesure où les marchés sont de plus en plus concurrentiels et les entreprises se doivent d'être réactives et d'exploiter les données disponibles afin d'appuyer leurs décisions stratégiques.

Il est ainsi à la fois difficile et primordial pour un décideur d'analyser des volumes importants de données afin d'en dégager des informations utiles, comme des indicateurs de performance et les indicateurs de prise de décision pour la maintenance conditionnelle.

Dans le cadre d'analyse de données, la classification est une étape importante, son objectif est de regrouper les objets d'un ensemble de données en classes homogènes selon leur ressemblance. Il existe deux types d'approches, la classification supervisée et la classification non supervisée. La distinction entre ces deux approches vient de la connaissance ou non des classes. En effet, pour une approche non supervisée, les classes sont à trouver ou définir de manière automatique (Cormack, 1971; Johnson, 1967) [59] [60], alors qu'une approche supervisée part du principe que les classes sont connues, ayant été préalablement définies par un expert (Borko et Bernick, 1963; Yang et Liu, 1999) [61] [62]. Cette seconde approche est appelée catégorisation.

- La classification supervisée qui consiste à classer des éléments dans des classes connues (par exemple état normal et état anormal). On parlera aussi d'apprentissage supervisé.
- La classification non supervisée qui consiste à regrouper les éléments ayant des comportements similaires dans des classes, inconnues au départ. On parlera alors de clustering, de segmentation ou d'apprentissage non supervisé.

La méthode de classification par l'analyse logique des données « Logical Analysis of Data » (LAD), étudiée dans ce mémoire, est une méthode de classification supervisée.

L'objectif principal de la méthode LAD est de classifier de nouvelles observations à partir d'un ensemble d'observations déjà classifié. Chaque observation est caractérisée

par un vecteur des valeurs d'attributs et la classe à laquelle appartient cette observation. La spécificité de cette méthode comporte la découverte d'un patron dont sa présence dans une observation conclue l'appartenance à une classe considérée.

Brièvement, le principe du LAD comporte l'identification des patrons capables de classifier correctement toutes les observations.

2.2.1 Historique

Les débuts de LAD peuvent être datés des années 80, quand l'équipe du Prof. Hammer au centre recherche opérationnelle de l'Université de Rutgers (RUTCOR) dans le New Jersey a découvert la possibilité d'utiliser des caractéristiques spécifiques de fonctions booléennes pour l'extraction de la connaissance des données. La première présentation publique de ces idées a été faite dans une conférence a Passau, Allemagne (Hammer, 1986) [63], dont le contenu a été publier plus tard (Crama et *al*, 1988) [64].

La recherche théorique dans cette direction est poursuivie à RUTCOR par l'équipe de M. Hammer. En attendant, la méthodologie a évolué du côté pratique, une procédure pour transformer des données brutes en format booléen a été créée, pour permettre d'appliquer le LAD sur des données arbitraires au lieu d'être limitée à des données booléennes, par la suite des techniques pour le traitement du bruit et des données absentes ont été ajoutées. Les premières expériences empiriques ont été testées avec la méthode LAD en 1994 (Boros et *al*, 1994) [65]. Ces résultats ont donné l'appui pratique pour encourager de poursuivre davantage la recherche dans ce domaine. Un rapport technique en 1996 (Boros et *al*, 1996) [66] a fourni une description détaillée et complète de l'exécution de LAD en tant qu'une méthode concrète pour l'analyse logique de données et classification, avec des résultats des expériences empiriques sur les ensembles de données connus, aussi bien que quelques études pilotes dans le domaine médical. Le même rapport de 1996 a été actualisé et édité dans un journal international (Boros et *al*, 2000) [67].

Après 1996, LAD est devenu également l'un des sujets de recherches dans différents domaines, la première application de la méthodologie LAD en génie industriel a été initiée par la professeure Soumaya Yacout et son équipe en 2005 ce qui a abouti à la création du logiciel cbmLAD dédié spécialement pour la maintenance conditionnelle en 2007 (David S. 2007) [1]. La première version de cbmLAD est limitée à la discrimination entre deux classes seulement, voir Fig. 2.1.

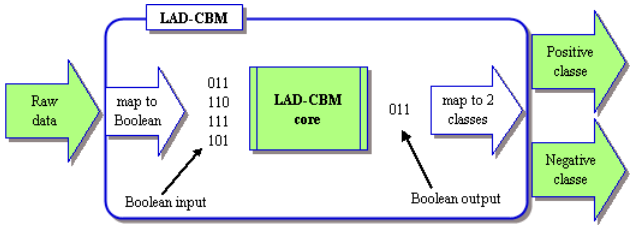


Figure 2.1: Processus général de CBM-LAD

2.2.2 La méthodologie

L'Analyse Logique des Données (LAD) a pour objectif de rechercher à partir des données disponibles un modèle explicatif ou prédictif avec une ou plusieurs relations logiques "patrons" satisfaisantes par un grand nombre d'observations dans une même classe.

Ces relations logiques caractérisent d'une certaine façon les observations ayant des propriétés communes. Par exemple, les observations peuvent être des mesures de température, pression, niveau d'huile pour une machine donnée sous forme de vecteurs de chiffres; et la propriété commune, d'avoir un état donnée de la machine (classe).

Exemple : considérons le cas d'un responsable de service de maintenance qui s'intéresse à la nature d'une observation dont il veut déterminer si elle est reliée à l'état défaillant de la machine ou à l'état normal. Imaginons qu'il souhaite construire une règle lui permettant de prévoir, à l'avance, sur la base de mesure des paramètres simples, l'état de la machine. Pour cela, il peut procéder par apprentissage à partir de données. Cela

consiste, pour lui, à recueillir des observations à partir de l'historique de la machine dans un état donné comme l'état normal ou de l'état défaillant. Sur la base de ce corpus qu'on appellera « données d'apprentissage », il mettra en œuvre une méthode d'apprentissage qui l'aiderait à bâtir son modèle d'identification automatique de l'état d'une machine à partir des nouvelles observations.

La méthodologie de LAD décrite par Boros (Boros et *al*. 2000) [67], a été adaptée pour la conception de logiciel cbmLAD qui se compose des modules suivants :

- Lecture des données et de préparation.
- Extraction de l'information et de patrons.
- La mise en place d'un classificateur.
- Essais de classification sur des données de test.
- Classification des nouvelles observations.

Ces étapes sont indiquées en fig. 2.2, et expliquées dans le paragraphe suivant.

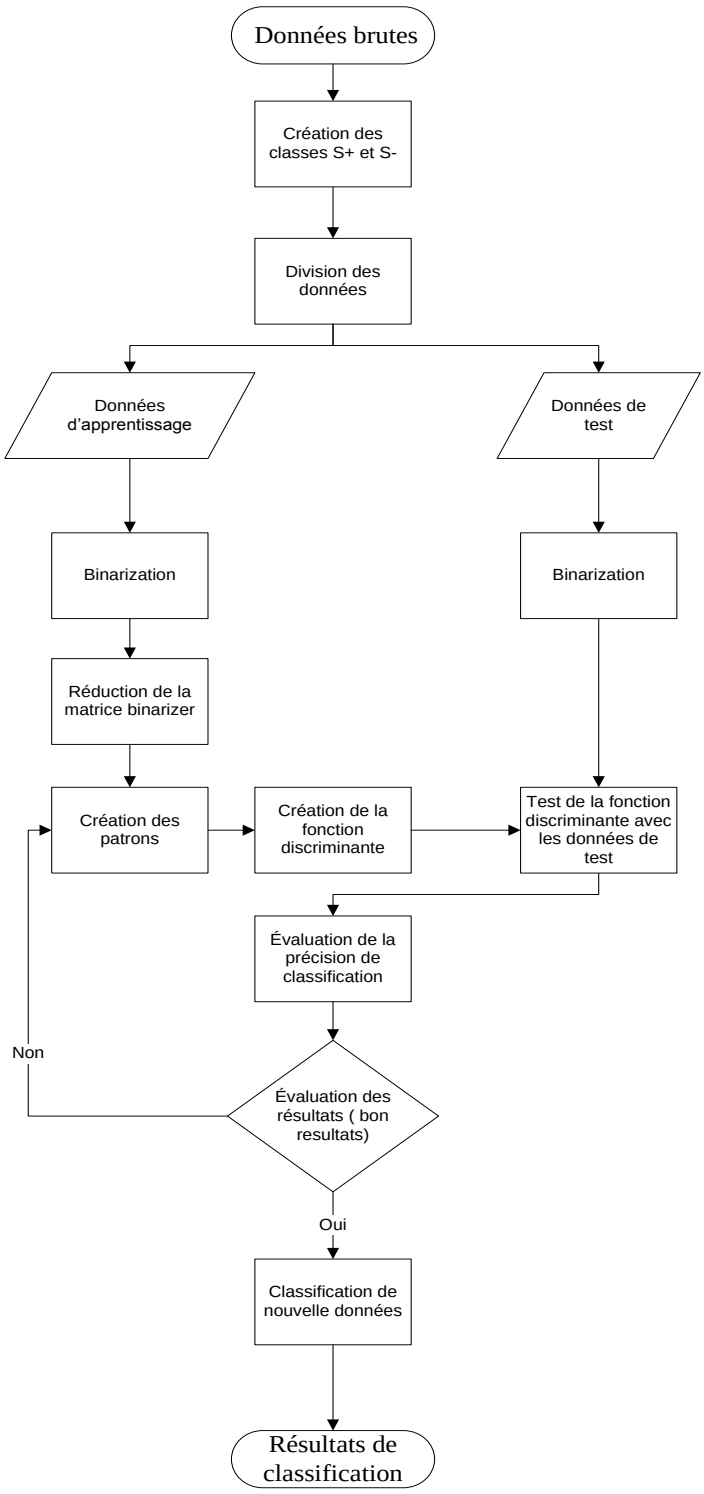


Figure 2.2- Diagramme de cbm-LAD

2.2.2.1 La lecture des données et la préparation

L'application cbmLAD lit les données des observations ou des signaux obtenus par des capteurs ou recueillies à des inspections périodiques. Les données représentent les mesures de nombreux indicateurs tels que les fréquences de vibration, la température, la composition de l'huile de lubrification. Chaque observation est accompagnée de sa classe, c'est-à-dire par l'indication si la machine est dans un état normal ou en détérioration.

La lecture et la préparation des données ont deux objectifs. Tout d'abord, les données sont transformées en un format binaire. La façon la plus simple pour accomplir cette étape est de trouver un ou plusieurs points de coupe, et de classer les observations soit au-dessus ou au-dessous de ces points seuils.

Le deuxième objectif est de réduire la taille de l'archive binaire, créé par la binarisation, en éliminant le plus grand nombre d'indicateurs inutiles. Pour se faire il faut transformer les données binarisées en une matrice représentative du modèle d'optimisation type couverture d'ensemble (en anglais set covering problem) (Taha H. 1975) [68].

2.2.2.2 Extraction de l'information et reconnaissance des patterns

Une caractéristique spécifique de cbmLAD est la détection de patrons logiques (patterns) qui distinguent les observations recueillies pendant que la machine est en détérioration, de toutes les autres observations. Un pattern caractéristique de cet état est une combinaison des valeurs d'indicateurs, par exemple, une forte concentration de particules métalliques dans les lubrifiants ou d'un niveau élevé de fréquences de vibration. cbmLAD détecte ces patrons qui sont capables de caractériser correctement les nouvelles observations dans l'état correspondant.

2.2.2.3 La mise en place d'un classificateur

Les patrons générés pendant un état de détérioration (ou normale) de la machine, sont appelés des patrons positifs (ou négatifs), ils fournissent des indications sur l'état de détérioration (ou normale), donc ils peuvent être utilisés pour la construction d'une règle de classement.

Plus une nouvelle observation est couverte par des patrons positifs (ou négatifs), plus elle indique que la machine est en état de détérioration (ou normale).

2.2.2.4 Essais de classification

Les essais constituent une étape qui consiste à faire la lecture des observations qui sont prises pendant l'état de détérioration, et d'autres prises pendant l'état normal, par la suite cbmLAD classe ces observations dans les deux classes (détérioration / normal) et comparer l'état réel à l'état (classe) obtenu par cbmLAD, une analyse statistique est effectuée pour calculer le pourcentage de réponses correctes. Si ce pourcentage n'est pas satisfaisant, une recherche des causes est effectuée afin d'améliorer le taux de classement. Par la suite une nouvelle observation est classée à l'aide du classificateur défini à l'étape 2.2.2.3.

La règle de classement est calculée par une fonction appelée fonction discriminante et notée (Δ) (Boros et *al.* 2000) [67], avec

$$\Delta = \sum_{k=1}^{r} w_k^+ P_k - \sum_{l=1}^{s} w_l^- N_l \qquad (1)$$

P_1, $P_2 \ldots P_k$ sont égales à 1 si les patrons 1, 2,3….k respectivement générées à l'état normal de la machine couvre la nouvelle observation et égales à zéro autrement, et N_1, $N_2 \ldots N_k$ sont égales à 1 si les patrons 1, 2,3….k respectivement générées à l'état détérioration de la machine couvre la nouvelle observation et égales à zéro autrement. Les poids w^+_k (w^-_l) représentent la fraction d'observations couvertes par le patron P_k (N_l) dans les données d'apprentissage.

Après la définition de la fonction discriminante, on doit la tester avec les données de test, la valeur de Δ est utilisée pour classer les observations selon les trois cas suivants (Alexe et *al.* 2002) [69] :

- Si $\Delta < 0$, l'observation est classée comme négatif (état normal).
- Si $\Delta > 0$, l'observation est classée comme positif (état de détérioration).
- Si $\Delta = 0$, l'observation reste non classifiée.

Suite à cette classification, nous pouvons mesurer l'exactitude (Accuracy) de la fonction discriminante, qui est mesurée par la méthode décrite dans le paragraphe suivant.

2.2.2.5 Évaluation de la qualité de classification

Il existe différentes mesures pour évaluer la qualité de classification. La plupart de ces mesures sont basées sur la matrice de confusion du tableau 2.1. qui croise la classe réelle des observations avec la classe prédite par la fonction discriminante.

La précision de la méthode de classification est évaluée en utilisant les données de la matrice (Kohavi and Provost, 1998) [70]. Le tableau 2.1 présente la matrice de confusion pour deux valeurs de classe (positive, négative). Ainsi,

- **a** est le nombre de classements corrects des observations de classe **négative**.
- **b** est le nombre de classements incorrects des observations de classe **négative**.
- **c** est le nombre de classements incorrects des observations de classe **positive**.
- **d** est le nombre de classements corrects des observations de classe **positive**.

Tableau 2.1- Matrice de confusion

Classe d'origine	**Classe d'affectation**	
	Négative	**Positive**
Négative	a	b
Positive	c	d

La qualité de la classification est définie par :

$$ACC = (a + d) / (a + b + c + d) \qquad (2)$$

Toutes les étapes du cbmLAD sont présentées dans le diagramme Figure 2.2.

2.2.3 Traitement des valeurs manquantes pour CBM-LAD

2.2.3.1 Remplacement de valeur manquante par le Min et le Max de la même classe

Suivant les techniques d'imputation décrites dans la littérature, la méthode qui permet de tenir compte de l'incertitude et de ne pas faire la moindre hypothèse sur les données est celle qui remplace la valeur manquante par toutes les valeurs possible [71]. L'idée est la suivante : puisqu'on ne connaît pas la valeur manquante, le plus simple est encore d'utiliser toutes les valeurs possibles. Ainsi toutes les valeurs observées d'un attribut seront utilisées pour créer autant de nouvelles observations, ne différant que par les valeurs de remplacement de la valeur manquante. L'incertitude liée à l'imputation est effectivement prise en compte. En revanche cela se fait de manière déterministe.

L'autre version de cette technique, dans laquelle une information de classe est prise en compte est identique à celle de la première méthode, à la différence près qu'on ne s'intéresse qu'aux valeurs des observations appartenant à la même classe.

Si nous avons des observations avec plus d'attributs avec une valeur manquant, nous ferons notre imputation d'un premier attribut, puis réaliseront le remplacement de l'attribut suivant, etc., jusqu'à ce que toutes les valeurs d'attributs inconnus soient remplacées par de nouvelles valeurs d'attributs connus.

Grzymala-Busse et Hu (2001) [72] notent que ces méthodes sont prometteuses, mais soulignent que puisqu'en imputant la valeur manquante par toutes les valeurs possibles d'un attribut, on peut obtenir autant d'informations que possible, mais la taille de la table qui en résultent peuvent augmenter de façon exponentielle, ce qui provoque un blocage du système en raison de l'insuffisance de la mémoire et le temps pour faire les calculs.

2.2.3.2 Notre cadre méthodologique

Notre cadre méthodologique est celui de la classification supervisée par la méthode de LAD, donc nous cherchons à obtenir un classificateur robuste et performant à partir de données incomplètes, peu nous importe de compléter la matrice des données avec des valeurs aussi proches que possible de la réalité, à laquelle nous n'avons pas accès.

Notre objectif est de construire un bon classificateur à partir de données incomplètes. Pour cela, nous proposons une nouvelle méthode d'imputation des données manquantes basée sur les valeurs existantes et sur le principe de la théorie de LAD.

2.2.3.3 Description de notre méthode

Pour réaliser cet objectif dans le cadre de l'application de LAD, E. Boros (E. Boros et al. 1999) [73], propose une méthode qui consiste à remplacer les valeurs manquantes, par des valeurs logiques (0 ou 1, puisque il traite une matrice binarisée), sous la condition que ce remplacement ne produit pas une redondance des observations et de ne pas avoir une même observation dans deux classes différentes, c'est à dire que lorsqu'on utilise cette méthode, une table de données cohérente ne doit pas être convertie en une table incohérente.

La table est incohérente quand elle contient au moins une paire d'observations contraires, c'est à dire, des observations caractérisées par les mêmes valeurs de tous les attributs, avec des valeurs différentes d'une décision (classe).

En s'inspirant de cette démarche nous proposons la méthode Min-Max pour imputer les valeurs manquantes numériques.

Cette façon nous a donné l'idée de procéder avec le même principe mais on minimise l'espace de solution possible pour utiliser la méthode de remplacement avec toutes les valeurs possibles, et garde une matrice de données complète cohérente, cette méthode est basée sur l'utilisation des valeurs Min et Max de l'attribut comme valeurs de remplacement, en supposant toujours que le pire des cas la valeur manquante sera remplacée par les valeurs extrêmes.

Puisque la présence des valeurs manquantes nécessite la modification du concept de couverture des patrons, Boros a introduit la notion du patron robuste voir [74], pour caractériser la qualité des patrons produits par LAD après remplacement des valeurs binaires manquantes par les deux valeurs logique 0 ou 1.

Un patron robuste, c'est un patron qui couvre une observation quel que soit la valeur utilisée pour imputée ses valeurs manquantes, pour l'application cbmLAD, seuls les patrons robustes sont pris en compte.

Nous avons choisi de compléter chaque attribut incomplet de façon à produire des patrons robustes qui couvrent les observations à valeur manquante sans être influencés par les valeurs de remplacement utilisées.

L'idée sous-jacente est que l'absence de certaines valeurs, relativement à un attribut, diminue le nombre de points de coupure relatif à cet attribut. En conséquence nous proposons une méthode pour imputée les données manquantes, attribut par attribut, afin de garder les mêmes points de coupure définis avant imputation, dans la mesure du possible.

Pour illustrer notre méthode, on considère que $\mathcal{E}^+ = \{w_{11}, w_{12}.....w_{1n}\}$, est un l'ensemble des observations de la classe positive et $\mathcal{E}^- = \{w_{01}, w_{02}.....w_{0m}\}$ représente l'ensemble des observations de la classe négative, $\mathcal{E}^+$ et $\mathcal{E}^-$ constituant la base d'apprentissage notée par $\mathcal{E}$. Supposons que la base ait un certain nombre d'observations ayant des valeurs d'attributs manquantes. Considérons l'attribut symbolique A qui prend ses valeurs dans l'ensemble $S_0 = \{v_{A01}, v_{A02}.....v_{A0l}\}$ pour la classe négative et $S_1 = \{v_{A11}, v_{A12}.....v_{A1k}\}$ pour la classe positive.

Notons :

$\mathcal{E} = \{w_{11}, w_{12}.....w_{1n}, w_{01}, w_{02}.....w_{0m}\}$, l'ensemble des observations

Les deux valeurs limites que peut prendre l'attribut A sont définies par :

$v^+_{A0} = Max\{v_A \quad v \in S_0\}$,

$v^{-}_{A0 = Min}\{v_A \quad v \in S_0\}$,

$v^{+}_{A1 = Max}\{v_A \quad v \in S_1\}$

Et $v^{-}_{A1 = Min}\{v_A \quad v \in S_1\}$

Si WA^{m} est une observation avec valeur manquantes pour l'attribut A, dans le pire des cas la valeur manquante de WA^{m} sera remplacée par les deux valeurs extrêmes que peut prendre A (v^{-}_{A0} ou v^{+}_{A0}).

Nous avons recours aux valeurs extrêmes Min et Max, parce que ce sont les seuls valeurs qui peuvent remplacer les valeurs manquantes, sans ajouter d'autre point de coupure, vu qu'ils sont déjà utilisés pour définir les points de coupures de l'attribut considéré.

Pour illustrer notre méthode, on considère l'attribut A, qui est représentée par les valeurs suivantes :

$\mathcal{E}^{+} = \{x_1, x_2, x_3, x_4\}$ et $\mathcal{E}^{-} = \{y_1, y_2, ?, y_4\}$ avec y_3 comme valeur manquante.

On suppose que les valeurs des deux classes sont ordonnées de la façon suivante :

$$x_1 \leq x_2 \leq y_2 \leq y_1 \leq x_3 \leq y_4 \leq x_4$$

Puisque il y a quatre alternances d'état, soit :

- De l'état positif avec la valeur x_2 vers l'état négatif de valeur y_2.
- De l'état négatif avec la valeur y_1 vers l'état positif de valeur x_3.
- De l'état positif avec la valeur x_3 vers l'état négatif de valeur y_4.
- De l'état négatif avec la valeur y_4 vers l'état positif de valeur x_4.

Il existe quatre points de coupure que nous établissons à l'aide de la formule (3) suivante :

$$P_c = ½ (V_{S-1} + V_S) \qquad (3)$$

Avec P_c est le point de coupure et V_{S-1} la valeur de l'indicateur dont l'état de départ et V_S est la valeur de l'indicateur qui présente le signe contraire.

Donc suivant le formule (3), les points de coupures de l'attribut A sont :

$$P_{c1} = ½ (x_2 + y_2),$$
$$P_{c2} = ½ (y_1 + x_3),$$
$$P_{c3} = ½ (x_3 + y_4) \text{ et}$$
$$P_{c4} = ½ (y_4 + x_4)$$

Les valeurs limites que peut prendre l'attribut A pour les deux classes sont définies par :

$v^{+}_{A0 = Max} \{v_A \quad v \in \mathcal{E}^{+}\}$

$v^{+}_{A0} = x_4$

$v^{-}_{A0 = Min} \{v_A \quad v \in \mathcal{E}^{+}\},$

$v^{-}_{A0} = x_1$

$v^{+}_{A1 = Max} \{v_A \quad v \in \mathcal{E}^{-}\},$

$v^{+}_{A1} = y_4$

Et $v^{-}_{A1 = Min} \{v_A \quad v \in \mathcal{E}^{-}\}$

$v^{-}_{A1} = y_2$

On remplace la valeur manquante y_3 par $y_{31} = y_2$ et $y_{32} = y_4$ donc l'ordonnancement des valeurs sera :

$$x_1 \leq x_2 \leq y_{31} \leq y_2 \leq y_1 \leq x_3 \leq y_4 \leq y_{32} \leq x_4$$

Les nouveaux points de coupures après remplacement de la valeur manquante de l'attribut A sont :

$$P_{c1} = ½ (x_2 + y_{31}) = ½ (x_2 + y_2),$$
$$P_{c2} = ½ (y_1 + x_3),$$
$$P_{c3} = ½ (x_3 + y_4) \text{ et}$$
$$P_{c4} = ½ (y_{32} + x_4) = ½ (y_4 + x_4)$$

On constat après cette démonstration qu'on a obtenu les même points de coupure après remplacement des valeurs manquantes.

2.2.3.4 Exemple d'application numérique

Le tableau 2.2 montre un exemple de base de données avec des valeurs manquantes, et les points de coupures calculés avec les valeurs observées de chaque attribut, et le tableau 2.5 illustre sur le même exemple l'imputation des valeurs manquantes par notre méthode. Dans cet exemple nous considérons un problème avec trois attributs, que nous noterons A, B, C. La matrice des données est composée de 6 observations reparties entre deux classes. Elle contient quatre valeurs manquantes pour les observations w_2, w_3, w_5 et w_6. Ainsi nous avons :

$\mathcal{E} = \{w_1, w_2 \ldots . w_6\}$, l'ensemble des observations

$\mathcal{E}^o = \{w_1, w_4\}$, l'ensemble des observations sans valeurs manquantes

$\mathcal{E}^m = \{w_2, w_3, w_5, w_6\}$, l'ensemble des observations avec des valeurs manquantes

Tableau 2.2- Exemple de base de données avec valeur manquantes

Observations	Attributs			Classe
	A	B	C	
W1	3.5	3.8	2.8	1
W2	2.6	**?**	5.2	1
W3	1	2.6	**?**	1
W4	3.5	1.6	3.8	0
W5	**?**	2.1	1	0
W6	2.4	2	**?**	0

Pour montrer l'efficacité de cette méthode on va suivre les étapes suivantes :

1- Calcule des points de coupure sans tenir compte des valeurs manquantes.
2- Calculer les valeurs Min-Max pour chaque attribut et pour chaque classe.
3- Remplacer les valeurs manquantes par le Min el la Max de l'attribut dans la même classe.
4- Recalculer les points de coupures.
5- Binariser la matrice obtenue en 3.

6- Trouver les patrons robustes.

7- Comparer les deux résultats avant et après imputation.

Les points de coupures calculés sans tenir compte les valeurs manquantes sont :

A = {1.7, 2.5, 3.05},
B = {2.35},
C = {1.9, 3.3, 4.5},

Remplacement des valeurs manquantes par la valeur min et la valeur max de chaque attribut pour les deux classes :

Tableau 2.3- La base de données du tableau 2.2 après remplacement des valeurs manquantes par la méthode Min-Max

Observations	Attributs			Classe
	A	B	C	
W_1	3.5	3.8	2.8	1
W_2	2.6	**2.6**	5.2	1
W_2'	2.6	**3.8**	5.2	1
W_3	1	2.6	**2.8**	1
W_3'	1	2.6	**5.2**	1
W_4	3.5	1.6	3.8	0
W_5	**2.4**	2.1	1	0
W_5'	**3.5**	2.1	1	0
W_6	2.4	2	**1**	0
W_6'	2.4	2	**3.8**	0

Les points de coupures:

A = {1.7, 2.5, 3.05},
B = {2.35},
C = {1.9, 3.3, 4.5},

Après la binarisation on obtient le tableau suivant :

Tableau 2.4- La base de données du tableau 2.3 après binarisation

Observations	Attributs							Classe
	a_1	a_2	a_3	b_1	c_1	c_2	c_3	
W_1	1	1	1	1	1	0	0	1
W_2	1	1	0	1	1	1	1	1
W_2’	1	1	0	1	1	1	1	1
W_3	0	0	0	1	1	0	0	1
W_3’	0	0	0	1	1	1	1	1
W_4	1	1	1	0	1	1	0	0
W_5	1	0	0	0	0	0	0	0
W_5’	0	0	1	0	0	0	0	0
W_6	1	0	0	0	0	0	0	0
W_6’	1	0	0	0	1	1	0	0

Génération des patrons :

Tableau 2.5- Les patrons obtenus à partir du tableau 2.4

Patron		Description	Les observations couvertes	Type du patron
Name	Sign			
N_1	-	not b_1	W_4 ,W_5 , $W_{5'}$, W_6 , $W_{6'}$	Robuste
N_2	-	a_1 not b_1	W_4 ,W_5 , W_6 , $W_{6'}$	Robuste
N_3	-	c_1 not b_1 not c_3	W_4 , $W_{6'}$	Normal
P_1	+	b_1	W_1 ,W_2 ,$W_{2'}$,W_3,$W_{3'}$	Robuste
P_2	+	b_1 c_1	W_1 ,W_2 ,$W_{2'}$,W_3,$W_{3'}$	Robuste
P_3	+	not a_1 b_1	W_3, W3’	Robuste
P_4	+	c_1 not c_2	W_1, W3	Normal

Puisque on a les mêmes points de coupure dans les deux matrices des données (le tableau 2.2 et le tableau 2.3), cela veut dire qu’on n’a pas changé la structure des classes.

On remarque qu'on a 5/7 des patrons générés après remplacement des valeurs manquantes sont robustes, puisque ils couvrent les observations avec valeurs manquantes, après remplacement avec la méthode Min-Max.

2.3 Évaluation des méthodes

Avec ou sans données manquantes, le but de toute procédure statistique est d'effectuer un prétraitement valide à propos des données traitées (Schafer, et Graham, 2002) [19]. Ainsi, les analyses ne s'attardent pas à vérifier si les méthodes de traitement comparées permettent de remplacer efficacement les valeurs manquantes. À cet effet, Schafer et Graham soulignent que pour effectuer une évaluation adéquate, une méthode de traitement des valeurs manquantes doit être testée dans le contexte spécifique des analyses qui sont visées. Ainsi, une méthode de traitement peut s'avérer efficace dans le contexte d'analyses de régression, mais inadéquate dans le contexte de classification, par exemple.

Dans le cadre de l'utilisation de l'analyse logique des données (LAD) dans le contexte de la maintenance conditionnelle, plusieurs paramètres doivent être considérés afin de porter un jugement sur l'efficacité d'une méthode de traitement des valeurs manquantes, dont les plus importants sont le nombre de données manquantes, le mécanisme des valeurs manquantes (section 1.6.3), ainsi que la nature des données (série temporelle dans le cas de la maintenance industrielle) .

Pour ces différentes caractéristiques, deux critères sont employés pour juger de l'efficacité des méthodes de traitement, soit le biais ainsi que la qualité de traitement à la sortie du processus de LAD. Une méthode de traitement sera d'autant plus efficace que la valeur de l'estimé obtenue se rapprochera de la valeur réelle et que la qualité de la classification LAD avec les valeurs manquante se rapproche de celle obtenue avec des données sans valeurs manquantes.

Dans le cadre de ce rapport l'évaluation de la qualité d'une méthode de traitement des valeurs manquantes, sera jugée par la qualité de classification à la sortie de cbmLAD,

puisqu'on ne connaît pas les vraies valeurs des données manquantes (le scénario le plus proche de la réalité).

2.4 Conclusion

Dans ce chapitre, nous avons fait une présentation de l'analyse logique des données comme technique d'intelligence artificielle utilisée dans le domaine de la maintenance conditionnelle, qui se distingue des autres méthodes de classification par le fait qu'elle utilise une technique combinatoire, pour générer et analyser en profondeur un petit sous-ensemble de combinaisons de variables, qui peut décrire le caractère positif ou négatif d'une observation, dans le but de trouver une fonction binaire, discriminante capable de distinguer les observations positives et celles qui sont négatives.

Pour la mise en œuvre de la méthodologie LAD, une étape primordiale consiste à faire un prétraitement des données avant de commencer la binarisation, cette phase permet de détecter et traiter les données aberrantes et/ou manquantes. Dans la deuxième partie de ce chapitre on a proposé de mettre en pratique une méthodologie de prétraitement des données à utiliser dans le cadre de cbm-LAD, basée sur le MIN-MAX.

Dans le chapitre 3, nous allons développer l'étude expérimentale et la présentation des résultats d'évaluation des performances de la méthode de substitution des données manquantes proposée.

CHAPITRE 3. RÉSULTATS EXPÉRIMENTAUX

3.1 Introduction

Pour compléter l'étude présentée ci-dessus, un certain nombre d'expérimentations ont été menées dans le contexte de la classification supervisée avec le logiciel cbm-LAD. Nous préférons désormais voir comment notre méthode se comporte de manière empirique. Á travers d'expériences sur des données réelles, nous souhaitons d'une part identifier les conditions qui sont les plus favorables à son application dans le cadre de l'analyse logique des données, et d'autre part juger de sa qualité de substitution des données manquantes, comparativement aux techniques existantes pour justifier l'intérêt qu'il peut y avoir à l'utiliser. Cela permet d'avoir une idée pour choisir la technique de substitution qui permet au logiciel cbm-LAD d'avoir la précision de classification pour un problème concret.

3.2 Expérimentations

La plupart des techniques d'imputation relatées dans la littérature statistique, cherchent à trouver des valeurs d'imputation les plus proches possibles des valeurs réelles, ce qui nécessite l'utilisation des données originales, pour la validation des résultats. Notre approche se base sur les méthodes d'imputation qui permettent de conserver la performance de la technique LAD, mesurée par la précision de bonne classification sans essayer de connaitre les valeurs réelles des données manquantes.

Dans le contexte de la classification supervisée avec cbm-LAD, la performance des techniques de substitution n'est évaluée ni sur la proximité entre les valeurs de substitution et les valeurs réelles, ni sur le respect de la distribution de certaines statistiques.

Dans cette optique, évaluer une technique d'imputation revient à évaluer la performance du modèle de classification construit à partir des données de la base que cette technique aura complétée. Pour cela nous avons choisi la précision (Accuracy en anglais), qui est

calculée à partir de la matrice de confusion (voir section **2.2.2.5**), étant donné que c'est la seule mesure de performance, utilisée dans les différentes études comparatives sur le sujet qui nous ont servi de référence [75], [76], [77]. En plus nous avons aussi considéré la marge de séparation dénotée par **MS**, comme une autre mesure de performance, afin d'être homogène avec les autres études de LAD.

La marge de séparation est définie comme une distance entre la valeur de la fonction de classification (discriminante) des données de la classe positive et les données de la classe négative les plus proches de la frontière de séparation, qui est représentée par une fonction de classification égale à zéro.

La marge de séparation est calculée par la différence entre la valeur minimale de la fonction de classification des observations positives et la valeur maximale de la fonction de classification des observations négatives, voir la formule de calcul (4).

$$\mathbf{MS} = \left| \mathbf{D^{+}_{min} - D^{-}_{max}} \right| \quad \mathbf{(4)}$$

Avec : D^{+}_{min} : Valeur minimale de la fonction discriminante des observations positives.

D^{-}_{max} : Valeur maximale de la fonction discriminante des observations négatives.

3.2.1 Protocole des expérimentations

Afin de réaliser des expériences aisément reproductibles, et d'éviter certains biais, nous avons fixé un protocole de test qui est en fait très proche de celui de Batista et Monard [76]. Pour chacune des bases de données disponibles nous commençons par créer (n) paires de bases de données apprentissage-test selon la procédure de validation croisée qui consiste à répéter plusieurs fois, sous des configurations pré définies, le schéma apprentissage test (fig. 3.3). Ensuite, pour les (n) sous bases de données d'apprentissage, en introduit artificiellement des valeurs manquantes selon le mécanisme MCAR, en fixant un certain taux de données manquantes (par exemple pour un taux 10 % des valeurs manquantes, on enlève d'une façon aléatoire 10 % du total des valeurs da la base de données d'apprentissage).

Les bases de test ne sont pas modifiées. Chacune des bases d'apprentissage est ensuite complétée par l'une des techniques d'imputation que nous souhaitons évaluer. Enfin pour chaque taux de données manquantes, un modèle de classification est construit avec le logiciel cbmLAD, pour chacune des bases de données d'apprentissage complètes qui lui sont associées. Il est alors évalué sur la base de données de test correspondant.

Le protocole a été conçu comme le montrent les figures 3.1, 3.2 et 3.3

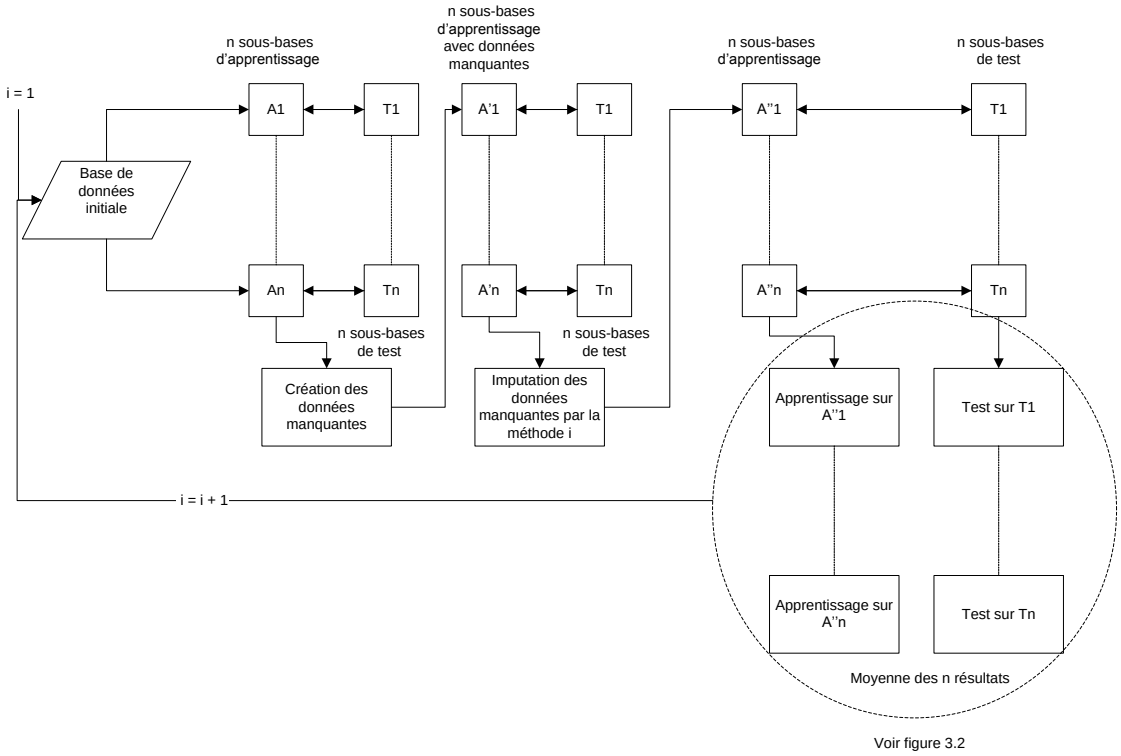


Figure 3.1- Protocole pour l'évaluation des performances de chaque technique de substitution des valeurs manquantes

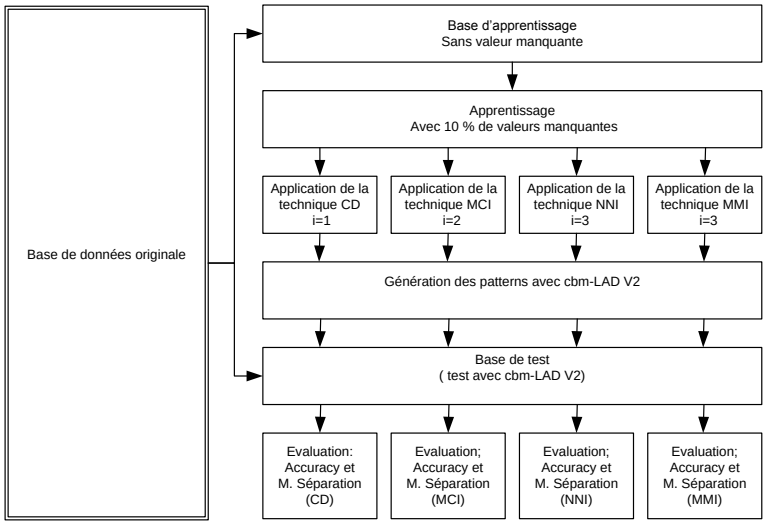
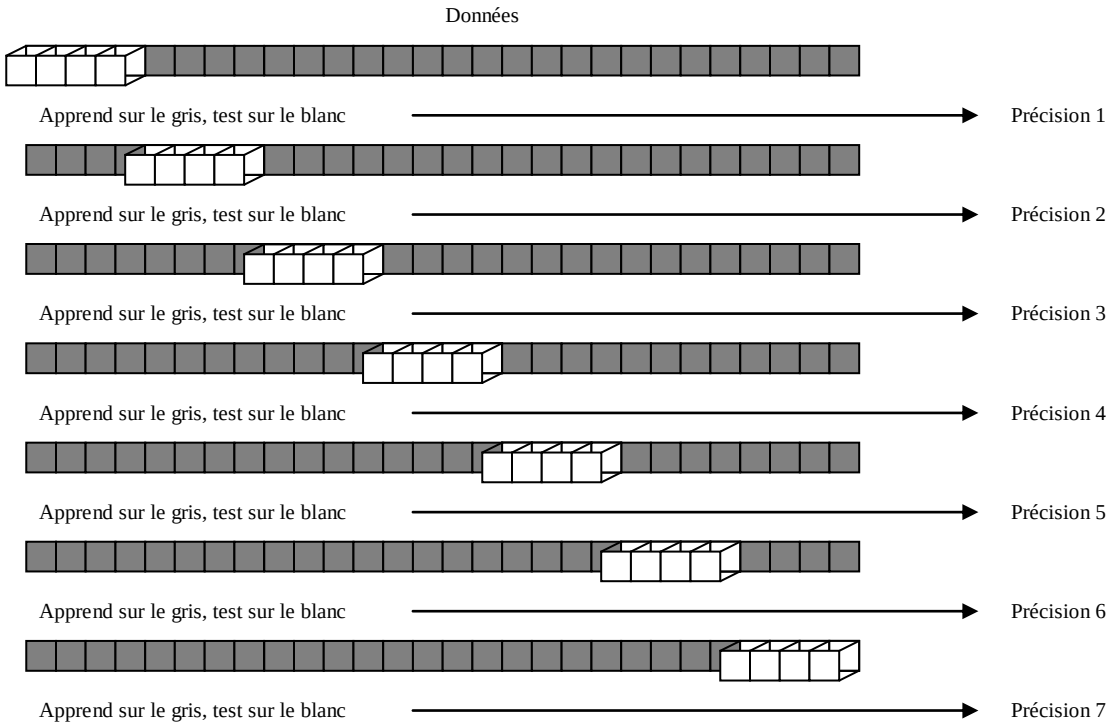


Figure 3.2- Exemple de configuration de test avec 10 % de valeurs manquantes



Précision moyenne = (Précision1 + Précision2 + Précision3 + …..+ Précision7)/ 7

Figure 3.3- Principe de la validation croisée avec n= blocks, (n = 7).

3.2.2 Méthode d'analyse des résultats

Notre objectif étant d'obtenir le meilleur taux de précision et la meilleure marge de séparation, nous comparons en fait les performances des classificateurs construits sur les bases de données complétées par les différentes méthodes d'imputation.

Nous utilisons le taux de précision (Accuracy en anglais et noté Acc dans la suite) pour comparer les différentes méthodes d'imputation. Cette mesure est mieux adaptée pour les problèmes dans lesquels les observations sont également réparties entre les différentes classes, c'est celle qui est utilisée dans les études empiriques de classification supervisée avec la méthode d'analyse logique des données (LAD). Pour les cas dans lesquels les observations sont inégalement réparties entre les différentes classes, la mesure la mieux adaptée est la moyenne des taux de reconnaissance de chaque classe (Balanced Accuracy pour les anglo-saxons), notée BalAcc [78]. Cette méthode n'a pas été utilisée, parce que la différence de répartition entre les classes n'est pas significative.

Pour avoir une analyse complète, nous utilisons aussi la marge de séparation entre les deux classes à la sortie du classificateur.

En outre, inclure au moins deux mesures d'évaluation dans le protocole expérimental nous permet une meilleure comparaison des techniques d'imputation. Pour pouvoir conclure quant à la supériorité d'une technique sur une autre, il nous faut nous appuyer sur des tests d'analyse statistique qui nous permettent de juger du caractère significatif des différences entre les résultats observés.

Nous aurons aussi plusieurs techniques à comparer en même temps, chacune ayant été évaluée sur plusieurs bases de données. Nous nous trouvons donc face à un problème de comparaison multiple.

Ayant introduit une nouvelle technique d'imputation des valeurs manquantes, il importe essentiellement de voir comment celle-ci se comporte par rapport aux techniques existantes, donc nous devons nous appuyer sur une certaine méthodologie statistique pour nous assurer que nos conclusions sont suffisamment fiables, et garantir pour une

probabilité d'erreur fixée que les différences observées reflètent des différences effectives entre les méthodes et qu'elles ne sont pas dues à une variation normale.

Comme méthodologie d'analyse statistique, nous avons adopté le test des rangs[1] appliqué au cas d'échantillons appariés (Wilcoxon matched-pairs signed-ranks test) [79] qu'est un test non paramétrique des rangs appariés pour des données appartenant à deux échantillons dépendants et n'ayant pas une distribution normale, et le test d'analyse de variance de Friedman[2] [80] [81], un équivalent non paramétrique d'ANOVA (ANalysis Of VAriance) car il ne fait aucune hypothèse sur la forme des distributions sous-jacentes. En fait le test de Friedman consiste à appliquer l'ANOVA sur les rangs des performances des classificateurs par technique de substitution, plutôt que directement sur leurs indices de performance.

Le test non paramétrique de Friedman nous permet de voir si l'on peut rejeter l'hypothèse nulle selon laquelle toutes les techniques mises en balance ont les mêmes performances. Si tel est le cas, il nous faudra utiliser le test non-paramétrique post hoc, une procédure de comparaison multiple est suggérée par [82], pour savoir quelles techniques se différencient.

Les tests non-paramétriques sont utilisés dès que l'effectif N d'un échantillon est inférieur à 30. En dessous de N=30, les tests nécessitent certaines hypothèses (normalité des distributions, égalité des variances, etc.). Ceci est particulièrement vrai quand les effectifs sont très faibles (N= 12 pour la base DGA, N= 12 pour la base Bearing et N=9 pour la base IRIS).

C'est dans ce cadre des petits échantillons que l'on a recours aux méthodes non paramétriques, dites encore « indépendantes de la distribution », qui reposent non pas sur les valeurs de la variable quantitative observée, mais sur les rangs qu'elles occupent dans la distribution. Les méthodes non paramétriques sont donc toujours valables, même

---

[1] Pour de plus amples explications concernant les tests statistiques utilisés : 1

[2] Pour de plus amples explications concernant les tests statistiques utilisés : 1

quand les distributions sont normales. Toutefois, elles sont alors moins efficaces que les tests paramétriques, ce qui implique que pour obtenir une même puissance les échantillons doivent être d'effectif plus élevé.

Afin de présenter des résultats synthétiques, nous avons donc décidé de considérer la moyenne du taux de précision et la marge de séparation de l'ensemble des bases de données, obtenues en appliquant les différentes méthodes d'imputation, et ensuite calculer les rangs correspondants, et ce, pour chaque technique d'imputation et chaque critère de performance.

3.2.3 Description des bases de données

Les expérimentations ont été menées sur trois bases de données réelles qui se divisent en deux catégories : Deux bases de données chronologiques et une base de données non chronologique, nous avons fait ce choix pour pouvoir vérifier le comportement des méthodes de substitution étudiées, pour les deux types de données.

Vu qu'il y a une difficulté à trouver une base de données non chronologique du domaine de la maintenance conditionnelle, on a utilisé la base de données IRIS (classification de trois types de fleur IRIS), parce que cette base est utilisée dans plusieurs études de classification supervisée, donc elle sera utilisée dans notre cas pour simuler le cas des observations non chronologiques. Pour les données chronologiques, on a utilisé deux bases de données du domaine d'ingénierie. Une description détaillée de chaque base de données s'en suit.

3.2.3.1 L'analyse des gaz dissous dans l'huile des transformateurs

Les gaz existants dans les transformateurs électriques sont produits par la dégradation de l'huile et d'autres matériaux [83]. Ces gaz entourent la source du défaut, puis se dissolvent ensuite dans l'huile, du fait de leur grande solubilité.

L'analyse de gaz dissous (Dissolved Gas *Analyses* -DGA) dans l'huile des transformateurs est le meilleur indicateur de l'état général d'un transformateur, puisqu'en présence d'une dégradation, la concentration des gaz augmente d'une manière

significative. Les gaz qui sont pris généralement en compte sont décrits dans le tableau 3.1.

Tableau 3.1- Les indicateurs de la base de données DGA.

Gas	Nom	Valeur nominale
H_2	hydrogène	100 ppm
CH_4	Méthane	120 ppm
C_2H_6	Ethane	65 ppm
C_2H_4	Ethylene	50 ppm
C_2H_2	Acétylène	35 ppm

Le DGA est une manière non destructive de diagnostiquer l'état du transformateur en analysant le volume des combustibles gaz dissous dans l'huile, le tableau 3.2 présente la base de données en question.

Étant donné deux types défauts, le positif (P) et le négatif (N) est simplement pour différencier deux classes (P veut dire la classe positive et N veut dire la classe négative), et n'a pas le sens utilisé au paravent en mathématique.

Tableau 3.2- La base de données analyse des gaz dissous dans l'huile (DGA)

Observations	H2	CH4	C2H6	C2H4	C2H2	Type de défaut	Classe
P1	24	13	43	5	319	Arcing	1
P2	127	24	32	0	81	Arcing	1
P3	9474	4066	6552	353	12997	Arcing	1
P4	441	207	224	43	261	Arcing	1
P5	212	38	47	15	78	Arcing	1
P6	800	1393	2817	304	3000	Arcing	1
P7	858	1324	2793	208	7672	Arcing	1
P8	274	27	33	5	97	Arcing	1
P9	1249	370	606	56	1371	Arcing	1
P10	307	22	33	2	109	Arcing	1
P11	127	107	154	11	224	Arcing	1
N1	266	584	862	328	1	Overheating	0
N2	65	61	143	16	3	Overheating	0
N3	16	87	395	75	30	Overheating	0
N4	199	770	1508	217	72	Overheating	0
N5	244	754	1281	172	27	Overheating	0
N6	117	167	481	48	7	Overheating	0
N7	137	369	1242	144	16	Overheating	0
N8	33	79	215	30	5	Overheating	0
N9	60	144	449	67	9	Overheating	0

3.2.3.2 Analyse vibratoire des roulements

La performance des roulements d'un moteur est très influente sur les performances du système moteur complet. Plus précisément, la présence de défauts de roulements se traduit souvent par une efficacité réduite, ou même des dommages graves. Afin de déterminer quand il est nécessaire de prendre un moteur hors-ligne pour la maintenance préventive, ces défauts doivent être diagnostiqués.

Chaque palier à roulement possède des fréquences de dommages caractéristiques pour la bague intérieure, le corps de roulement et la bague extérieure. Ces fréquences dépendent de la vitesse de rotation, des dimensions géométriques et du nombre de corps de roulement.

Un défaut peut être quantifié en mesurant l'amplitude de la fréquence du signal relevée aux fréquences de défaut de roulement ou à leurs harmoniques. L'évolution de la dégradation d'un défaut de roulement provoquera non seulement l'augmentation de l'amplitude de la vibration aux fréquences de roulements, mais générera également des vibrations aux harmoniques de ces fréquences ainsi qu'à des fréquences connexes dues à la modulation d'amplitude

Une machine ayant un roulement défectueux peut générer au moins cinq fréquences caractéristiques :

• Fréquence de rotation de l'arbre (f_s).

• Fréquence fondamentale du train FTF.

• Fréquence de passage des billes sur la piste extérieure BPFO. Cette fréquence est apparente dès le deuxième stade de dégradation, car la bague externe est plus proche du capteur.

• Fréquence de passage des billes sur la piste intérieure BPFI. Cette fréquence devient plus significative lorsque le défaut est avancé.

• Deux fois la fréquence de rotation des billes (2 X BSF). En effet, cette fréquence apparaîtra souvent à sa deuxième harmonique, car la bille est excitée deux fois par tour lorsqu'elle tourne sur elle-même, l'impact ayant lieu sur la bague externe et interne.

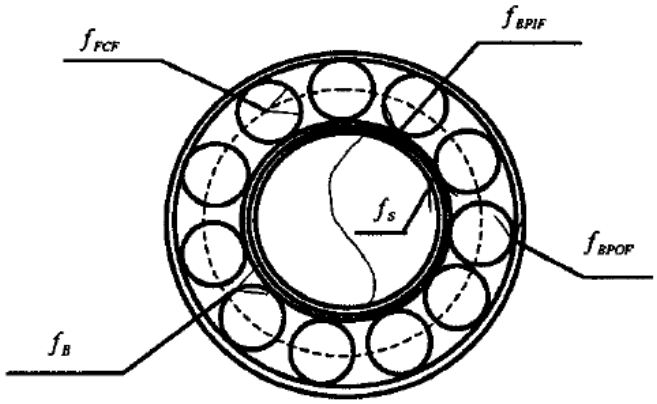


Figure 3.4- Les fréquences de base dans un palier.

Dans l'article de G. Goddu 1998 [84], le spectre de fréquence du signal de vibration de roulement est analysé pour la détection des différents défauts du moteur.

Les observations de la base de données en question, sont définies par le vecteur des attributs suivant : $X= [X_L, X_D, X_T]$.

Avec $X_L = X_{LOOSE}$, $X_D = X_{DAMAGE}$ et $X_T = X_{TIME}$ comme indicateurs de cette base de données et qui sont on fonction des fréquences mesurées (définie à la page avant).

Pour notre étude, on a retenu les deux classes suivantes (parce que le cbmLAD a seulement deux classes), desserrage (LOOSE) qui représente une situation facilement corrigible et dommage *(DAMAGE)* qui représente une situation où on doit remplacer le roulement, le tableau 3.3 présente la base de données en question.

Étant donné deux types défauts, le positif (P) et le négatif (N) est simplement pour différencier deux classes (P veut dire la classe positive et N veut dire la classe négative), et n'a pas le sens utilisé auparavant en mathématique.

Tableau 3.3- La base de données analyse vibratoire des roulements [82].

Observations	XL	XD	XT	Situation	Classe
P1	0.124	0.167	0.00749	Loose	1
P2	0.105	0.378	0.0765	Loose	1
P3	0.106	0.373	0.169	Loose	1
P4	0.108	0.372	0.214	Loose	1
P5	0.0132	0.665	0.0263	Loose	1
P6	0.0897	0.435	0.101	Loose	1
P7	0.103	0.393	0.189	Loose	1
P8	0.104	0.389	0.237	Loose	1
P9	0.00385	0.709	0.0518	Loose	1
P10	0.0544	0.549	0.127	Loose	1
P11	0.0861	0.431	0.227	Loose	1
P12	0.0958	0.426	0.266	Loose	1
P13	0.00407	0.721	0.0633	Loose	1
P14	0.0427	0.577	0.138	Loose	1
P15	0.0806	0.462	0.232	Loose	1
P16	0.0887	0.44	0.282	Loose	1
N1	0.848	0.059	0.0297	Damage	0
N2	0.388	0.253	0.079	Damage	0
N3	0.212	0.33	0.163	Damage	0
N4	0.174	0.348	0.216	Damage	0
N5	0.634	0.225	0.0479	Damage	0
N6	0.334	0.324	0.1	Damage	0
N7	0.197	0.351	0.195	Damage	0
N8	0.17	0.362	0.233	Damage	0
N9	0.345	0.451	0.0686	Damage	0
N10	0.255	0.42	0.12	Damage	0
N11	0.173	0.397	0.215	Damage	0
N12	0.153	0.393	0.252	Damage	0
N13	0.26	0.522	0.0828	Damage	0
N14	0.204	0.48	0.126	Damage	0
N15	0.159	0.42	0.228	Damage	0
N16	0.141	0.421	0.269	Damage	0
N17	0.861	0.0499	0.0598	Damage	0
N18	0.627	0.147	0.111	Damage	0
N19	0.384	0.257	0.163	Damage	0

N20	0.319	0.281	0.206	Damage	0
N21	0.792	0.0985	0.0743	Damage	0
N22	0.595	0.182	0.104	Damage	0
N23	0.368	0.269	0.183	Damage	0
N24	0.308	0.293	0.232	Damage	0
N25	0.629	0.225	0.0972	Damage	0
N26	0.501	0.263	0.12	Damage	0
N27	0.39	0.312	0.21	Damage	0
N28	0.283	0.324	0.253	Damage	0
N29	0.551	0.285	0.114	Damage	0
N30	0.447	0.311	0.136	Damage	0
N31	0.318	0.333	0.225	Damage	0
N32	0.272	0.346	0.26	Damage	0
N33	0.866	0.0451	0.0732	Damage	0
N34	0.702	0.116	0.117	Damage	0
N35	0.473	0.213	0.165	Damage	0
N36	0.385	0.251	0.206	Damage	0
N37	0.818	0.0813	0.0899	Damage	0
N38	0.671	0.138	0.117	Damage	0
N39	0.456	0.227	0.181	Damage	0
N40	0.38	0.259	0.222	Damage	0
N41	0.7	0.172	0.11	Damage	0
N42	0.589	0.212	0.14	Damage	0
N43	0.42	0.272	0.2	Damage	0
N44	0.353	0.29	0.253	Damage	0
N45	0.632	0.227	0.124	Damage	0
N46	0.537	0.254	0.147	Damage	0
N47	0.4	0.291	0.216	Damage	0
N48	0.343	0.308	0.26	Damage	0

3.2.3.3 Classification des fleurs iris de Fisher

Les données ont été recueillies par Edgar Anderson [85]. Ce sont les mesures en centimètres des variables suivantes : longueur du sépale (Sepal.Length), largeur du sépale (Sepal.Width), longueur du pétale (Petal.Length) et largeur du pétale (Petal.Width) pour trois espèces d'iris : Iris setosa, I. versicolor et I. virginica.

Dans le traitement du fameux fichier IRIS (Fisher, 1936) [86], on cherche à produire un regroupement en 3 classes des iris à partir de leur morphologie, la figure 3.4 montre les trois classes des fleurs iris.

Figure 3.5- Les trois classes des fleurs iris (Iris setosa, Iris versicolor et Iris virginica)[3].

Pour notre étude, on a retenu les deux classes suivantes, ***Iris setosa*** et ***Iris versicolor.*** On n'a pas retenu la 3eme classe parce que, le logiciel cbm-LAD ne traite que deux classes, le tableau 3.4 présente la base de données en question.

Tableau 3.4- La base de données fleurs iris de Fisher avec deux classes.

Observations	longueur du sépale	largeur du sépale	longueur du pétale	largeur du pétale	Type d'iris	Classe
P1	5.1	3.5	1.4	0.2	Setosa	1
P2	4.9	3	1.4	0.2	Setosa	1
P3	4.7	3.2	1.3	0.2	Setosa	1
P4	4.6	3.1	1.5	0.2	Setosa	1
P5	5	3.6	1.4	0.2	Setosa	1
P6	5.4	3.9	1.7	0.4	Setosa	1
P7	4.6	3.4	1.4	0.3	Setosa	1
P8	5	3.4	1.5	0.2	Setosa	1
P9	4.4	2.9	1.4	0.2	Setosa	1
P10	4.9	3.1	1.5	0.1	Setosa	1
P11	5.4	3.7	1.5	0.2	Setosa	1
P12	4.8	3.4	1.6	0.2	Setosa	1
P13	4.8	3	1.4	0.1	Setosa	1
P14	4.3	3	1.1	0.1	Setosa	1
P15	5.8	4	1.2	0.2	Setosa	1

[3] http://en.wikipedia.org/wiki/Iris_flower_data_set

P16	5.7	4.4	1.5	0.4	Setosa	1
P17	5.4	3.9	1.3	0.4	Setosa	1
P18	5.1	3.5	1.4	0.3	Setosa	1
P19	5.7	3.8	1.7	0.3	Setosa	1
P20	5.1	3.8	1.5	0.3	Setosa	1
P21	5.4	3.4	1.7	0.2	Setosa	1
P22	5.1	3.7	1.5	0.4	Setosa	1
P23	4.6	3.6	1	0.2	Setosa	1
P24	5.1	3.3	1.7	0.5	Setosa	1
P25	4.8	3.4	1.9	0.2	Setosa	1
P26	5	3	1.6	0.2	Setosa	1
P27	5	3.4	1.6	0.4	Setosa	1
P28	5.2	3.5	1.5	0.2	Setosa	1
P29	5.2	3.4	1.4	0.2	Setosa	1
P30	4.7	3.2	1.6	0.2	Setosa	1
P31	4.8	3.1	1.6	0.2	Setosa	1
P32	5.4	3.4	1.5	0.4	Setosa	1
P33	5.2	4.1	1.5	0.1	Setosa	1
P34	5.5	4.2	1.4	0.2	Setosa	1
P35	4.9	3.1	1.5	0.2	Setosa	1
P36	5	3.2	1.2	0.2	Setosa	1
P37	5.5	3.5	1.3	0.2	Setosa	1
P38	4.9	3.6	1.4	0.1	Setosa	1
P39	4.4	3	1.3	0.2	Setosa	1
P40	5.1	3.4	1.5	0.2	Setosa	1
P41	5	3.5	1.3	0.3	Setosa	1
P42	4.5	2.3	1.3	0.3	Setosa	1
P43	4.4	3.2	1.3	0.2	Setosa	1
P44	5	3.5	1.6	0.6	Setosa	1
P45	5.1	3.8	1.9	0.4	Setosa	1
P46	4.8	3	1.4	0.3	Setosa	1
P47	5.1	3.8	1.6	0.2	Setosa	1
P48	4.6	3.2	1.4	0.2	Setosa	1
P49	5.3	3.7	1.5	0.2	Setosa	1
P50	5	3.3	1.4	0.2	Setosa	1
N1	7	3.2	4.7	1.4	Versicolor	0
N2	6.4	3.2	4.5	1.5	Versicolor	0

N3	6.9	3.1	4.9	1.5	Versicolor	0
N4	5.5	2.3	4	1.3	Versicolor	0
N5	6.5	2.8	4.6	1.5	Versicolor	0
N6	5.7	2.8	4.5	1.3	Versicolor	0
N7	6.3	3.3	4.7	1.6	Versicolor	0
N8	4.9	2.4	3.3	1	Versicolor	0
N9	6.6	2.9	4.6	1.3	Versicolor	0
N10	5.2	2.7	3.9	1.4	Versicolor	0
N11	5	2	3.5	1	Versicolor	0
N12	5.9	3	4.2	1.5	Versicolor	0
N13	6	2.2	4	1	Versicolor	0
N14	6.1	2.9	4.7	1.4	Versicolor	0
N15	5.6	2.9	3.6	1.3	Versicolor	0
N16	6.7	3.1	4.4	1.4	Versicolor	0
N17	5.6	3	4.5	1.5	Versicolor	0
N18	5.8	2.7	4.1	1	Versicolor	0
N19	6.2	2.2	4.5	1.5	Versicolor	0
N20	5.6	2.5	3.9	1.1	Versicolor	0
N21	5.9	3.2	4.8	1.8	Versicolor	0
N22	6.1	2.8	4	1.3	Versicolor	0
N23	6.3	2.5	4.9	1.5	Versicolor	0
N24	6.1	2.8	4.7	1.2	Versicolor	0
N25	6.4	2.9	4.3	1.3	Versicolor	0
N26	6.6	3	4.4	1.4	Versicolor	0
N27	6.8	2.8	4.8	1.4	Versicolor	0
N28	6.7	3	5	1.7	Versicolor	0
N29	6	2.9	4.5	1.5	Versicolor	0
N30	5.7	2.6	3.5	1	Versicolor	0
N31	5.5	2.4	3.8	1.1	Versicolor	0
N32	5.5	2.4	3.7	1	Versicolor	0

N33	5.8	2.7	3.9	1.2	Versicolor	0
N34	6	2.7	5.1	1.6	Versicolor	0
N35	5.4	3	4.5	1.5	Versicolor	0
N36	6	3.4	4.5	1.6	Versicolor	0
N37	6.7	3.1	4.7	1.5	Versicolor	0
N38	6.3	2.3	4.4	1.3	Versicolor	0
N39	5.6	3	4.1	1.3	Versicolor	0
N40	5.5	2.5	4	1.3	Versicolor	0
N41	5.5	2.6	4.4	1.2	Versicolor	0
N42	6.1	3	4.6	1.4	Versicolor	0
N43	5.8	2.6	4	1.2	Versicolor	0
N44	5	2.3	3.3	1	Versicolor	0
N45	5.6	2.7	4.2	1.3	Versicolor	0
N46	5.7	3	4.2	1.2	Versicolor	0
N47	5.7	2.9	4.2	1.3	Versicolor	0
N48	6.2	2.9	4.3	1.3	Versicolor	0
N49	5.1	2.5	3	1.1	Versicolor	0
N50	5.7	2.8	4.1	1.3	Versicolor	0

En résumé, les principales caractéristiques de chacune des bases sont décrites dans le tableau 3.5.

Tableau 3.5- Description des bases de données utilisées.

Nom	DGA *	Bearing	IRIS *
Nombre de classe	3	2	3
Nombre d'observation	30	64	150
Nombre d'attribut	5	3	4
Valeurs manquantes	Non	Non	Non
Nombre de classes utilisé	2	2	2
Nombre d'observations utilisé	20	64	100
Type d'attribut	Réelle	Réelle	Réelle
Propriété	Chronologique	Chronologique	Non chronologique
Référence	[83]	[84]	[85]

(* : On a utilisée seulement deux classes, puisque cbmLAD ne traite que deux classes)

3.3 Résultats expérimentaux et discussions

3.3.1 Base de données d'analyse des gaz dissous (DGA)

L'ensemble de bases de données de **DGA** (Dissolved gas analysis**),** contient 20 observations. Chaque observation est classifiée comme défaut d'arc électrique **(Arcing)** ou défaut de surchauffage **(Overheating)**. Dans notre cas, l'ensemble des observations positives est l'ensemble des observations de la classe défaut d'arc électrique **(Arcing**) et l'ensemble d'observations négatives est l'ensemble des observations de la classe du défaut de surchauffage **(Overheating)**. Chaque observation est décrite à l'aide de 5 attributs (indicateurs).

L'ensemble des données appartenant à la classe positive dans la base des données de l'apprentissage et de test sont 8 et 3 respectivement (72 % pour l'apprentissage contre 28 % pour le test), et l'ensemble des données appartenant à la classe négative dans la base des données de l'apprentissage et de test sont 6 et 3 respectivement (66 % pour l'apprentissage contre 34 % pour le test), donc le total des observations sera 20 observations pour les deux classes utilisées.

Suivant le principe de la validation croisée exposé au paragraphe 3.2.1, on a construit les différentes paires de bases de données apprentissage-test, résultats présentés dans le tableau 3.6.

Tableau 3.6- Les bases d'apprentissage-test pour DGA transformer Data set.

Test	Les observations de la base d'apprentissage Total = 14	Les observations de la base de test Total = 6
1	P : 4 ----> 11 ; N : 4 ---->9	P :1 ----> 3 ; N : 1 ---->3
2	P : 4 ----> 11 ; N : 1---->3 + 7---->9	P :1 ----> 3 ; N : 4 ---->6
3	P : 4 ----> 11 ; N : 1---->6	P :1 ----> 3 ; N : 7 ---->9
4	P : 1 ----> 3 + 7----> 11 ; N : 4 ---->9	P :4----> 6 ; N : 1 ---->3
5	P : 1---> 3 + 7---> 11; N : 1---->3 + 7---->9	P : 4----> 6 ; N : 4 ---->6
6	P : 1---> 3 + 7---> 11; N : 1---->6	P : 4----> 6 ; N : 7 ---->9
7	P : 1 ----> 6 + 10----> 11 ; N : 4 ---->9	P :7----> 9 ; N : 1 ---->3
8	P : 1 ----> 6 + 10---> 11; N : 1-->3 + 7--->9	P : 7----> 9 ; N : 4 ---->6
9	P : 1 ----> 6 + 10---> 11; N : 1---->6	P : 7----> 9 ; N : 7 ---->9
10	P : 2 ----> 9 ; N : 4 ---->9	P :10----> 1 ; N : 1 ---->3
11	P : 2 ----> 9; N : 1-->3 + 7--->9	P : 10----> 1 ; N : 4 ---->6
12	P : 2 ----> 9 ; N : 1---->6	P : 10----> 1 ; N : 7 ---->9

Pour la base de données **DGA**, nous avons 12 paires de bases apprentissage-test, on a testé les 4 taux de valeurs manquantes suivants : 5 %, 10 %, 20 %, 40 %, créant ainsi 192 sous-bases de données d'apprentissage, pour le test des 4 méthodes de substitution.
Le nombre total des bases d'apprentissage est calculé de la façon suivante :
Le nombre total des bases = (nombre de taux de valeur manquante) X (nombre de paires de bases apprentissage-test) X (nombre de méthodes de substitution)
Le nombre total des bases = 4*12*4 = 192 sous bases de données d'apprentissage.

3.3.1.1 Résultats

Les résultats obtenus avec un même taux de données manquantes et pour une même sous-base de données d'apprentissage, mais avec différentes techniques de substitution sont présentés dans le tableau 3.7 et la figure 3.6 pour le taux de précision (Accuracy), et dans le tableau 3.8 et la figure 3.7 pour la mesure de séparation.

Les résultats du taux de précision (Accuracy), correspondant à une même technique de substitution, sont moyennés sur 12 tests (le nombre total des bases d'apprentissage-test tableau 3.6) suivant le même protocole de test décrit dans la section 3.2.1.

Chaque colonne correspond à une technique de substitution (CD, MCI, NNI, MMI), parmi les résultats de la précision (séparation) dans une même colonne, plus la précision (séparation) d'une méthode de substitution est grand, meilleure est la méthode.

Nota : les résultats présentés dans ce chapitre sont obtenus par le logiciel cbmLAD V2, développé en 2010, par l'équipe de recherche de Mme Soumaya Yacout.

Test	Acc Sans valeurs manquantes	Acc avec la méthode CD				Acc avec la méthode MCI				Acc avec la méthode NNI				Acc avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	100.00%	100.00%	100.00%	66.67%	66.67%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	83.33%
T2	83.33%	33.33%	33.33%	66.67%	66.67%	83.33%	83.33%	100.00%	100.00%	83.33%	83.33%	83.33%	66.67%	83.33%	83.33%	83.33%	83.33%
T3	100.00%	100.00%	100.00%	83.33%	83.33%	100.00%	100.00%	100.00%	83.33%	100.00%	100.00%	83.33%	83.33%	100.00%	100.00%	100.00%	83.33%
T4	100.00%	100.00%	66.67%	66.67%	66.67%	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	66.67%	66.67%	100.00%	100.00%	66.67%	66.67%
T5	100.00%	83.33%	66.67%	66.67%	66.67%	100.00%	100.00%	100.00%	66.67%	100.00%	66.67%	83.33%	66.67%	100.00%	83.33%	83.33%	66.67%
T6	100.00%	100.00%	66.67%	83.33%	83.33%	100.00%	100.00%	100.00%	83.33%	100.00%	100.00%	83.33%	66.67%	100.00%	100.00%	100.00%	83.33%
T7	100.00%	100.00%	66.67%	66.67%	66.67%	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	66.67%	66.67%
T8	100.00%	83.33%	66.67%	66.67%	66.67%	100.00%	83.33%	83.33%	66.67%	100.00%	83.33%	83.33%	66.67%	100.00%	83.33%	83.33%	66.67%
T9	100.00%	100.00%	66.67%	66.67%	66.67%	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	66.67%
T10	100.00%	100.00%	100.00%	50.00%	50.00%	100.00%	100.00%	50.00%	50.00%	100.00%	100.00%	50.00%	50.00%	100.00%	100.00%	50.00%	50.00%
T11	83.33%	33.33%	33.33%	50.00%	50.00%	83.33%	83.33%	66.67%	66.67%	83.33%	83.33%	66.67%	83.33%	83.33%	83.33%	83.33%	83.33%
T12	100.00%	100.00%	100.00%	50.00%	50.00%	100.00%	100.00%	83.33%	100.00%	100.00%	100.00%	83.33%	66.67%	100.00%	100.00%	100.00%	100.00%
Moyenne	97.22%	86.11%	72.22%	65.28%	65.28%	97.22%	95.83%	90.28%	76.39%	97.22%	93.06%	81.94%	73.61%	97.22%	94.44%	84.72%	75.00%

Tableau 3.7- Résultats de Acc pour les méthodes CD – MCI – NNI et MMI de la base de données DGA.

Test	M.S. Sans valeurs manquantes	Séparation avec la méthode CD				Séparation avec la méthode MCI				Séparation avec la méthode NNI				Séparation avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	0.35	0.33	0.33	1.00	1.00	0.35	0.33	0.33	0.33	0.67	0.67	0.33	0.33	0.33	0.64	1.00	0.67
T2	1.33	0.67	0.67	1.00	1.00	1.33	1.33	0.33	0.33	1.33	1.33	1.33	0.33	1.33	1.33	1.33	0.33
T3	1.33	1.67	1.33	0.33	0.33	1.33	1.33	1.33	0.67	1.67	1.67	1.67	0.67	1.33	1.33	1.33	0.67
T4	0.92	0.33	1.00	1.00	1.00	0.92	0.92	0.67	1.00	0.67	0.67	1.00	1.00	0.93	0.51	1.00	1.00
T5	0.67	0.67	0.33	1.00	1.00	0.97	0.67	1.00	1.33	0.67	1.96	1.33	1.33	0.67	1.74	1.66	1.33
T6	1.63	1.67	1.33	1.00	1.00	1.63	1.63	1.33	1.00	1.33	1.33	2.00	1.00	1.63	1.89	1.66	1.00
T7	0.92	0.38	1.00	1.00	1.00	0.92	0.66	1.00	0.91	1.00	0.44	0.33	1.00	0.93	0.82	1.00	1.00
T8	1.00	0.94	1.00	1.00	1.00	1.00	1.63	1.66	1.33	1.00	1.96	1.33	1.33	1.00	1.89	1.67	1.33
T9	1.96	1.71	2.00	0.33	0.33	1.96	1.96	1.33	1.00	2.00	1.44	1.92	1.00	1.97	1.89	1.67	1.00
T10	1.33	0.33	0.33	0.00	0.00	0.33	0.33	2.00	2.00	1.33	1.33	2.00	2.00	1.33	1.35	2.00	2.00
T11	1.33	0.00	0.00	0.00	0.00	1.34	1.34	1.33	1.33	1.34	1.34	1.06	0.67	1.34	1.34	1.01	0.69
T12	1.33	1.33	1.33	0.00	0.00	1.33	1.33	1.33	0.27	1.34	1.34	1.29	1.15	1.34	1.34	1.01	0.36
Moyenne	1.12	0.93	1.00	0.85	0.85	1.15	1.16	1.00	0.88	1.15	1.27	1.25	0.89	1.12	1.34	1.37	0.93

Tableau 3.8- Résultats de Marge de séparation pour les méthodes CD – MCI – NNI et MMI de la base de données DGA.

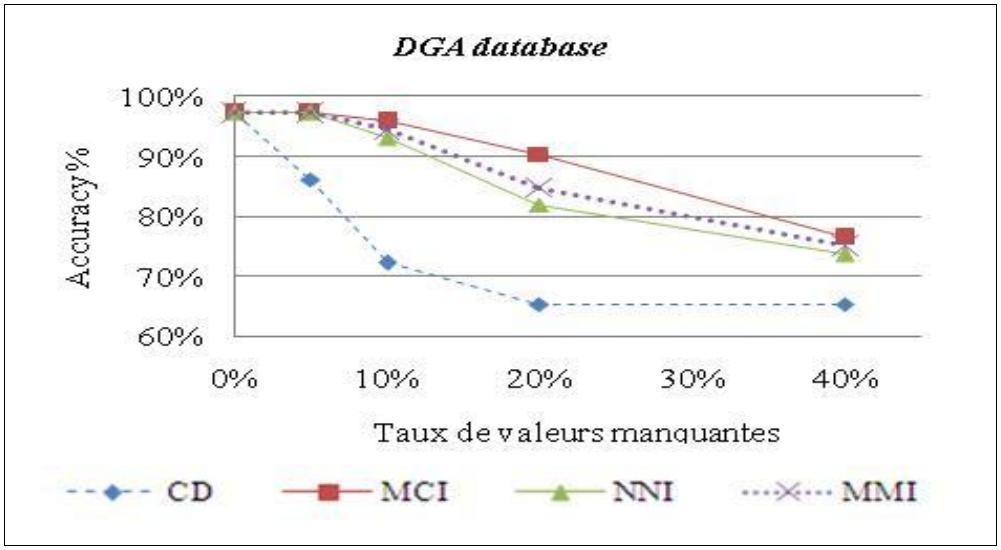


Figure 3.6- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (La précision).

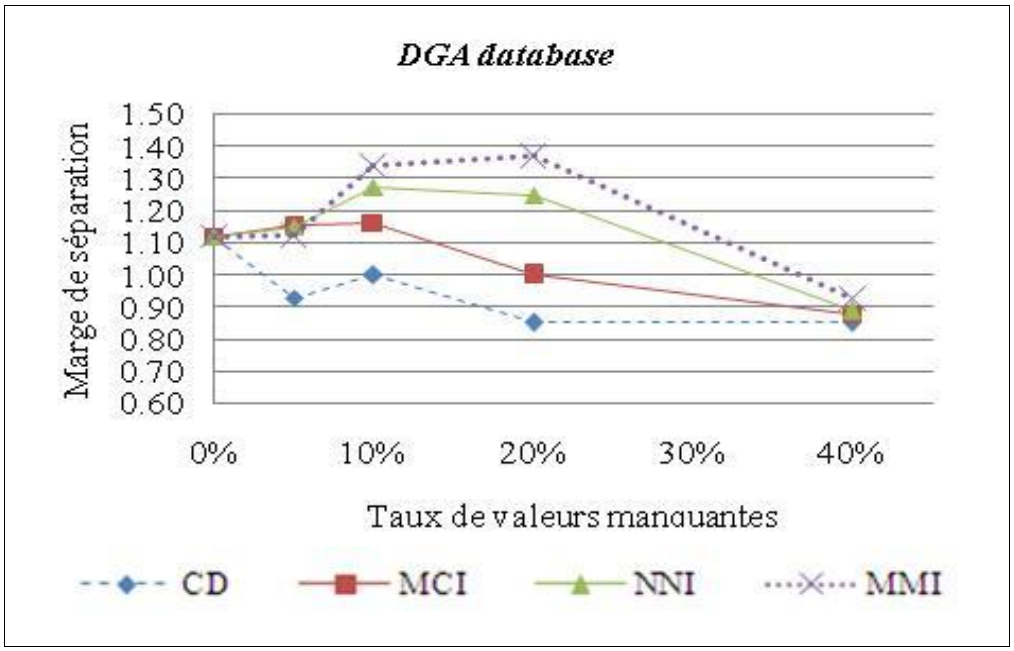


Figure 3.7- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (Marge de séparation)

3.3.1.2 Test T de Wilcoxon

Une série de tests de Wilcoxon[4] pour données appariées, en anglais connu sous le nom de « Wilcoxon Matched-Pair Signed Ranks », ont été effectués pour déterminer si les différences des résultats pour la précision et la marge de séparation en raison de la modification du taux des valeurs manquantes, étaient appréciables sur le plan statistique. Le test de Wilcoxon permet d'analyser des paires de données et la mesure initiale et finale d'une seule base de données.

L'hypothèse nulle : il n'y a pas de différence entre le résultat de la précision de la base de données sans traitement des valeurs manquantes et le résultat de la même base de données avec traitement de valeurs manquantes.

Le tableau 3.9 résume tous les tests de wilcoxon réalisé pour les différents taux des valeurs manquantes et pour les quatre méthodes d'imputation.

Tableau 3.9- Résumé des tests de Wilcoxon pour la précision (la base de données DGA)

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
Originale / 5% CD	10	0	4	Accepter H0	0.125
Originale / 10% CD	36	0	8	Rejeter H0	0.007812
Originale / 20% CD	78	0	12	Rejeter H0	0.0004883
Originale / 40% CD	78	0	12	Rejeter H0	0.0004883
Originale / 5% MCI	0	0	0	Accepter H0	1
Originale / 10% MCI	1	0	1	Accepter H0	1
Originale / 20% MCI	12.5	2.5	5	Accepter H0	0.3125
Originale / 40% MCI	52.5	2.5	10	Rejeter H0	0.009766
Originale / 5% NNI	0	0	0	Accepter H0	1
Originale / 10% NNI	3	0	2	Accepter H0	0.5
Originale / 20% NNI	36	0	8	Rejeter H0	0.007812
Originale / 40% NNI	45	0	9	Rejeter H0	0.003906
Originale / 5% MMI	0	0	0	Accepter H0	1
Originale / 10% MMI	3	0	2	Accepter H0	0.5
Originale / 20% MMI	15	0	5	Accepter H0	0.0625
Originale / 40% MMI	45	0	9	Rejeter H0	0.003906

[4] Pour de plus amples explications concernant les tests statistiques utilisés, voir annexe n°1

Le tableau 3.9 nous montre que suivant le test de wilcoxon avec un seuil significatif de 5 %, la méthode CD est rejetée même avec un petit taux des valeurs manquantes (à partie de 10 % des valeurs manquantes). Ces premiers résultats montrent le désavantage évident de la méthode CD, elle réduit significativement la taille de la base de données (le nombre des observations), par conséquent, elle élimine de l'information importante pour définir un modèle de classification pour cbm-LAD.

Tableau 3.10- Résumé des tests de Wilcoxon pour la marge de séparation (la base de données DGA)

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
Originale / 5% CD	48	7	10	Rejeter H0	0.03711
Originale / 10% CD	36	9	9	Accepter H0	0.1289
Originale / 20% CD	54	12	11	Accepter H0	0.06738
Originale / 40% CD	54	12	11	Accepter H0	0.06738
Originale / 5% MCI	3	3	3	Accepter H0	1
Originale / 10% MCI	10	5	5	Accepter H0	0.625
Originale / 20% MCI	23	22	9	Accepter H0	1
Originale / 40% MCI	44.5	21.5	11	Accepter H0	0.3203
Originale / 5% NNI	11	25	5	Accepter H0	0.3828
Originale / 10% NNI	22	33	10	Accepter H0	0.625
Originale / 20% NNI	20	46	11	Accepter H0	0.2783
Originale / 40% NNI	50	28	12	Accepter H0	0.4238
Originale / 5% MMI	6	15	6	Accepter H0	0.4375
Originale / 10% MMI	17	38	10	Accepter H0	0.3223
Originale / 20% MMI	15	40	10	Accepter H0	0.2324
Originale / 40% MMI	51.5	26.5	12	Accepter H0	0.3394

Les résultats du tableau 3.10 nous montrent qu'il n'y a pas de différence entre la marge de séparation pour la base originale et la marge de séparation obtenue par chaque méthode de traitement des valeurs manquantes, avec les différents taux de valeurs manquantes pour le seuil de signification fixé à $p<0,05$.

Autrement dit, la différence observée entre les résultats des traitements et la base originale pourrait être attribuable au hasard.

Ceci n'exclut pas que l'on puisse aboutir à une conclusion différente si on teste les grands taux de valeurs manquantes.

On refait le même test de wilcoxon pour comparer les méthodes de traitement deux à deux, le tableau 3.12 représente les résultats de comparaison de la précision à partir du tableau 3.11, et le tableau 3.14 représente les résultats de la marge de séparation à partir du tableau 3.13.

Tableau 3.11- Résultats de la précision par méthode et par taux de valeurs manquantes

Taux de VM	**La précision**			
	CD	**MCI**	**NNI**	**MMI**
0%	97.22%	97.22%	97.22%	97.22%
5%	86.11%	97.22%	97.22%	97.22%
10%	72.22%	95.83%	93.06%	94.44%
20%	65.28%	90.28%	81.94%	84.72%
40%	65.28%	76.39%	73.61%	75.00%

Tableau 3.12- Résultats du test des rangs appariés de Wilcoxon pour la précision

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
CD / MCI	0	10	4	Accepter H0	0.125
CD / NNI	0	10	4	Accepter H0	0.125
CD / MMI	0	10	4	Accepter H0	0.125
MCI / NNI	6	0	3	Accepter H0	0.25
MCI / MMI	6	0	3	Accepter H0	0.25
NNI / MMI	0	6	3	Accepter H0	0.25

Tableau 3.13- Résultats de marge de séparation par méthode et par taux de valeurs manquantes

Taux de VM	**Marge de séparation**			
	CD	**MCI**	**NNI**	**MMI**
0%	1.12	1.12	1.12	1.12
5%	0.93	1.15	1.15	1.12
10%	1.00	1.16	1.27	1.34
20%	0.85	1.00	1.25	1.37
40%	0.85	0.88	0.89	0.93

Tableau 3.14- Résultats du test des rangs appariés de Wilcoxon pour la marge de séparation

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
CD / MCI	0	10	4	Accepter H0	0.125
CD / NNI	0	10	4	Accepter H0	0.125
CD / MMI	0	10	4	Accepter H0	0.125
MCI / NNI	0	6	3	Accepter H0	0.25
MCI / MMI	1	9	4	Accepter H0	0.25
NNI / MMI	1	9	4	Accepter H0	0.25

3.3.1.3 Discussion

La comparaison des différents graphiques de la figure 3.6, correspondants aux résultats obtenus avec la base de données DGA pour un même critère d'évaluation, permet de mettre en évidence l'influence de la méthode de substitution ainsi que le taux des valeurs manquantes.

L'augmentation du taux des valeurs manquantes de 5% à 40% a provoqué une diminution de la précisionde 31.94 % pour la méthode CD, de 20.83% pour la méthode MCI, de 24.61% pour la méthode NNI et de 22.22 % pour la méthode MMI accompagné d'une diminution de la marge de séparation de 24.10 % pour la méthode CD, de 21.42 % pour la méthode MCI, de 20.53% pour la méthode NNI et de 16.96 % pour la méthode MMI par rapport à la précision et la marge de séparation initiale (la base de données sans valeurs manquantes).

La méthode MMI a provoqué une amélioration de a précision et de la marge de séparation par rapport aux autres méthodes.

Les Meilleurs résultats sont obtenus par les méthodes MCI et MMI, même si l'écart entre les moyennes n'est pas significatif suivant les résultats du test de Wilcoxon, pour un seuil de signification de 0.05.

Pour la marge de séparation la méthode MMI se démarque bien par rapport aux autres méthodes.

3.3.2 Base de données analyse vibratoire des roulements

L'ensemble de bases de données analyse vibratoire des roulements, contient 64 observations. Chaque observation est classifiée comme ***Loose*** ou ***Damage***. Dans notre cas, l'ensemble des observations positives est l'ensemble des observations de la classe ***Loose*** et l'ensemble d'observations négatives est l'ensemble des observations de la classe ***Damage***. Chaque observation est décrite à l'aide de 3 attributs (indicateurs).

L'ensemble des données appartenant à la classe positive dans la base des données de l'apprentissage et de test sont 40 et 8 respectivement, et l'ensemble des données appartenant à la classe négative dans la base des données de l'apprentissage et de test sont 8 et 8 respectivement, donc le total des observations sera 64 observations pour les deux classes (74 % pour l'apprentissage et 26 % pour le test).

Suivant le principe de la validation croisée exposé au paragraphe **3.2.1**, on a construit les différentes paires de bases de données apprentissage-test, résultats présentés dans le tableau 3.15.

Tableau 3.15- Les bases d'apprentissage-test pour la base de données analyse vibratoire des roulements

Test	Les observations de la base d'apprentissage Total = 48	Les observations de la base de test Total = 16
1	P : 9 ----> 48 ; N : 9 ---->16	P :1 ----> 8 ; N : 1 ---->8
2	P : 1 ----> 8 + 17---->48 ; N : 9 ---->16	P :9 ----> 16 ; N : 1 ---->8
3	P : 1 ----> 16 + 25---->48 ; N : 9 ---->16	P :17 ----> 24 ; N : 1 ---->8
4	P : 1 ----> 24 + 33---->48 ; N : 9 ---->16	P :25 ----> 32 ; N : 1 ---->8
5	P : 1 ----> 32 + 41---->48 ; N : 9 ---->16	P :33----> 40 ; N : 1 ---->8
6	P : 1 ----> 40 ; N : 9 ---->16	P :41----> 48 ; N : 1 ---->8
7	P : 9 ----> 48 ; N : 1 ---->8	P :1 ----> 8 ; N : 9 ---->16
8	P : 1 ----> 8 + 17---->48 ; N :1 ---->8	P :9 ----> 16 ; N : 9 ---->16
9	P : 1 ----> 16 + 25---->48 ; N : 1 ---->8	P :17 ----> 24 ; N : 9 ---->16
10	P : 1 ----> 24 + 33---->48 ; N : 1 ---->8	P :25 ----> 32 ; N : 9 ---->16
11	P : 1 ----> 32 + 41---->48 ; N : 1 ---->8	P :33----> 40 ; N : 9 ---->16
12	P : 1 ----> 40 ; N : 1 ---->8	P :41----> 48 ; N : 9 ---->16

Pour la base de données analyse vibratoire des roulements, nous avons 12 paires de bases apprentissage-test, on a testé les 4 taux de valeurs manquantes suivants : 5 %, 10 %, 20%, 40%, créant ainsi 192 sous bases de données d'apprentissage, pour le test des 4 méthodes de substitution.

Le nombre total des bases d'apprentissage est calculé de la façon suivante :

Le nombre total des bases = (nombre de taux de valeur manquante) X (nombre de paires de bases apprentissage-test) X (nombre de méthodes de substitution)

Le nombre total des bases = 4*12*4 = 192 sous bases de données d'apprentissage.

3.3.2.1 Résultats

Les résultats obtenus avec les différents taux de données manquantes et pour la même sous base de données, mais avec différentes techniques de substitution sont présentés dans le tableau 3.16 et la figure 3.8 pour la précision et le tableau 3.17 et la figure 3.9 pour la mesure de séparation.

Les résultats de la précision (séparation), correspondant à une même technique de substitution, sont moyennés sur plusieurs tests suivant le même protocole de test décrits dans la section 3.2.1.

Chaque colonne correspond à une technique de substitution (CD, MCI, NNI, MMI), parmi les résultats de la précision (séparation) dans une même colonne, plus la précision (séparation) d'une méthode de substitution est grand, meilleure est la méthode.

Test	Acc Sans valeurs manquantes	Acc avec la méthode CD				Acc avec la méthode MCI				Acc avec la méthode NNI				Acc avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	93.75%	100.00%	100.00%	100.00%	62.50%	93.75%	93.75%	93.75%	100.00%	100.00%	100.00%	100.00%	100.00%	93.75%	93.75%	93.75%	100.00%
T2	56.25%	56.25%	56.25%	56.25%	50.00%	56.25%	56.25%	56.25%	62.50%	56.25%	56.25%	56.25%	62.50%	56.25%	56.25%	56.25%	62.50%
T3	93.75%	93.75%	93.75%	93.75%	100.00%	93.75%	93.75%	93.75%	93.75%	56.25%	56.25%	56.25%	62.50%	93.75%	93.75%	93.75%	93.75%
T4	93.75%	93.75%	93.75%	93.75%	100.00%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	100.00%	93.75%	93.75%	93.75%	93.75%
T5	93.75%	93.75%	93.75%	93.75%	100.00%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%
T6	93.75%	93.75%	93.75%	93.75%	100.00%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%	93.75%
T7	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T8	93.75%	93.75%	93.75%	50.00%	50.00%	93.75%	93.75%	100.00%	81.25%	93.75%	93.75%	100.00%	50.00%	93.75%	93.75%	100.00%	81.25%
T9	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T11	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T12	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Moyenne	93.23%	93.75%	93.75%	90.10%	88.54%	93.23%	93.23%	93.75%	93.23%	90.63%	90.63%	91.15%	88.54%	93.23%	93.23%	93.75%	93.23%

Tableau 3.16- Résultats de Acc pour les méthodes CD – MCI – NNI et MMI pour base de données analyse vibratoire des roulements.

Test	Acc Sans valeurs manquantes	Séparation avec la méthode CD				Séparation avec la méthode MCI				Séparation avec la méthode NNI				Séparation avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	0.67	0.69	0.69	0.62	0.33	1.02	1.02	1.02	0.68	0.63	0.63	0.94	0.66	1.01	1.01	1.02	0.66
T2	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
T3	1.33	1.33	1.33	1.00	1.00	1.33	1.33	1.33	1.33	1.00	1.00	1.00	1.00	1.33	1.33	1.33	1.33
T4	1.33	1.33	1.33	1.00	1.00	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.00	1.33	1.33	1.33	1.33
T5	1.33	1.33	1.33	1.00	1.00	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33
T6	1.33	1.33	1.33	1.00	1.00	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33
T7	1.00	0.67	0.67	1.67	1.03	1.00	0.66	0.66	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02
T8	1.34	1.34	1.34	0.00	0.00	1.34	1.34	1.34	1.00	1.34	1.34	0.92	0.00	1.34	1.34	1.34	1.00
T9	1.67	1.67	1.67	1.67	1.33	1.67	1.67	1.67	1.67	1.67	1.67	1.88	1.67	1.67	1.67	1.67	1.67
T10	1.67	1.67	1.67	1.67	1.33	1.67	1.67	1.67	1.34	1.67	1.67	1.67	1.34	1.67	1.67	1.67	1.67
T11	1.67	1.67	1.67	1.67	1.33	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67
T12	1.67	1.67	1.67	1.67	1.33	1.67	1.67	1.67	1.67	1.67	1.67	1.55	1.34	1.67	1.67	1.67	1.67
Moyenne	1.22	1.19	1.19	1.00	0.74	1.26	1.22	1.22	1.15	1.18	1.18	1.19	1.00	1.26	1.26	1.26	1.19

Tableau 3.17- Résultats de la marge de séparation pour les méthodes CD – MCI – NNI et MMI pour la base de données analyse vibratoire des roulements.

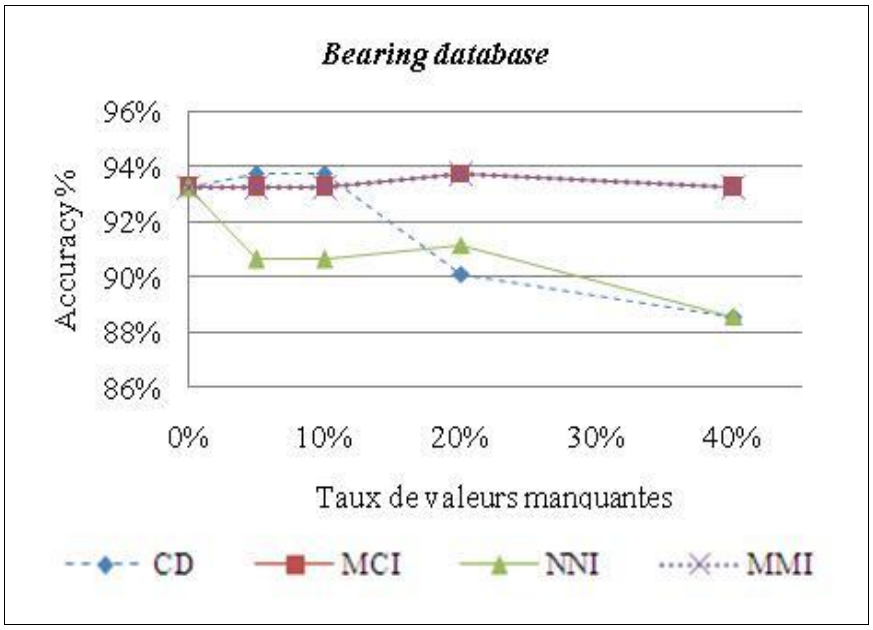


Figure 3.8- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (La précision).

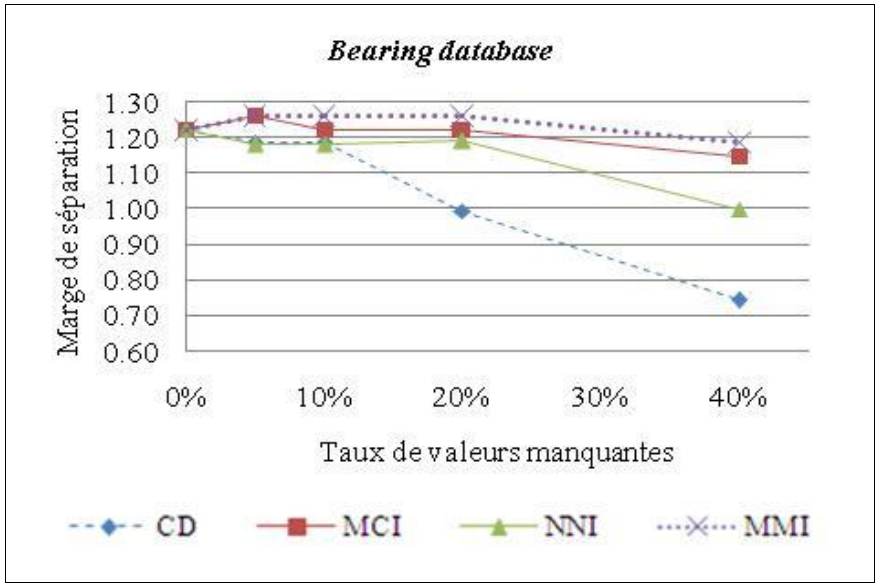


Figure 3.9- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (Marge de séparation)

3.3.2.2 Test statistique

Une série de tests de Wilcoxon[5] pour paires appariées, en anglais connu sous le nom de « Wilcoxon Matched-Pair Signed Ranks », ont été effectués pour déterminer si les différences des résultats pour la précision et la marge de séparation en raison de la modification du taux de valeurs manquantes, étaient appréciables sur le plan statistique. Le test de Wilcoxon permet d'analyser des paires de données et la mesure initiale et finale d'une seule base de données.

L'hypothèse nulle : il n'y pas de différence entre le résultat de la précision de la base de données sans traitement des valeurs manquantes et le résultat de la même base de données avec traitement des valeurs manquantes.

Le résultat des tests statistiques de wilcoxon pour les différents taux de valeurs manquantes, réalisés à partir du tableau 3.16, sont représentés sur le tableau 3.18

Tableau 3.18- Résumé des tests de Wilcoxon pour la précision (la base de données analyse vibratoire des roulements)

Couple des bases de test	T^+	T^-	N	Décision pour p = 0.05	Accepter H0 pour p ≤
Originale / 5% CD	0	1	1	Accepter H0	1
Originale / 10% CD	0	1	1	Accepter H0	1
Originale / 20% CD	2	1	2	Accepter H0	1
Originale / 40% CD	16	12	7	Accepter H0	0.8125
Originale / 5% MCI	0	0	0	Accepter H0	1
Originale / 10% MCI	0	0	0	Accepter H0	1
Originale / 20% MCI	0	1	1	Accepter H0	1
Originale / 40% MCI	3	3	3	Accepter H0	1
Originale / 5% NNI	2	1	2	Accepter H0	1
Originale / 10% NNI	2	1	2	Accepter H0	1
Originale / 20% NNI	3	3	3	Accepter H0	1
Originale / 40% NNI	9	6	5	Accepter H0	0.8125
Originale / 5% MMI	0	0	0	Accepter H0	1
Originale / 10% MMI	0	0	0	Accepter H0	1
Originale / 20% MMI	0	1	1	Accepter H0	1
Originale / 40% MMI	3	3	3	Accepter H0	1

[5] Pour de plus amples explications concernant les tests statistiques utilisés, voir annexe n°1

Le tableau 3.18 nous montre que pour la méthode CD et la méthode NNI, le seuil *p* pour accepter $H_{0,}$ commence à diminuer à partir d'un taux de valeur manquante de 40%, par contre pour les méthodes MMI et MCI, le seuil *p* pour accepter H_0 reste stable à 1, cela confirme la supériorité de la méthode MMI sur la méthode CD et la méthode NNI, résultats qui restent à confirmer avec des taux de valeurs manquantes supérieurs à 40 %. Ces premiers résultats montrent le désavantage évident de la méthode CD, elle réduit significativement la taille de la base de données (le nombre des observations), par conséquent, elle élimine de l'information importante pour définir un modèle de classification pour cbm-LAD.

Le résultat des tests statistiques de wilcoxon sur la marge de séparation, pour les différents taux de valeurs manquantes, réalisés à partir du tableau 3.17, sont présentés au tableau 3.19.

L'hypothèse nulle : il n'y a pas de différence entre le résultat de la marge de séparation de la base de données sans traitement des valeurs manquantes et le résultat de la même base de données avec traitement de valeurs manquantes.

Tableau 3.19- Résumé des tests de Wilcoxon pour la marge de séparation (la base de données analyse vibratoire des roulements)

Couple des bases de test	T^+	T^-	N	**Décision pour p = 0.05**	**Accepter H0 pour $p \leq$**
Originale / 5% CD	2	1	2	Accepter H0	1
Originale / 10% CD	2	1	2	Accepter H0	1
Originale / 20% CD	22	6	7	Accepter H0	0.2188
Originale / 40% CD	77	1	12	Rejeter H0	0.0009766
Originale / 5% MCI	0	1	1	Accepter H0	1
Originale / 10% MCI	1	2	2	Accepter H0	1
Originale / 20% MCI	1	2	2	Accepter H0	1
Originale / 40% MCI	9	1	4	Accepter H0	0.25
Originale / 5% NNI	3	0	2	Accepter H0	0.5
Originale / 10% NNI	3	0	2	Accepter H0	0.5
Originale / 20% NNI	10	5	5	Accepter H0	0.625
Originale / 40% NNI	21	0	6	Rejeter H0	0.03125
Originale / 5% MMI	0	1	1	Accepter H0	1
Originale / 10% MMI	0	1	1	Accepter H0	1
Originale / 20% MMI	0	1	1	Accepter H0	1
Originale / 40% MMI	4	2	3	Accepter H0	0.75

Les résultats du tableau 3.19 nous montrent qu'il n'y a pas de différence entre la marge de séparation pour la base originale et la marge de séparation obtenue par chaque méthode de traitement des valeurs manquantes est pour différents taux de valeurs manquantes pour le seuil de signification fixé à $p<0.05$.

Ceci n'exclut pas que l'on puisse aboutir à une conclusion différente si on teste les grands taux de valeurs manquantes.

Autrement dit, une différence observée entre les résultats des traitements et la base originale pour un taux des valeurs manquantes de 40 %, pour les méthodes CD et NNI, ce qui confirme les résultats statistiques sur la précision pour les deux méthodes en question (rejeter H0 pour un taux de valeur manquante de 40 %).

On refait le même test de wilcoxon pour comparer les méthodes de traitement deux à deux, le tableau 3.21 représente les résultats de comparaison de la précision à partir du tableau 3.20, et le tableau 3.23 représente les résultats de la marge de séparation à partir du tableau 3.22.

Tableau 3.20- Résultats de la précision par méthode et par taux de valeurs manquantes

Taux de VM	**La précision**			
	CD	**MCI**	**NNI**	**MMI**
0%	93.23%	93.23%	93.23%	93.23%
5%	93.75%	93.23%	90.63%	93.23%
10%	93.75%	93.23%	90.63%	93.23%
20%	90.10%	93.75%	91.15%	93.75%
40%	88.54%	93.23%	88.54%	93.23%

Tableau 3.21- Résultats du test des rangs appariés de Wilcoxon pour la précision

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
CD / MCI	3	7	4	Accepter H0	0.625
CD / NNI	5	1	3	Accepter H0	0.5
CD / MMI	3	7	4	Accepter H0	0.625
MCI / NNI	10	0	4	Accepter H0	0.125
MCI / MMI	0	0	0	Accepter H0	1
NNI / MMI	0	10	4	Accepter H0	0.125

Tableau 3.22- Résultats de marge de séparation par méthode et par taux de valeurs manquantes

Taux de VM	Marge de séparation			
	CD	MCI	NNI	MMI
0%	1.22	1.22	1.22	1.22
5%	1.19	1.26	1.18	1.26
10%	1.19	1.22	1.18	1.26
20%	1.00	1.22	1.19	1.26
40%	0.74	1.15	1.00	1.19

Tableau 3.23- Résultats du test des rangs appariés de Wilcoxon pour la marge de séparation

Couple des bases de test	T^+	T^-	N	Décision pour $P = 0.05$	Accepter H0 pour $P \leq$
CD / MCI	0	10	4	Accepter H0	0.125
CD / NNI	3	7	4	Accepter H0	0.625
CD / MMI	0	10	4	Accepter H0	0.125
MCI / NNI	10	0	4	Accepter H0	0.125
MCI / MMI	0	6	3	Accepter H0	0.25
NNI / MMI	0	10	4	Accepter H0	0.125

3.3.2.3 Discussion

La comparaison des différents graphiques de la Figure 3.8, correspondants aux résultats obtenus avec la base de données analyse vibratoire des roulements pour un même critère d'évaluation, permet de mettre en évidence l'influence de la méthode de substitution ainsi que le taux des valeurs manquantes.

L'augmentation du taux des valeurs manquantes de 5 % à 40 % a provoqué une diminution de la précision de 4.64 % pour la méthode CD, de 0% pour la méthode MCI, de 4.64 % pour la méthode NNI et de 0 % pour la méthode MMI accompagnée d'une diminution de la marge de séparation de 42.62 % pour la méthode CD, de 5.73 % pour la méthode MCI, de 18.03 % pour la méthode NNI et de 2.45 % pour la méthode MMI par rapport à la précision et la marge de séparation initiale (de la base de données sans valeurs manquantes).

Pour les taux des valeurs manquantes inférieurs à 40 %, la méthode MMI a provoqué une amélioration de la précision de 0.52 % et de la marge de séparation de 3.27 %, par rapport aux autres méthodes.

Les meilleurs résultats sont obtenus par les méthodes MCI et MMI (voir figure 3.8 et 3.9), même si l'écart entre les moyennes n'est pas significatif suivant les résultats du test de Wilcoxon, pour un seuil de signification de 0.05.

Pour la marge de séparation la méthode MMI se démarque bien par rapport aux autres méthodes.

3.3.3 Base de données IRIS

L'ensemble **IRIS**, contient 100 observations. Chaque observation est classifiée comme Iris-setosa ou Iris-versicolor. Dans notre cas, l'ensemble des observations positives est l'ensemble des observations de la classe Iris-setosa et l'ensemble d'observations négatives est l'ensemble des observations de la classe Iris-versicolor. Chaque observation est décrite à l'aide de 4 attributs (indicateurs).

L'ensemble des données appartenant à la classe positive dans la base des données de l'apprentissage et de test sont 30 et 20 respectivement, et l'ensemble des données appartenant à la classe négative dans la base des données de l'apprentissage et de test sont 30 et 20 respectivement, donc le total des observations serait 100 observations pour les deux classes. Suivant le principe de la validation croisée exposé au paragraphe 3.2.1, on a construit les différentes paires de bases de données apprentissage-test, les résultats sont présentés dans le tableau 3.24.

Tableau 3.24- Les bases d'apprentissage-test pour IRIS Data set

Test	Les observations de la base d'apprentissage Total = 60	Les observations de la base de test Total = 40
1	P : 21----> 50 ; N : 21 ---->50	P :1 ----> 20 ; N :1 ----> 20
2	P : 21----> 50 ; N : 41 ---->20	P :1 ----> 20 ; N : 21 ---->40
3	P : 21----> 50 ; N : 11 ---->40	P :1 ----> 20 ; N : 41 ---->10
4	P : 41----> 20 ; N : 21 ---->50	P :21 ----> 40 ; N :1 ----> 20
5	P : 41----> 20 ; N : 41 ---->20	P :21 ----> 40 ; N : 21 ---->40
6	P : 41----> 20 ; N : 11 ---->40	P :21 ----> 40 ; N : 41 ---->10
7	P : 11----> 40 ; N : 21 ---->50	P : 41 ----> 10 ; N :1 ----> 20
8	P : 11----> 40 ; N : 41 ---->20	P : 41 ----> 10 ; N : 21 ---->40
9	P : 11----> 40 ; N : 11 ---->40	P : 41 ----> 10 ; N : 41 ---->10

Pour la base de données fleurs IRIS, nous avons 9 paires de bases apprentissage-test, on a testé les 4 taux de valeurs manquantes suivants : 5 %, 10 %, 20 %, 40 %, créant ainsi 144 sous bases de données d'apprentissage, pour le test des 4 méthodes de substitution.

Le nombre total des bases d'apprentissage est calculé de la façon suivante :

Le nombre total des bases = (nombre de taux de valeur manquante) X (nombre de paires de bases apprentissage-test) X (nombre de méthodes de substitution)

Le nombre total des bases = 4*9*4 = 144 sous bases de données d'apprentissage.

3.3.3.1 Résultats

Les résultats obtenus avec un même taux de données manquantes et pour une même sous base de données, mais avec différentes techniques de substitution sont présentés dans le tableau 3.25 pour la précision et le tableau 3.26 pour la mesure de séparation.

Les résultats de la précision, correspondant à une même technique de substitution, sont moyennés sur plusieurs tests suivant le même protocole de test décrit dans la section 3.2.1.

Chaque colonne correspond à une technique de substitution (CD, MCI, NNI, MMI), parmi les résultats de la précision (séparation) dans une même colonne, plus la précision (séparation) d'une méthode de substitution est grand, meilleure est la méthode.

Test	Acc Sans valeurs manquantes	Acc avec la méthode CD				Acc avec la méthode MCI				Acc avec la méthode NNI				Acc avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	100.00%	100.00%	100.00%	90.00%	82.50%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	100.00%	97.50%	100.00%	100.00%	100.00%	100.00%
T2	100.00%	100.00%	100.00%	90.00%	90.00%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T3	100.00%	100.00%	100.00%	90.00%	90.00%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	100.00%	100.00%
T4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T6	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	97.50%	100.00%	100.00%	100.00%	100.00%
T7	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	95.00%	100.00%	100.00%	100.00%	100.00%
T8	100.00%	100.00%	100.00%	100.00%	95.00%	100.00%	100.00%	100.00%	95.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
T9	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Moyenne	100.00%	100.00%	100.00%	96.67%	95.28%	100.00%	100.00%	100.00%	96.11%	100.00%	100.00%	100.00%	97.78%	100.00%	100.00%	100.00%	100.00%

Tableau 3.25- Résultats de Acc pour les méthodes CD – MCI – NNI et MMI pour la base de données IRIS

Test	Acc Sans valeurs manquantes	Séparation avec la méthode CD				Séparation avec la méthode MCI				Séparation avec la méthode NNI				Séparation avec la méthode MMI			
		Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes				Taux des valeurs manquantes			
		5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%	5%	10%	20%	40%
T1	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	0.67	1.00	1.00	1.00	0.66	1.00	1.00	1.00	1.00
T2	2.00	2.00	2.00	1.00	1.00	2.00	2.00	2.00	0.67	2.00	2.00	2.00	1.33	2.00	1.67	1.67	1.67
T3	1.34	1.34	1.34	0.67	1.34	1.34	1.34	1.34	0.67	1.34	1.34	1.67	0.67	1.34	1.34	1.34	1.00
T4	1.67	1.67	1.33	1.33	1.33	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.33	1.67	1.67	1.67	1.67
T5	2.00	2.00	2.00	2.00	1.67	2.00	2.00	2.00	1.67	2.00	2.00	2.00	1.67	2.00	2.00	2.00	2.00
T6	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67	1.67
T7	1.34	1.34	1.34	1.34	1.67	1.34	1.34	1.34	1.34	1.34	1.34	1.34	0.33	1.67	1.67	1.34	1.34
T8	1.33	1.33	1.33	1.33	0.67	1.33	1.33	1.33	0.67	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33
T9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Moyenne	1.48	1.48	1.45	1.26	1.22	1.48	1.48	1.48	1.11	1.48	1.48	1.52	1.11	1.52	1.48	1.45	1.41

Tableau 3.26- Résultats de la marge de séparation pour les méthodes CD – MCI – NNI et MMI pour la base de données IRIS

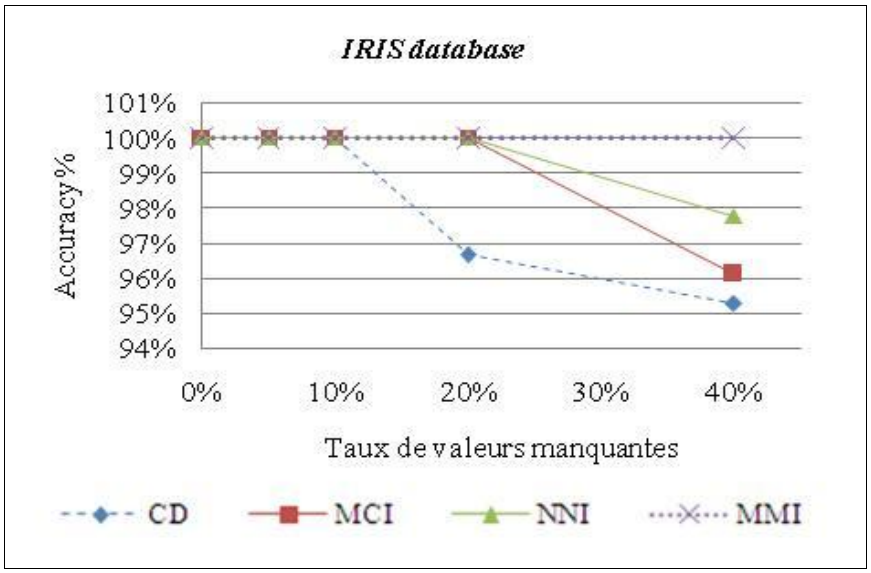


Figure 3.10- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (La précision).

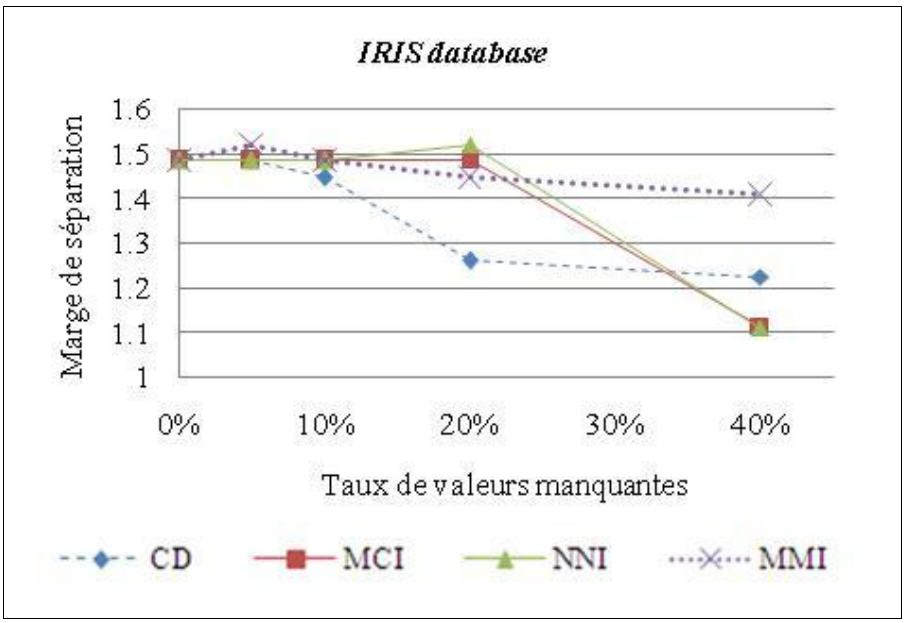


Figure 3.11- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes (Marge de séparation)

3.3.3.2 Test statistique

Une série de tests de Wilcoxon[6] pour paires appariées, en anglais connu sous le nom de « Wilcoxon Matched-Pair Signed Ranks », ont été effectués pour déterminer si les différences des résultats pour la précision et la marge de séparation en raison de la modification du taux de valeurs manquantes, étaient appréciables sur le plan statistique. Le test de Wilcoxon permet d'analyser des paires de données et la mesure initiale et finale d'une seule base de données.

L'hypothèse nulle : il n'y a pas de différence entre le résultat de la précision de la base de données sans traitement des valeurs manquantes et le résultat de la même base de données avec traitement de valeurs manquantes.

Le résultat des tests statistiques de wilcoxon pour les différents taux de valeurs manquantes, réalisés à partir du tableau 3.25, sont présentés au tableau 3.27

Tableau 3.27- Résumé des tests de Wilcoxon pour la précision (la base de données IRIS)

Couple des bases de test	T^+	T^-	N	**Décision pour $p = 0.05$**	**Accepter H0 pour $p \leq$**
Originale / 5% CD	0	0	0	Accepter H0	1
Originale / 10% CD	0	0	0	Accepter H0	1
Originale / 20% CD	6	0	3	Accepter H0	0.25
Originale / 40% CD	10	0	4	Accepter H0	0.125
Originale / 5% MCI	0	0	0	Accepter H0	1
Originale / 10% MCI	0	0	0	Accepter H0	1
Originale / 20% MCI	0	0	0	Accepter H0	1
Originale / 40% MCI	10	0	4	Accepter H0	0.125
Originale / 5% NNI	0	0	0	Accepter H0	1
Originale / 10% NNI	0	0	0	Accepter H0	1
Originale / 20% NNI	0	0	0	Accepter H0	1
Originale / 40% NNI	10	0	4	Accepter H0	0.125
Originale / 5% MMI	0	0	0	Accepter H0	1
Originale / 10% MMI	0	0	0	Accepter H0	1
Originale / 20% MMI	0	0	0	Accepter H0	1
Originale / 40% MMI	0	0	0	Accepter H0	1

[6] Pour de plus amples explications concernant les tests statistiques utilisés, voir annexe n°1

Le tableau 3.27 nous montre que pour les méthodes CD,NNI et MCI, le seuil *p* pour accepter $H_{0,}$ commence à diminuer à partir d'un taux de valeur manquante de 40%, par contre pour la méthode MMI , le seuil *p* pour Accepter H_0 reste stable à 1, cela confirme la supériorité de la méthode MMI , résultats qui restent à confirmer avec des taux de valeurs manquantes supérieurs à 40%. Ces premiers résultats montrent le désavantage évident de la méthode CD puisque le seuil *p* pour accepter H_0, commence à diminuer à partir d'un taux de valeur manquante de 20%.

Le résultat des tests statistiques de wilcoxon sur la marge de séparation, pour les différents taux de valeurs manquantes, réalisés à partir du tableau 3.26, sont présentés au tableau 3.28.

L'hypothèse nulle : il n'y pas de différence entre le résultat de la marge de séparation de la base de données sans traitement des valeurs manquantes et le résultat de la même base de données avec traitement de valeurs manquantes.

Tableau 3.28- Résumé des tests de Wilcoxon pour la marge de séparation

(La base de données IRIS)

Couple des bases de test	T^+	T^-	N	**Décision pour p = 0.05**	**Accepter H0 pour $p \leq$**
Originale / 5% CD	0	0	0	Accepter H0	1
Originale / 10% CD	1	0	1	Accepter H0	1
Originale / 20% CD	6	0	3	Accepter H0	0.25
Originale / 40% CD	20	1	6	Accepter H0	0.0625
Originale / 5% MCI	0	0	0	Accepter H0	1
Originale / 10% MCI	0	0	0	Accepter H0	1
Originale / 20% MCI	0	0	0	Accepter H0	1
Originale / 40% MCI	15	0	5	Accepter H0	0.0625
Originale / 5% NNI	0	0	0	Accepter H0	1
Originale / 10% NNI	0	0	0	Accepter H0	1
Originale / 20% NNI	0	1	1	Accepter H0	1
Originale / 40% NNI	21	0	6	Accepter H0	0.03125
Originale / 5% MMI	0	1	1	Accepter H0	1
Originale / 10% MMI	2	1	2	Accepter H0	1
Originale / 20% MMI	1	0	1	Accepter H0	1
Originale / 40% MMI	3	0	2	Accepter H0	0.5

Les résultats du tableau 3.28 nous montrent qu'il n'y a pas de différence entre la marge de séparation pour la base originale et la marge de séparation obtenue par chaque méthode de traitement des valeurs manquantes est pour différents taux de valeurs manquantes pour le seuil de signification fixé à $p < 0.05$.

Ceci n'exclut pas que l'on puisse aboutir à une conclusion différente si on teste les grands taux de valeurs manquantes.

Autrement dit, une différence observée pour le seuil *p* pour accepter $H_{0,}$ il commence à diminuer à partir d'un taux de valeur manquante de 40 %, et particulièrement pour la méthode CD, il commence à diminuer à partir de 20 %.

On refait le même test de wilcoxon pour comparer les méthodes de traitement deux à deux, le tableau 3.30 représente les résultats de comparaison de la précision à partir du tableau 3.29, et le tableau 3.32 représente les résultats de la marge de séparation à partir du tableau 3.31.

Tableau 3.29- Résultats de la précision par méthode et par taux de valeurs manquantes

Taux de VM	**La précision**			
	CD	**MCI**	**NNI**	**MMI**
0%	100.00%	100.00%	100.00%	100.00%
5%	100.00%	100.00%	100.00%	100.00%
10%	100.00%	100.00%	100.00%	100.00%
20%	96.66%	100.00%	100.00%	100.00%
40%	95.27%	96.11%	97.78%	100.00%

Tableau 3.30- Résultats du test des rangs appariés de Wilcoxon pour la précision

Couple des bases de test	T^+	T^-	N	**Décision pour $P = 0.05$**	**Accepter H0 pour $P \leq$**
CD / MCI	0	3	2	Accepter H0	0.5
CD / NNI	0	3	2	Accepter H0	0.5
CD / MMI	0	3	2	Accepter H0	0.5
MCI / NNI	0	1	1	Accepter H0	1
MCI / MMI	0	1	1	Accepter H0	1
NNI / MMI	0	1	1	Accepter H0	1

Tableau 3.31- Résultats de marge de séparation par méthode et par taux de valeurs manquantes

Taux de VM	Marge de séparation			
	CD	MCI	NNI	MMI
0%	1.48	1.48	1.48	1.48
5%	1.48	1.48	1.48	1.52
10%	1.45	1.48	1.48	1.48
20%	1.26	1.48	1.52	1.45
40%	1.22	1.11	1.11	1.41

Tableau 3.32- Résultats du test des rangs appariés de Wilcoxon pour la marge de séparation

Couple des bases de test	T^+	T^-	N	Décision pour $P = 0.05$	Accepter H0 pour $P \leq$
CD / MCI	2	4	3	Accepter H0	0.75
CD / NNI	2	4	3	Accepter H0	0.75
CD / MMI	0	10	4	Accepter H0	0.125
MCI / NNI	0	1	1	Accepter H0	1
MCI / MMI	1	5	3	Accepter H0	0.5
NNI / MMI	2	4	3	Accepter H0	0.75

3.3.3.3 Discussion

La comparaison des différents graphiques de la Figure 3.10, correspondants aux résultats obtenus avec la base de données IRIS pour un même critère d'évaluation, permet de mettre en évidence l'influence de la méthode de substitution ainsi que le taux des valeurs manquantes.

L'augmentation du taux des valeurs manquantes de 5% à 40% a provoqué une diminution de la précision de 4.72 % pour la méthode CD, de 3.89% pour la méthode MCI, de 2.22% pour la méthode NNI et de 0% pour la méthode MMI accompagné d'une diminution de la marge de séparation de 17.56 % pour la méthode CD, de 25% pour la méthode MCI, de 25% pour la méthode NNI et de 4.72% pour la méthode MMI par rapport à la précision et la marge de séparation initiale (de la base de données sans valeurs manquantes).

Les meilleurs résultats sont obtenus par les méthodes MMI (voir figure 3.8 et 3.9), même si l'écart entre les moyennes n'est pas significatif suivant les résultats du test de Wilcoxon, pour un seuil de signification de 0.05.

Pour la marge de séparation la méthode MMI se démarque bien par rapport aux autres méthodes.

3.4 Robustesse de la méthode de substitution proposée

Les résultats des trois ensembles de données, sont présentés au tableau 3.33, on observe que la qualité de résultats (La précision / Marge de séparation), est fort dépendante du taux des valeurs manquantes dans l'ensemble de données d'apprentissage.

Tableau 3.33- Les performances des techniques de substitution en fonction du taux des valeurs manquantes pour les trois bases de données.

Taux VM	Data-base	La précision				Séparation			
		Méthode de substitution des VM				Méthode de substitution des VM			
		CD	NNI	MCI	MMI	CD	NNI	MCI	MMI
5%	IRIS	100.00%	100.00%	100.00%	100.00%	1.48	1.48	1.48	1.52
	BEARING	93.75%	90.63%	93.23%	93.23%	1.19	1.18	1.26	1.26
	DAG	86.11%	97.22%	97.22%	97.22%	0.93	1.15	1.15	1.12
10%	IRIS	100.00%	100.00%	100.00%	100.00%	1.45	1.48	1.48	148
	BEARING	93.75%	90.63%	93.23%	93.23%	1.19	1.18	1.22	1.26
	DAG	72.22%	93.06%	95.83%	94.44%	1	1.27	1.16	1.34
20%	IRIS	96.66%	100.00%	100.00%	100.00%	1.26	1.52	1.48	1.45
	BEARING	90.10%	91.15%	93.75%	93.75%	1	1.19	1.22	1.26
	DAG	65.28%	81.94%	90.28%	84.72%	0.85	1.25	1	1.37
40%	IRIS	95.27%	97.78%	96.11%	100.00%	1.22	1.11	1.11	1.41
	BEARING	88.54%	88.54%	93.23%	93.23%	0.74	1	1.15	1.19
	DAG	65.28%	73.61%	76.39%	75.00%	0.85	0.89	0.88	0.93

La Figure 3.12 regroupe les graphiques donnant la moyenne des performances de l'ensemble des techniques en fonction du taux de valeurs manquantes pour chacune des

bases de données de l'étude. Une tendance émerge de ces graphiques : la moyenne des performances a tendance à décroître lorsque le taux de valeurs manquantes augmente. Ceci est bien en accord avec l'idée intuitive selon laquelle la dégradation de la qualité des données s'accompagne d'une dégradation de la qualité des classificateurs construits à partir de ces données. Notons cependant que dans certains cas, les performances sont stables, ce qui est le signe d'une certaine robustesse.

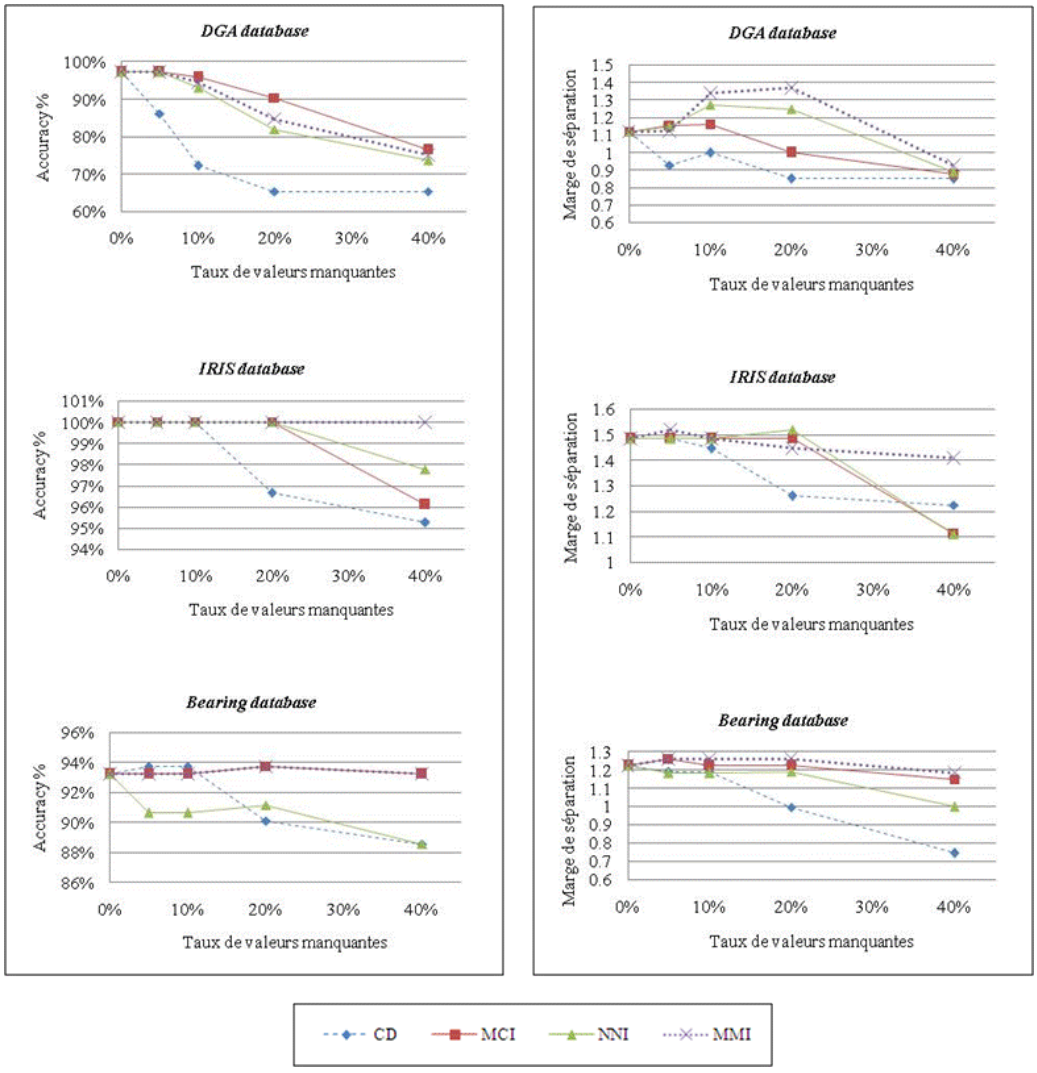


Figure 3.12- Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes pour chacune des bases de données étudiées (La précision à gauche et Marge de séparation à droite)

La comparaison des différents graphiques d'une même colonne, correspondants aux résultats obtenus avec les trois bases de données pour un même critère d'évaluation, permet de mettre en évidence l'influence de la méthode de substitution ainsi que le taux des valeurs manquantes.

En regardant de manière plus détaillée, sans utilisation d'un test statistique, lors de l'analyse des résultats avec les différents taux de données manquantes, il s'avère que quand le taux de données manquantes augmente, le taux de bonnes classifications obtenu décroît, résultats supportés par la théorie statistique de l'apprentissage [87]. Il semble évident que le résultat dépend ainsi de la qualité des données d'entrée.

Avec un taux faible de données manquantes, les méthodes ne se différencient pas significativement. Par exemple, avec 5% de valeurs manquantes, les taux de bonnes classifications obtenus sur une même base de données sont très similaires. À 40% de données manquantes, il y a un écart significatif entre les résultats obtenus par les différentes méthodes.

La moyenne des performances sur une même base de données n'évoluera pas de la même façon en fonction du taux de valeurs manquantes. L'exemple le plus marquant est celui de la méthode CD. Avec les trois bases de données, quel que soit le critère d'évaluation, la moyenne des performances de cette méthode décroît d'une façon remarquable, ce qui concorde avec plusieurs résultats de la littérature (Little, 1988) [18], alors que la méthode MMI semble stable pour la majorité des bases.

Rappelons que CD est une méthode basée sur l'élimination des observations avec valeur manquante. Il est donc vraisemblable que si le nombre d'observations est insuffisant dû à la proportion de valeurs manquantes est grande, ses performances soient relativement faibles.

Pour les trois ensembles de données, nous avons validé les résultats obtenus par un test statistique non paramétrique de Friedman qui ne fait aucune hypothèse sur la forme des distributions sous-jacentes. Comme de nombreux tests non paramétriques, il travaillera non pas sur les valeurs numériques des résultats, mais sur leurs rangs, une fois ces résultats convenablement réunis dans un tableau approprié et nous avons utilisé les tests post hoc

évoqués ci-dessus pour évaluer statistiquement les différences observées quand le test de Friedman conclut qu'il existe des différences significatives.

Pour réaliser le test de Friedman, on doit suivre les étapes suivantes :

1. Calculer les performances des techniques de substitution en fonction du taux des valeurs manquantes pour les trois bases de données (tableau 3.33).
2. Classer les résultats de la précision et da marge de séparation en rang et par ordre (tableau 3.34).
3. Calculer la somme des rangs. (tableau 3.35).
4. Calculer la statistique de Friedman (tableau 3.36).

Tableau 3.34- Le classement des résultats de la précision et da marge de séparation en rang et par ordre.

Taux VM	**Data-base**	**La précision**				**Séparation**			
		Méthode de substitution des VM				**Méthode de substitution des VM**			
		CD	**NNI**	**MCI**	**MMI**	**CD**	**NNI**	**MCI**	**MMI**
5%	**IRIS**	2.5	2.5	2.5	2.5	2	2	2	4
	BEARING	4	1	2.5	2.5	2	1	3.5	3.5
	DAG	1	3	3	3	1	3.5	3.5	2
10%	**IRIS**	2.5	2.5	2.5	2.5	1	3	3	3
	BEARING	4	1	2.5	2.5	2	1	3	4
	DAG	1	2	4	3	1	3	2	4
20%	**IRIS**	1	3	3	3	1	4	3	2
	BEARING	1	2	3.5	3.5	1	2	3	4
	DAG	1	2	4	3	1	3	2	4
40%	**IRIS**	1	3	2	4	3	1.5	1.5	4
	BEARING	1.5	1.5	3.5	3.5	1	2	3	4
	DAG	1	2	4	3	1	3	2	4

En notant R_X, la somme des rangs de la colonne i (par méthode x) pour chaque taux de valeurs manquantes.

Tableau 3.35- Somme des rangs.

Somme des rangs	La précision				Séparation			
	5%	***10%***	***20%***	***40%***	***5%***	***10%***	***20%***	***40%***
$\mathbf{R_{CD}}$	7.5	7.5	3	3.5	5	4	3	5
$\mathbf{R_{NNI}}$	6.5	5.5	7	6.5	6.5	7	9	6.5
$\mathbf{R_{MCI}}$	8	9	10.5	9.5	9	8	8	6.5
$\mathbf{R_{MMI}}$	8	8	9.5	10.5	9.5	11	10	12

Le calcul de la statistique de Friedman sera donc :

Tableau 3.36- Résultats du test de Friedman

Critères	La précision				**Séparation**			
	5%	***10%***	***20%***	***40%***	***5%***	***10%***	***20%***	***40%***
a	3	3	3	3	3	3	3	3
b	4	4	4	4	4	4	4	4
SRj^2	226.5	231.5	258.5	255	238.5	258.5	254	253.5
X_r^2	**0.3**	**1.3**	**6.7**	**6**	**2.7**	**5**	**5.8**	**5.7**
P = 5%	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4
Ex-æquo	6	6	2	4	4	0	0	2
Correction	0.5	0.633	0.966	0.933	0.93	1	1	0.966
X_r^2 (après correction)	**0.6**	**2.053**	**6.935**	**6.428**	**2.8928**	**5**	**5.8**	**5.90**
Décision	Accepter H_0	Accepter H_0	Accepter H_0	Accepter H_0	Accepter H_0	Accepter H_0	Accepter H_0	Accepter H_0

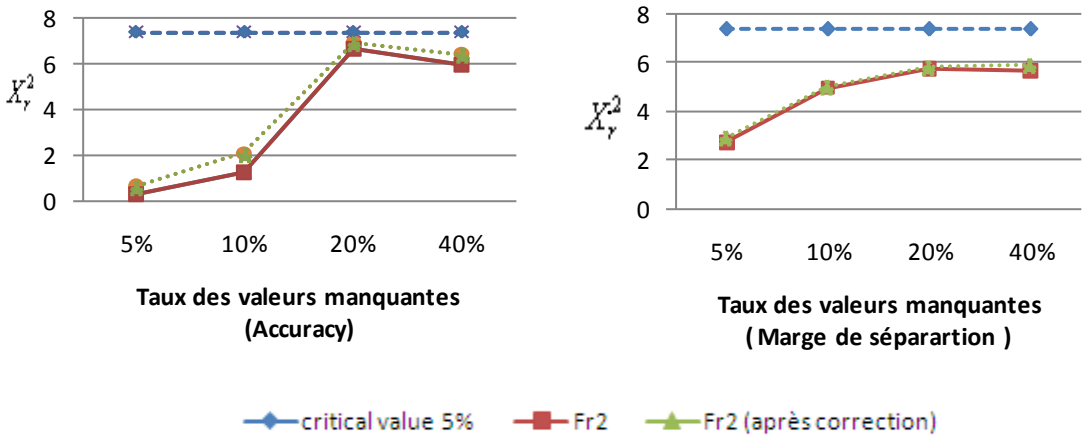


Figure 3.13- Résultats du test de Friedman

De ces résultats, il paraît que la technique que nous avons proposée obtient des performances tout à fait satisfaisantes. Elle s'avère statistiquement supérieure à la quasi-totalité des autres techniques, pour au moins un critère de performance.

Seule la méthode MCI obtient un total des rangs supérieur par rapport au total des rangs de la méthode MMI, pour la précision, bien que la différence observée ne soit pas significative. Il est à noter que le comportement de la méthode MMI est beaucoup plus intéressant avec des taux de valeurs manquantes élevés.

Concernant les résultats des tests statistiques, toutes les méthodes sont jugées équivalentes par le test de Friedman pour un taux de valeurs manquantes inférieur à 40 %.

CHAPITRE 4. CONCLUSION ET PERSPECTIVES

L'analyse logique des données est une nouvelle méthodologie pour détecter les informations structurelles sur un ensemble de données. Elle peut offrir une solution à divers problèmes de classification dans le domaine de la maintenance conditionnelle.

CBM-LAD est un outil de classification puissant pourvu que la base de données d'apprentissage ne soit pas affectée par des valeurs manquantes, il semble clair que la qualité de la classification, dans le contexte de CBM-LAD, dépend de la méthode de traitement des données manquantes en aval. Tout cela est en accord avec notre hypothèse selon laquelle la qualité des données utilisées joue un rôle non négligeable dans la prise de décision dans le domaine industriel.

Dans le cadre de ce mémoire, nous avons pu fournir une solution au problème de classification avec valeurs manquantes. La méthode de substitution développée a, en effet, conduit à d'excellents résultats avec plusieurs jeux de données.

Le comportement de notre approche de substitution des valeurs manquantes par la méthode de MMI, est beaucoup plus intéressant avec des taux de valeurs manquantes élevés, on comparaison avec les autres méthodes de substitution testées dans ce mémoire.

L'étude comparative de différentes techniques est un point crucial, pour pouvoir juger la pertinence de la méthode proposée, car elle permet de justifier empiriquement certains choix théoriques, nous avons fréquemment recours pour l'analyse quantitative des résultats dans ce mémoire, aux tests statistiques non-paramétriques.

À travers la revue de littérature, on constate que les méthodes de substitution des valeurs manquantes, sont nombreuses et qu'il n'existe pas de recettes définitives, et on doit agir de cas par cas. Cependant, nous sommes à même d'affirmer que la démarche que nous proposons conduit à des substitutions satisfaisantes tant d'un point de vue théorique qu'empirique. En outre, la méthode de substitution MMI, étant donné sa particularité de

créer des bases de données complètes à partir des valeurs déterminées par la connaissance que l'on a des données, ainsi que l'information des classes, et sa relative simplicité (ne demande pas des calculs complexes c.à.d. un temps d'exécution rapide et une faible consommation de la mémoire), fournit des avantages non négligeables dans l'analyse logique des données et ouvre ainsi une voie intéressante dans le traitement des données manquantes pour la maintenance conditionnelle.

Le protocole des tests que nous avons adopté est tout de même séduisant, car il permet de comparer les algorithmes sur une base de test commune, mais il y a deux remarques à faire par rapport à ce schéma de protocole : on ne traite toujours que les valeurs manquantes totalement aléatoires dans ce cas et seulement sur le fichier d'apprentissage, cette approche présente l'inconvénient de ne pas correspondre à un scénario réaliste. En pratique, les bases de données sur lesquelles un classificateur peut être appris contiennent des données manquantes, mais les futurs exemples qu'il faudra classer aussi.

Afin de pouvoir généraliser les résultats ainsi obtenus, il serait intéressant d'étendre l'analyse sur des taux de valeurs manquantes plus grandes ainsi que sur des bases de données plus grande.

L'autre piste à creuser dans ce cas est la gestion de données manquantes lorsque nous appliquons le modèle de classification c.à.d. lorsque les observations du fichier test elles-mêmes ne sont pas décrites complètement. L'idée mérite d'être creusée à mon avis. La gestion des données manquantes est au moins aussi importante lors de la phase de classement que lors de la phase d'apprentissage.

BIBLIOGRAPHIE

[1] Brown, M. L., and J. F. Kros, "The Impact of Missing Data on Data Mining," in John Wang, (ed.), Data Mining: Opportunities and Challenges, Hershey, PA: IRM Press, 2003, pp. 174–198.

[2] Salamanca D., Yacout S., 2007, "Condition based maintenance with logical analysis of data", 7e Congrès International de Génie industriel.

[3] Barlow, R. E. et Hunter, L. C. (1960). Optimum preventive maintenance policies. Oper. Res., 8(1) :90–100.

[4] Simmons, G. J. et Pollock, S. M. (2005). Marginally monotonic maintenance policies for a multi-state deteriorating machine with probabilistic monitoring, and silent failures. IEEE Transactions on Reliability, 54(3) :489–497.

[5] Kobbacy, K. A. et Jeon, J. (2002). Generalized non-stationary preventive maintenance model for derteriorating repairable systems. Quality and Reliability Engineering International, 18 :367–372.

[6] Mobley K 1989 Introduction to Predictive Maintenance. Van Nostrand Reinhold, New York.

[7] Alsyouf, I. (2004). Cost Effective Maintenance for Competitive Advantages. Doctoral Thesis.Sweden, Växjö: Växjö University Press.

[8] Fararooy, S. and Allan, J. (1995). Condition-Based Maintenance of Railway Signalling Equipment.In International Conference Electric Railways in a United Europe.

[9] Basseville, Michèle et Marie-Odile Cordier (1996). Surveillance et diagnostic de systèmes dynamiques : approches complémentaires du traitement de signal et de l'intelligence artificielle. Technical Report 2861. INRIA. Rennes, France.

[10] Nicolas PALLUAT Méthodologie de surveillance dynamique à l'aide des réseaux neuro-fous temporels. PhD thesis, L'UFR des Sciences et Techniques de l'Université de Franche-Comté, 2006.

[11] Elizabeth Vannan « Quality Data An Improbable Dream. A process for reviewing and improving data quality makes for reliable and usable results ».. EDUCAUSE QUARTERLY, CUMREC conferences, Centre for Education Information, 2001

[12] Paul Jermyn, Maurice Dixon and Brian J Read « Preparing Clean Views of Data for Data Mining ». 12th ERCIM Workshop on Database Research, Amsterdam, 2-3 November 1999.

[13] Brion P. et Clairin R. (1997) - Manuel de sondages : Applications aux pays en développement, INSEE et CPED, Paris.

[14] Chiang et al., 2003 L.H. Chiang, R.J. Pell and M.B. Seasholtz, Exploring process data with the use of robust outlier detection algorithms, Journal of Process Control 13 (2003) (5), pp. 437–449.

[15] Frank E. Grubbs, « Procedures for Detecting Outlying Observations in Samples ». Technometrics, Vol. 11, No. 1 (Feb., 1969), pp. 1-21

[16] Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley & Sons, Inc , pp. 11-13.

[17] Simon G.A. et Simonoff J.S. (1986) - Diagnostic plots for missing data in least squares regression, Journal of the American Statistical Association, 81, 501-509.

[18] Little R.J.A. (1988) - A test of missing completely at random for multivariate data with missing values, Journal of the American

[19] Schafer J. L. and J. W. Graham. Missing data: Our view on the state of the art. Psychological Methods, 7(2):147-177, 2002.

[20] M. Hu, S.M. Salvucci et M.P. Cohen: Evaluation of some popular imputation algorithms, in Section on Survey Research Methods, pages 309-313, 2000. American Statistical Association.

[21] Kline, R.B., 1998. Principles and Practice of Structural Equation Modelling. Guilford Press, New York.

[22] Q. Song et M. Shepperd : A new imputation method for small software project data sets. Journal of Systems and Software, 80(1):51-62, 2007.

[23] Malhotra, N.K., 1987. Analyzing marketing research data with incomplete information on the dependent variable. Journal of Marketing Research 24, 74–84.

[24] Kim, J.O., Curry, J., 1977. The treatment of missing data in multivariate analysis. Sociological Methods and Research 6, 215–241.

[25] Kaufman, C.J., 1988. The application of logical imputation to household measurement. Journal of the Market Research Society 30, 453–466.

[26] M. Magnani : Techniques for dealing with missing data in knowledge discovery tasks. Research report, University of Bologna, Computer Science Department, 2003.

[27] Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: Chapman and Hall.

[28] Roth, P.L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-560.

[29] Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40(2), 229–252.

[30] Raymond, M. R. (1986). Missing data in evaluation research.Evaluation & the Health Professions, 9(4), 395–420.

[31] Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measuremen, 47, 13–26.

[32] Allison. P. D. (2001). Missing Data (Quantitative Applications in the Social Sciences). Sage University Papers Series on Quantitative Applications in the Social Sciences,07-136. Sage, Thousand Oaks CA, .

[33] Rubin D.B. (1987) - Multiple imputation for nonresponse in survey, Wiley.

[34] Raaijmakers, Q.A.W. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data : Introducing the relative mean substitution approach. Educational and Psychological Measurement, 59, 725-748.

[35] Kromrey, J.D., Heines, C.V., 1994. Nonrandomly missing data in multiple regression: an empirical comparison of common missing-data. Educational and Psychological Measurement 54 (3), 573–593.

[36] Pigott, T.D. (2001). A review of methods for missing data. Educational research and evaluation, 7: 353-383.

[37] Chan, L.S., Dunn, O.J., 1972. The treatment of missing values in discriminant analysis. I. The sampling experiment. Journal of the American Statistical Association 67, 473–477.

[38] Chan, L.S., Gilman, J.A., Dunn, O.J., 1976. Alternative approaches to missing values in discriminant analysis. Journal of the American Statistical Association 71, 842–844.

[39] Frane, J.W., 1976. Some simple procedures for handling missing data in multivariate analysis. Psychometrika 41, 409–415.

[40] Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. Psychological Methods, 6 (4), 317-329.

[41] Chen J, Shao J (2000). Nearest neighbor imputation for survey data. Journal of Official Statistics, Vol.16, No.2, 2000. pp. 113-131

[42] Engels JM, Diehr P (2003). Imputation of missing longitudinal data: a comparison of methods - Journal of Clinical Epidemiology, Volume 56 (2003), Issue 10, Pages 968-976 Statistical Association, 83, 1198-1202

[43] Shichao Zhang. (2008). Parimputation: From imputation and null-imputation to partially imputation. IEEE Intelligent Informatics Bulletin, Vol 9(1), 2008: 32-38.

[44] Zhang, S.C, Qin, Y.S., Zhang, J.L., Zhu, X.F., Zhang, C.Q. (2008).Missing Value Imputation Based on Data Clustering. Transactions on Computational Science Journal, LNCS 4750, pp 128-138.

[45] Rubin, D. (1978) Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse., 1–23. U.S. Department of Commerce.

[46] Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91 (434), 473-489.

[47] Rubin, D.B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," Journal of the American Statistical Association, 81, 366-374.

[48] Fichman, M. et Cummings, J.N., (2003). Multiple imputation for missing data : Making the most of what you know. Organizational Research Methods, 6: 282-308.

[49] Rubin D., « The bayesian boostrap », Ann. Statistics, vol. 9, p. 130-134, 1981.

[50] Freund Y., « Boosting a week learning algorithm by majority », Information and Computation, 1995.

[51] DeSarbo, W.S., Green, P.E., Carroll, J.D., 1986. Missing data in product-concept testing. Decision Sciences 17, 163–185.

[52] Lee, S.Y., Chiu, Y.M., 1990. Analysis of multivariate polychoric correlation models with incomplete data. British Journal of Mathematical and Statistical Psychology 43, 145–154.

[53] A. Dempster, N. Laird et D. Rubin : Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1-38, 1977.

[54] Z. Ghahramani et M.I. Jordan : Supervised learning from incomplete data via an EM approach. In J.D. Cowan, G. Tesauro et J. Alspector, éditeurs : Advances in Neural Information Processing Systems 6, pages 120-127. Morgan Kaufman, 1994.

[55] Laird, N.M., 1988. Missing data in longitudinal studies. Statistics in Medicine 7, 305–315.

[56] Ruud, P.A., 1991. Extensions of estimation methods using the EM algorithm. Journal of Econometrics 49, 305–341.

[57] Y. Zou, A. An et X. Huang : Evaluation and automatic selection of methods for handling missing data. In IEEE International Conference on Granular Computing, 2005.

[58] Daniel T. Larose (2005) - Des Données à la connaissance, une introduction au data mining- pp 34-35-36 ,ISBN : 2-7117-4855-3, code-barre : 978 2 7117 4855 6 Edition Vuibert Informatique

[59] Cormack, R. M. (1971). A review of classification (with discussion). The Royal Statistical Society 3, 321–367.

[60] Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika 32, 241–254.

[61] Borko, H., and Bernick, M. D. (1963). Automatic document classification. Journal of the Association for Computing Machinery, 10:1151–1162.(p. 160)

[62] Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1(1-2), 69–90.

[63] P.L. Hammer, (1986). Partially Defined Boolean Functions and Cause-Effect Relationships. Proc. Int'l Conf. Multi-Attrubute Decision Making Via OR-Based Expert Systems.

[64] Y. Crama, P.L. Hammer, and T. Ibaraki,(1988). Cause-Effect Relationships and Partially Defined Boolean Functions. Annals of Operations Research, vol. 16, pp. 299-326.

[65] Boros. E., Hammer, P. L., Kogan, A., Mayoraz, E., & Muchnik, I. (1994). Logical analysis of data—overview. RUTCOR- Center For Operation Research. Rutgers University, RTR 1–94.

[66] Boros, E., Ibaraki, T., & Makino, K. (1996). Extensions of partially defined Boolean functions with missing data. Rutgers University. RUTCOR Research Report, RRR 06-96.

[67] Boros E., Hammer P., Ibaraki T., Kogan A., Mayoraz E., Muchnick I. (2000). An Implementation of Logical Analysis of Data. IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 2, March/ April, pp. 292-306.

[68] Taha H. (1975). Integer programming: Theory, applications and computations (pp. 326). New York: Academic Press.

[69] Alexe, G., Alexe, S., Hammer, P. L., & Kogan, A. (2002). Comprehensive vs. comprehensible classifiers in logical analysis of data. RUTCOR Research Report, Rutgers University, RRR 9-2002.

[70] R. Kohavi and F. Provost, Glossary of terms. Editorial for the special issue on application of machine learning and the knowledge of discovery process, *Machine Learning* **30** (1998), pp. 271–274.

[71] Grzymala-Busse, J. W.: On the unknown attribute values in learning from examples. Proc. of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 16–19, 1991, Lecture Notes in Artificial Intelligence, vol. 542. Springer-Verlag, Berlin Heidelberg New York (1991) 368–377.

[72] J. W. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. In Proceedings of the 2nd International Conference on Rough Sets and Current Trends in Computing, RSCTC'00, pages 378–385, 2001

[73] E. Boros, T. Ibaraki, and K. Makino, "Logical Analysis of Binary Data with Missing Bits," *Artificial Intelligence*, vol. 107, no. 2, pp. 219–263, 1999

[74] Boros E., P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik. An Implementation of Logical Analysis of Data. IEEE Transactions on Knowledge and DataEngineering, 12(2):292{306, 2000.

[75] E. Acuna and C. Rodriguez. The treatment of missing values and its effect in the classifier accuracy. In Classification, Clustering and Data Mining Applications, pages 639–648. Springer-Verlag, 2004.

[76] G. Batista and M. C. Monard. An analysis of four missing data treatment methods of supervised learning. Applied Artificial Intelligence, 6(3):309–327, 2003.

[77] J. W. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. In Proceedings of the 2nd International Conference on Rough Sets and Current Trends in Computing, RSCTC'00, pages 378–385, 2000.

[78] DR Velez, BC White, AA Motsinger, WS Bush, MD Ritchie, SM Williams and JH Moore, A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, Genet Epidemiol (2007).

[79] Whitely E, Ball J: Statistics review 6: Nonparametric methods. Critical Care 2002, 6:509-513.

[80] J. Demsar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, (7) :1–30, 2006.

[81] Bewick V, Cheek L, Ball J. Statistics review 10: further nonparametric methods. Critical Care. 2004;8(3):196 –199.

[82] Sheldon, M.R., Fillyaw, M.J. and Thompson, W.D. (1996) the use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures design. Physiother. Res. Int. **1**, 221–228

[83] Siva Sarma, D., & Kalyani, G. N. S. (2007). Application of AI techniques for non-destructive evaluation of power transformers using DGA. International Journal of Innovations in EnergySystems and Power, 2(1), 37–43.

[84] G. Goddu, B. Li, M.Y. Chow, Motor bearing fault diagnosis by a fundamental frequency amplitude based fuzzy decision system, in: Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society, August–September 1998, Aachen, Germany, pp. 1961–1965.

[85] Anderson E. The irises of the Gaspe Peninsula. Bulletin of the American Iris Society, 59 :2-5, 1935.

[86] Fisher, R. 1936. The use of multiple measurements in taxonomic problem. Ann. Eugenics, 7: 179-188.

[87] V.N. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.

[88] Ford, B.L., 1976. Missing data procedures: a comparative study. In: Statistical Reporting Serviceunknown: book, U.S. Department of Agriculture, Washington, DC.

[89] Zar, J. H., 1999: Biostatistical analysis, Prentice-Hall Inc, Englewood Cliffs, New Jersey

ANNEXE 1 : EXPLICATION DES TESTS STATISTIQUES UTILISÉS

Un test non paramétrique est un test d'hypothèse pour lequel il n'est pas nécessaire de spécifier la forme de la distribution de la population étudiée.
Les méthodes non paramétriques requièrent peu d'hypothèses concernant la population étudiée et elles ignorent notamment l'hypothèse classique de la normalité de la population.
Dans notre contexte on a utilisé les deux tests non paramétriques suivants :

1- Le test des rangs Wilcoxon : pour savoir l'effet de variation du taux des valeurs manquantes sur la même base de données, afin de comparer les différentes méthodes de traitement des données manquantes sur la même base de données.
2- Le test de Friedman : ce test permet de comparer plusieurs méthodes de traitement des données manquantes entre elles (A, B, C, D) pour plusieurs bases de données.

1- Test des rangs appliqué au cas d'échantillons appariés (*Wilcoxon matched-pairs signed-ranks test*)

Le test T de WICOXON est un test des rangs pariés pour des données appartenant à deux échantillons dépendants; c'est l'équivalent du test T de Student utilisé pour la comparaison d'échantillons dépendants issus d'une distribution normale. Le test de Wilcoxon donne plus de poids à une paire qui montre une large différence entre les deux conditions qu'à une paire ayant une faible différence. Cela implique que l'on puisse dire quel membre d'une paire est plus grand que l'autre (donner le signe de la différence), mais aussi que l'on puisse ranger les différences en ordre croissant.

1-1 Méthode

Pour pouvoir expliquer ce test, on va utiliser du tableau 1, qui présente les résultats de la précision de deux base de données A et B, *di* = différence entre chaque paire, représentant la différence entre les valeurs de la précision appariées obtenues lors des deux traitements. Chaque paire a un *di*.

On range tous les *di* sans tenir compte de son signe. Dans ce cas, lorsque l'on range les *di*, un di de -2 est affecté d'un rang inférieur à celui d'un di de -3 ou +3. Puis réaffecter à chaque rang le signe de la différence.

Ensuite, on détermine la somme des rangs des différences positives $T^{+} = \sum R^{+}$ et la somme des rangs des différences négatives $T^{-} = \sum R^{-}$.

Si les traitements A et B sont équivalents, donc si H0 est vraie, la somme des rangs ayant un signe positif et celle des rangs ayant un signe négatif devraient être à peu près égales. Mais si la somme des rangs de signes positifs est très différente de celle des rangs de signes négatifs, nous en déduirons que le traitement A diffère du traitement B, et rejetterons l'hypothèse nulle. Donc, il y a rejet de H_0 lorsque la somme des rangs de signe négatif ou que celle des rangs de signe positif est faible.

Tableau.1 - Résultats du test des rangs appariés de Wilcoxon pour la précision avec un taux des valeurs manquantes de 5%, on utilisant la méthode d'imputation CD.

Test	Acc Sans valeurs manquantes (A)	Acc avec valeurs manquantes (B)	Différance *d*= A-B	Rang de $\lvert d \rvert$	Rang avec le signe de *d*
T1	100.00%	100.00%	0%	-	-
T2	83.33%	33.33%	50%	3.5	3.5
T3	100.00%	100.00%	0%	-	-
T4	100.00%	100.00%	0%	-	-
T5	100.00%	83.33%	16.67%	1.5	1.5
T6	100.00%	100.00%	0%	-	-
T7	100.00%	100.00%	0%	-	-
T8	100.00%	83.33%	16.67%	1.5	1.5
T9	100.00%	100.00%	0%	-	-
T10	100.00%	100.00%	0%	-	-
T11	83.33%	33.33%	50%	3.5	3.5
T12	100.00%	100.00%	0%	-	-

Il est possible que les deux valeurs de la précision d'une quelconque paire soient égales. Il n'y a pas de différence observée entre les deux traitements pour cette paire (d = 0). De telles paires sont abandonnées. N est alors égal au nombre de paires dont la différence entre les traitements n'est pas nulle.

Dans le cas ou deux ou plus des différences observées entre paire peuvent être égales entre elles. On donne alors le même rang à ces valeurs liées. Le rang affecté est la moyenne des rangs qu'auraient eu les diverses valeurs si elles avaient différentes. Ainsi, deux des paires observées présentent les différences suivantes : 16.67 et 16.67. Chaque paire aura le rang 1.5, car (1 + 2) / 2 = 1.5. La différence suivante aura alors le rang 3, puisque les rangs 1 et 2 ont déjà été utilisés.

T représente la valeur minimale de la somme des rangs positifs et la somme des rangs négatifs ($T = min\{T^+, T^-\}$). La table annexe 2, donne les valeurs critiques de T et leurs

niveaux de signification associés pour N. Si le T observé est égal ou inférieur à la valeur donnée dans la table pour un niveau de signification et pour le nombre de différences non nulles N, l'hypothèse nulle peut être rejeté à ce niveau de signification.

Pour l'exemple du tableau 3.3, On a :

$T^+ = 3.5 + 1.5 + 1.5 + 3.5 = 10$

$T^- = 0$

Avec T^+ = la somme des rangs positifs

T^- = la somme des rangs négatifs.

N= le nombre total des valeurs de *d* non nulles = 4.

$T = min\ \{T^+, T^-\} = 0$

La table A2.1[7] des valeurs critiques du test de wilcoxon, montre que pour N = 4, un $T_{0.5\ (2)},4 = 2$ nous permet de rejeter l'hypothèse nulle au seuil $P = 0{,}5$ pour un test bilatéral, c'est-a-dire la distribution avec traitement des valeurs manquantes n'est pas équivalente a la distribution sans valeurs manquantes. Par conséquent, nous conclurions, dans cet exemple, que la présence de 50% des valeurs manquantes affecte la précision de classificateur cbmLAD.

2- Test de FRIEDMAN.

Le test de FRIEDMAN est un test de comparaison de données appartenant à plus de deux échantillons dépendants. Ce test permet de comparer plusieurs méthodes de traitements des valeurs manquantes entre elles (A, B, C, D) pour plusieurs bases de données.

[7] Voir annexe n°2

Ce test distribue les données en un tableau à double entrée ayant ***b*** rangées et ***a*** colonnes. Les rangées représentent les différentes bases de données et les colonnes les différentes méthodes de traitement des données manquantes. Les données sont rangées. La détermination des rangs se fait pour chaque rangée séparément. Donc pour k méthodes de traitement des valeurs manquantes, les rangs de chaque rangée se répartissent entre 1 et k. Si deux ou plusieurs données sont identiques, nous leur attribuons le rang moyen. Par exemple, les données 86.11%, 97.22%, 97.22% et 97.22%, ne peuvent pas être classées comme 1, 2, 3 et 4. Les données identiques se voient attribuer un rang chacune (i.e., 2, 3, 4), puis on calcule le rang moyen des données identiques (i.e., (2+3+4)/3=3).

Supposons que l'on voulait étudier les résultats de 3 bases de données avec 4 méthodes de traitements des données manquantes. Pour un taux des valeurs manquantes de 5%, les résultats sont présentés dans le tableau suivant :

Tableau.2 - Résultats la précision avec un taux des valeurs manquantes de 5%, on appliquant quatre méthodes d'imputation pour trois bases de données.

Data-base	**La précision**			
	Méthode de substitution des VM			
	CD	**NNI**	**MCI**	**MMI**
IRIS	100,00%	100,00%	100,00%	100,00%
BEARING	93,75%	90,63%	93,23%	93,23%
DAG	86,11%	97,22%	97,22%	97,22%

L'étape suivante consiste à transformer les données de chaque ligne en rangs en attribuant le rang le plus petit à la donnée la plus petite, etc. Puisqu'il n'y a que 4 méthode de traitement, le rang supérieur ne doit pas dépasser 4. On obtient alors le tableau suivant :

Tableau.3 – Les résultats de la précision transformés en rangs.

Data-base	La précision			
	Méthode de substitution des VM			
	CD	**NNI**	**MCI**	**MMI**
IRIS	2.5	2.5	2.5	2.5
BEARING	4	1	2.5	2.5
DAG	1	3	3	3
Somme des rangs	**R_{CD} =7.5**	**R_{NNI} =6.5**	**R_{MCI} =8**	**R_{MMI} =8**

Pour comparer l'efficacité des traitements, nous devrons donc nous intéresser à la somme des rangs par traitements R_i. Ainsi, la statistique du test de Friedman s'écrit (Zar, J. H. 1999) [89] :

$$X_r^2 = \frac{12}{ba(a+1)} \sum_{i=1}^{a} R_i^2 - 3b(a+1) \qquad (5)$$

Le ***a*** représente le nombre de méthodes de traitement, R_i somme des rangs de chaque colonne, et le ***b*** représente le nombre de lignes (nombre des bases de données).

Avec cette formulation, on comprend que X_r^2 traduit l'idée d'une variabilité inter-traitements, elle correspond à la dispersion des rangs moyens conditionnels autour de la moyenne globale. Si les traitements se valent tous, nous obtiendrons $X_r^2 = 0$. Plus ils se démarqueront les uns des autres, plus X_r^2 prendra une valeur élevée.

La région critique du test correspond aux grandes valeurs de X_r^2, soit au risque p :

$$R.C.: X_r^2 \geq (X_r^2)_{p,a,b}$$

Lorsqu'il y a des ex-æquo à l'intérieur d'une ligne de valeurs, le principe des rangs moyens est utilisé c.-à-d. on réalise la péréquation des rangs affectés aux traitements qui présentent la même valeur. La statistique de test doit être corrigée pour tenir compte de l'ajustement, avec les caractéristiques suivantes : la statistique corrigée est toujours supérieure ou égale à la statistique non corrigée, s'il n'y a pas aucun ex-æquo, la statistique corrigée et non corrigée doivent être identiques, la correction sera d'autant plus sensible que le nombre d'ex-æquo est élevé.

Pour la ligne n_i : nous notons G_i le nombre de valeurs différentes, la valeur n_g étant répétée t_{ig} fois.

Le facteur de correction est donc (Zar, J. H. 1999) [89]:

$$C = 1 - \frac{\sum_{i=1}^{n} \sum_{g=1}^{Gi} \left(t_{ig}^{3} - t_{ig} \right)}{b \left(k^{3} - a \right)} \qquad (6)$$

La statistique de Friedman ajustée pour les ex-æquo s'écrit alors (Zar, J. H. 1999) [89] :

$$X^{2}_{r\ ajustée} = \frac{X_r^{2}}{C} \qquad (7)$$

Exemple de calcul :

Les résultats de la précision pour les quatre méthodes appliquées sur les trois bases de données sont représentés dans le tableau 2.

Si l'hypothèse nulle est vraie, la répartition des rangs dans chacune des colonnes doit être la même. C'est à dire que l'on doit s'attendre à avoir la même fréquence de 1, 2, 3 et 4 dans

chacune des colonnes, ce qui a pour conséquence que la somme des rangs dans devrait être à peu près la même.

En notant R, la somme des rangs de la colonne i (par méthode) pour chaque taux de valeurs manquantes.

A partir du tableau 2, les résultats du test de Friedman sont représentés par le tableau 3.

Tableau.4 – Les résultats du test de Friedman.

Critères	Valeur	Formule
a	4	
b	3	
$\sum Rj^2$	226.5	
X_r^2	0.3	(5)
Valeur critique à *p= 0.05*	7.4	
Ex-æquo	6	
Correction d'Ex-æquo	0.5	(6)
X_r^2 (après correction)	0.6	(7)
Décision	Accepter H_0	

Le seuil critique du test à 5% est égal à 7,4 (voir La table A2.2 [8]), Les méthodes de traitement présentent des performances identiques au risque 5%.

Si l'hypothèse nulle (que tous les échantillons, colonnes, proviennent de la même population) est vraie, la distribution des rangs dans chaque colonne sera due à la chance, et les différents rangs apparaîtront avec la même fréquence. Le total des rangs par colonne (R_j) sera aléatoire. Mais, si les observations sont dépendantes d'au moins une des conditions (si H_0 est fausse), alors le total des rangs par colonnes devrait varier d'une colonne à l'autre. Le test de Friedman teste si les totaux des rangs par colonne différent significativement.

[8] Voir annexe n°2

ANNEXE 2 : TABLES STATISTIQUES UTILISÉES

Annexe 2
Table A2.1 :Valeurs critiques pour le test de Wilcoxon

n	$\alpha(2)$: 0.50 / $\alpha(1)$: 0.25	0.20 / 0.10	0.10 / 0.05	0.05 / 0.025	0.02 / 0.01	0.01 / 0.005	0.005 / 0.0025	0.001 / 0.0005
4	2	0						
5	4	2	0					
6	6	3	2	0				
7	9	5	3	2	0			
8	12	8	5	3	1	0		
9	16	10	8	5	3	1	0	
10	20	14	10	8	5	3	1	
11	24	17	13	10	7	5	3	0
12	29	21	17	13	9	7	5	1
13	35	26	21	17	12	9	7	2
14	40	31	25	21	15	12	9	4
15	47	36	30	25	19	15	12	6
16	54	42	35	29	23	19	15	8
17	61	48	41	34	27	23	19	11
18	69	55	47	40	32	27	23	14
19	77	62	53	46	37	32	27	18
20	86	69	60	52	43	37	32	21
21	95	77	67	58	49	42	37	25
22	104	86	75	65	55	48	42	30
23	114	94	83	73	62	54	48	35
24	125	104	91	81	69	61	54	40
25	136	113	100	89	76	68	60	45
26	148	124	110	98	84	75	67	51
27	160	134	119	107	92	83	74	57
28	172	145	130	116	101	91	82	64
29	185	157	140	126	110	100	90	71
30	198	169	151	137	120	109	98	78
31	212	181	163	147	130	118	107	86
32	226	194	175	159	140	128	116	94
33	241	207	187	170	151	138	126	102
34	257	221	200	182	162	148	136	111
35	272	235	213	195	173	159	146	120
36	289	250	227	208	185	171	157	130
37	305	265	241	221	198	182	168	140
38	323	281	256	235	211	194	180	150
39	340	297	271	249	224	207	192	161
40	358	313	286	264	238	220	204	172
41	377	330	302	279	252	233	217	183
42	396	348	319	294	266	247	230	195
43	416	365	336	310	281	261	244	207
44	436	384	353	327	296	276	258	220
45	456	402	371	343	312	291	272	233
46	477	422	389	361	328	307	287	246
47	499	441	407	378	345	322	302	260
48	521	462	426	396	362	339	318	274
49	543	482	446	415	379	355	334	289
50	566	503	466	434	397	373	350	304
51	590	525	486	453	416	390	367	319
52	613	547	507	473	434	408	384	335
53	638	569	529	494	454	427	402	351
54	668	592	550	514	473	445	420	368
55	688	615	573	536	493	465	438	385
56	714	639	595	557	514	484	457	402
57	740	664	618	579	535	504	477	420
58	767	688	642	602	556	525	497	438
59	794	714	666	625	578	546	517	457
60	822	739	690	648	600	567	537	476

Réf : Zar, J. H., 1999: Biostatistical analysis, Prentice-Hall Inc, Englewood Cliffs, New Jersey

Table A2.2 :Valeurs critiques pour le test de Friedman

a (n)	b (M)*	α: 0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
3	2	3.000	4.000							
3	3	2.667	4.667	(6.000)	6.000					
3	4	2.000	4.500	6.000	6.500	(8.000)	(8.000)	8.000		
3	5	2.800	3.600	5.200	6.400	(8.400)	8.400	(10.000)	(10.000)	10.000
3	6	2.330	4.000	5.330	7.000	8.330	9.000	(10.330)	10.330	12.000
3	7	2.000	3.714	5.429	7.143	8.000	8.857	10.286	11.143	12.286
3	8	2.250	4.000	5.250	6.250	7.750	9.000	9.750	12.000	12.250
3	9	2.000	3.556	5.556	6.222	8.000	9.556	10.667	11.556	12.667
3	10	1.800	3.800	5.000	6.200	7.800	9.600	10.400	12.200	12.600
3	11	4.636	3.818	4.909	6.545	7.818	9.455	10.364	11.636	13.273
3	12	1.500	3.500	5.167	6.167	8.000	9.500	10.167	12.167	12.500
3	13	1.846	3.846	4.769	6.000	8.000	9.385	10.308	11.538	12.923
3	14	1.714	3.571	5.143	6.143	8.143	9.000	10.429	12.000	13.286
3	15	1.733	3.600	4.933	6.400	8.133	8.933	10.000	12.133	12.933
4	2	3.600	5.400	(6.000)	6.000					
4	3	3.400	5.400	6.600	7.400	8.200	(9.000)	(9.000)	9.000	
4	4	3.000	4.800	6.300	7.800	8.400	9.600	(10.200)	10.200	11.100
4	5	3.000	5.160	6.360	7.800	9.240	9.960	10.920	11.640	12.600
4	6	3.000	4.800	6.400	7.600	9.400	10.200	11.400	12.200	12.800
4	7	2.829	4.886	6.429	7.800	9.343	10.371	11.400	12.771	13.800
4	8	2.550	4.800	6.300	7.650	9.450	10.350	11.850	12.900	13.800
4	9			6.467	7.800	9.133	10.867	12.067		14.467
4	10			6.360	7.800	9.120	10.800	12.000		14.640
4	11			6.382	7.909	9.327	11.073	12.273		14.891
4	12			6.400	7.900	9.200	11.100	12.300		15.000
4	13			6.415	7.985	7.369	11.123	12.323		15.277
4	14			6.343	7.886	9.343	11.143	12.514		15.257
4	15			6.440	8.040	9.400	11.240	12.520		15.400
5	2			7.200	7.600	8.000	8.000			
5	3			7.467	8.533	9.600	10.133	10.667		11.467
5	4			7.600	8.800	9.800	11.200	12.000		13.200
5	5			7.680	8.960	10.240	11.680	12.480		14.400
5	6			7.733	9.067	10.400	11.867	13.067		15.200
5	7			7.771	9.143	10.514	12.114	13.257		15.657
5	8			7.800	9.300	10.600	12.300	13.500		16.000
5	9			7.733	9.244	10.667	12.444	13.689		16.356
5	10			7.760	9.280	10.720	12.480	13.840		16.480
6	2			8.286	9.143	9.429	9.714	10.000		
6	3			8.714	9.857	10.810	11.762	12.524		13.286
6	4			9.000	10.286	11.429	12.714	13.571		15.286
6	5			9.000	10.486	11.743	13.229	14.257		16.429
6	6			9.048	10.571	12.000	13.619	14.762		17.048
6	7			9.122	10.674	12.061	13.857	15.000		17.612
6	8			9.143	10.714	12.214	14.000	15.286		18.000
6	9			9.127	10.778	12.302	14.143	15.476		18.270
6	10			9.143	10.800	12.343	14.229	15.600		18.514

Réf : Zar, J. H., 1999: Biostatistical analysis, Prentice-Hall Inc, Englewood Cliffs, New Jersey