TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

DBO_5	Demande biochimique en oxygène sur 5 jours
DCO	Demande chimique en oxygène
DDC	Dépassement du débit de conception
KNN	K-nearest neighbors
Logreg	Logistic regression imputation
MA	Manque aléatoire
MDO	Manque dépend des observations
MELCC	Ministère de l'Environnement et de la Lutte contre les Changements Climatiques
MEX	Mexique
MES	Matière en suspension
MICE	Multivariate imputation by chained equations
MNA	Manque non aléatoire
NRMSE	Normalized root mean squared error
OOB	Out-Of-Bag
PFC	Proportion of falsely classified
PMM	Predictive mean matching
Polr	Proportional odds model
Polyreg	Polytomous regression imputation
Ptot	Phospore total
TRT	Type de traitement

LISTE DES SYMBOLES ET UNITÉS DE MESURE

s	seconde
m	minute
kg	kilogramme
m^3	mètre cube
cm	centimètre
MPa	Mégapascal
mD	millidarcy
Hz	Herz
dB	Décibel
$mg \cdot L^{-1}$	milligramme par litre

INTRODUCTION

L'acquisition de données permet de constituer des bases de données en vue d'être exploitées dans le processus de prise de décision. Cependant, un dysfonctionnement des processus d'acquisitions peut générer des lacunes dans ces bases de données. C'est un problème auquel doivent faire face la plupart des domaines qui suscitent l'intérêt aujourd'hui. C'est notamment le cas en médecine, en biologie, en sciences humaines ainsi qu'en environnement. Des raisons différentes peuvent être à l'origine d'une donnée manquante, comme le mauvais fonctionnement d'un appareil de mesure, une erreur humaine suite à une mesure manuelle ou encore une donnée aberrante qui aurait été supprimée. Il est donc bien souvent inévitable de faire face à un manque de données, et ceci, peu importe le processus de collecte d'information. Les enjeux soulevés par le manque de données résident dans l'importance de l'information dans le processus de prise de décision. En effet, l'absence de certaines données peut engendrer une incertitude générant un biais sur le choix final, voir rendre impossible la décision.

Combler ces lacunes par de nouvelles prises de données est généralement impossible, dans la plupart des cas, du fait de contraintes financières ou temporelles. C'est la raison pour laquelle un grand nombre de travaux de recherche ont été récemment menés dans ce domaine. De ces travaux ont émergé plusieurs approches différentes au problème des données manquantes (Farhangfar, Kurgan et Pedrycz, 2007). Certaines méthodes couramment utilisées proposent d'omettre les enregistrements d'un système pour lesquels il manque un, ou plusieurs attributs. C'est par exemple le cas de méthodes statistiques usuelles comme les *listwise deletion methods* (méthodes de suppression par liste). Néanmoins, ces méthodes ne sont applicables que dans les cas où le pourcentage de données manquantes est faible. De plus, l'échelle grandissante des problématiques soulevées ces dernières années a précipité leur obsolescence (Pigott, 2001). En effet, l'émergence des nouvelles technologies de l'information et des sciences associées au traitement des *big data* (mégadonnées) ont marqués l'avènement des méthodes d'imputation de données manquantes. Cependant, la majorité de ces méthodes sont limitées au traitement d'un seul type de variable soit de type qualitatif, soit de type quantitatif. Par ailleurs, les méthodes d'imputation ne prennent pas toutes en compte les interactions et

non-linéarités qui peuvent exister au sein des variables qui définissent le système à l'étude. Dans une volonté de combler cette lacune de la littérature scientifique, Stekhoven et Bühlmann (2012) ont développé la méthode d'imputation missForest qui permet de traiter n'importe quel type de données (qualitatives, quantitatives et mixtes).

Cette méthode est basée sur le principe des *random forest* (forêts aléatoires) de Breiman (2001). Ces outils de régression et de classification sont réputés pour leur robustesse face aux données de type mixte, aux systèmes de grandes dimensions et aux structures de données complexes (Breiman, 2001). Cela soulève la question de l'applicabilité de la méthode missForest aux problématiques environnementales. En effet, compte tenu de la diversité des problématiques rencontrées dans ce domaine, les bases de données à traiter y sont d'une grande variabilité. Les enjeux abordés dans le cadre du présent travail sont donc associés à la gestion des données en environnement.

La présente recherche vise à déterminer si la méthode missForest apporte une solution au problème du manque de données rencontré en environnement et plus spécifiquement dans le contexte du suivi de la performance de traitement des stations d'épuration. Pour répondre à cet objectif principal, deux sous-objectifs ont été définis, soit : (1) évaluer comparativement missForest avec deux des principales méthodes d'imputation sur une dizaine de bases de données complètes et variées; et (2) évaluer l'applicabilité de la méthode missForest appliquée aux données enregistrées concernant les paramètres de traitement des stations d'épuration des eaux usées du Québec.

Le présent document est organisé en quatre chapitres. Le premier présente l'état des connaissances dans le domaine de l'imputation ainsi que le contexte de l'étude avant d'aborder les points théoriques nécessaires à la compréhension de ce document. Les deuxième et troisième chapitres abordent respectivement la méthodologie appliquée et les résultats obtenus. La discussion abordant notamment les perspectives et les recommandations sera présentée dans le quatrième chapitre. Elle sera suivie par la conclusion.

CHAPITRE 1

ÉTAT DES CONNAISSANCES

Ce premier chapitre dresse un état des connaissances liées au sujet d'étude, il est divisé en deux parties. En premier lieu, une revue de littérature pour le domaine de l'imputation de données manquantes est présentée. La deuxième partie entre dans le détail du fonctionnement des méthodes d'imputations abordées dans ce mémoire.

Tout processus de prise de décision nécessite une collecte d'information. Cette collecte d'information aboutit à l'établissement d'une base de données qui doit être analysée afin d'en extraire les connaissances pertinentes à la prise de décision. Cependant, pour des raisons qui diffèrent selon le secteur d'activité, il est fréquent que ces bases de données soient incomplètes. Or, une base de données incomplète complique l'analyse et peut donner lieu à des conclusions erronées. Elles nécessitent donc d'être prétraitées avant d'être analysées (Pyle, Editor et Cerra, 1999). En effet, de nombreuses méthodes d'analyse comme les analyses en composantes principales ou par partitionnement de données nécessitent des données complètes pour fonctionner (Liao *et al.*, 2014). Bien que certaines méthodes ne posent pas cette contrainte, des données manquantes engendrent systématiquement une perte d'efficacité statistique (Wang et Wang, 2010), d'où la nécessité de les considérer. Il existe deux approches différentes qui évitent d'avoir recours à une récolte de données supplémentaires : la suppression par cas et l'imputation (Luengo, García et Herrera, 2012). Cependant, en supprimant les enregistrements ou les attributs incomplets, les méthodes de suppression par cas risquent la perte d'information pertinente. C'est pourquoi cette étude se concentre sur l'alternative la plus fréquemment utilisée (Sessa et Syed, 2017), l'imputation de données manquantes.

1.1 Imputation de données manquantes

L'imputation de données manquantes consiste à combler les « trous » dans des bases de données incomplètes par des valeurs substituées et identifiées comme des « données imputées ». La manière de combler un manque de données diffère selon la méthode

d'imputation utilisée. Les premières méthodes d'imputation à avoir été utilisées sont basées sur des fondamentaux de mathématiques, c'est notamment le cas des méthodes de complétion par combinaison linéaire. La plus utilisée d'entre elles étant la méthode d'imputation par la moyenne qui se contente d'effectuer une moyenne sur les données observées (Wikistat.fr, 2015). Ces méthodes ont par la suite évolué vers des méthodes plus complexes analysant la distribution des données pour imputer.

Les méthodes d'imputation ont déjà fait leurs preuves dans plusieurs domaines. C'est notamment le cas en biologie (Celton *et al*., 2010; Gromski *et al*., 2014; Liao *et al*., 2014) et en médecine (Bousquet, 2012; Waljee *et al*., 2013; Dávila, 2015). Or, l'environnement est un domaine interdisciplinaire qui intègre les sciences de l'information, physiques et biologiques. Étant donné les similitudes qu'il existe entre les problématiques soulevées par ces domaines, il est possible que l'efficacité des méthodes d'imputation soit transposable aux sciences environnementales. C'est ce type de méthodes qui seront abordées dans la suite de ce chapitre.

1.1.1 Catégories de méthodes d'imputation

L''utilisation de méthodes d'imputation pour résoudre le problème de données manquantes est un travail qui suscite l'intérêt depuis plusieurs dizaines d'années. Les bases de cette discipline ont été établies par les travaux de Little et Rubin (1987), tout particulièrement dans le domaine des analyses statistiques. Suite à leurs travaux, de nombreuses études sont venues approfondir les connaissances en termes d'imputation de données jusqu'à ce que dans les années 90, les puissances de calcul de plus en plus accessibles ont permis l'arrivée des algorithmes d'apprentissage automatique. Ces méthodes ont révolutionné l'exploration de données, notamment grâce à leur capacité à traiter les problèmes de plus grandes dimensions (Bzdok, Altman et Krzywinski, 2018).

L'apprentissage automatique est un domaine issu de l'informatique et de l'intelligence artificielle. Contrairement aux méthodes statistiques usuelles telles que la méthode d'imputation par la moyenne, les méthodes d'apprentissage automatique obtiennent de

l'information à partir de données sans avoir recours à une programmation explicite. Elles ont donc besoin de moins d'intervention humaine. Afin de générer un modèle d'imputation, ces méthodes recherchent des schémas de données généralisables.

Suite à l'arrivée des algorithmes d'apprentissage automatique, de nombreuses méthodes d'imputation de données manquantes ont émergé (plus de 50 recensées dans la littérature). Parmi les différentes manières d'imputer récemment proposées, deux caractéristiques fondamentales différencient les méthodes d'imputation : les méthodes à imputation unique et à imputation multiple (Gómez-Carracedo *et al*., 2014), et les méthodes paramétriques et non paramétriques.

1.1.1.1 Méthodes à imputation unique et à imputation multiple

Méthodes à imputation unique : la plupart des méthodes d'imputation existantes entrent dans cette catégorie. Le principe de l'imputation unique vise à imputer une donnée manquante une seule fois; une seule valeur lui est donc associée. Cependant, les données imputées sont considérées comme étant les données qui auraient été observées si la base de données avait été complète, ce qui n'est jamais certain. Par conséquent, ces méthodes ne prennent pas en compte l'incertitude des données imputées (Zhang, 2016).

Méthodes à imputation multiple : ces méthodes effectuent plusieurs imputations pour reconstruire les données. À chaque donnée manquante est associé plusieurs valeurs, et toutes ces valeurs sont possiblement le résultat cherché (Buuren et Groothuis-Oudshoorn, 2011; Buuren et Oudshoorn, 1999). En d'autres termes, ces algorithmes génèrent plusieurs versions différentes de la base de données imputée. Une fois les imputations multiples terminées, une analyse est menée sur chaque base de données imputée et suite à cette analyse, les résultats sont combinés pour obtenir une base de données complète. Ce principe d'imputation prend en compte l'incertitude qui existe sur la valeur à imputer à chaque imputation et réduit donc le biais qui en découle. Le fonctionnement de ce type de méthodes est illustré à la Figure 1.1.

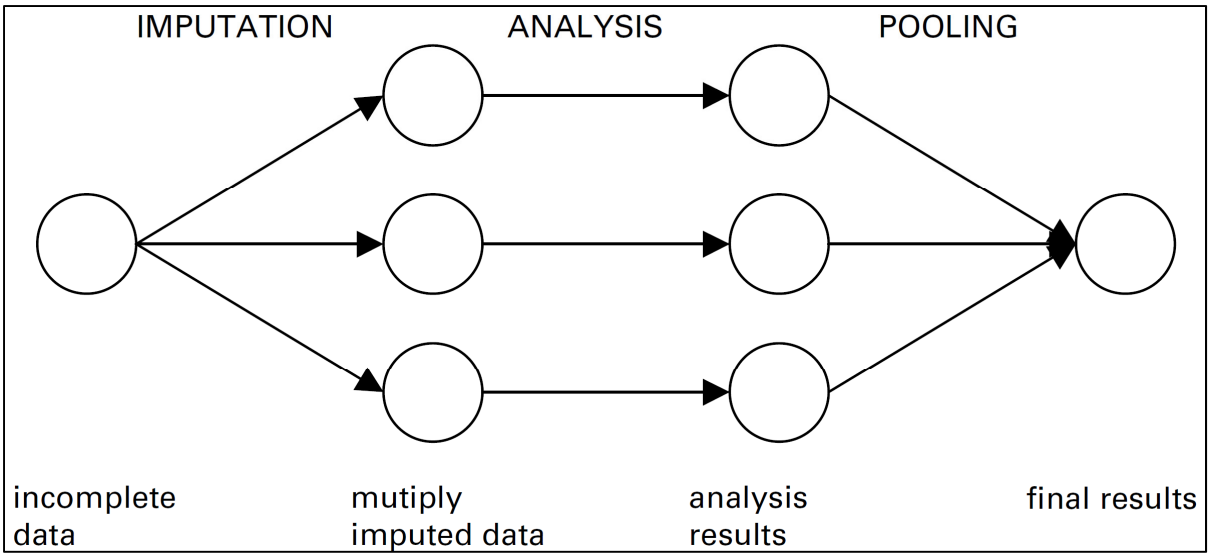

Figure 1.1 Processus des algorithmes d'imputation multiple
Tirée de Buuren et Oudshoorn (1999)

1.1.1.2 Méthodes de type paramétrique et non paramétrique

Les méthodes paramétriques imputent les données manquantes en faisant des suppositions sur la distribution des variables à partir des données observables. Cette distribution dépend du réglage d'un ensemble de paramètres fixes. Ce type de fonctionnement peut induire un biais car les suppositions faites sur la distribution des données ne sont pas nécessairement vérifiées. Par exemple, certaines de ces méthodes définissent les variables quantitatives par l'intermédiaire de termes linéaires et sans interactions, des termes non linéaires déterminants peuvent donc être omis. C'est notamment le cas de la régression linéaire. À l'inverse, les méthodes non paramétriques ne sont pas régies par des lois de probabilités paramétriques et ne font donc pas de supposition sur la distribution des données (Seaman, Bartlett et White, 2012; Shah *et al.*, 2014). La nature paramétrique ou non d'une méthode n'est pas liée au nombre d'imputations nécessaire à la reconstruction des données.

1.1.2 Conditions d'application

Parce que l'algorithme des méthodes d'imputation diffèrent dans leur manière d'imputer des données, leurs conditions d'application également. En effet, selon l'approche d'imputation utilisée, les caractéristiques de la base de données à imputer peuvent rendre impossible l'imputation des données. Deux critères identifiés à partir de la littérature sont susceptibles d'avoir un impact sur la capacité des méthodes à imputer : la structure de la base de données et le type de manque de données rencontré.

1.1.2.1 Caractéristiques structurelles des bases de données

Dans la littérature, plusieurs études suggèrent qu'il existe un lien entre la capacité des méthodes à imputer et les différentes caractéristiques structurelles des bases de données à imputer. C'est par exemple le cas pour la dimension de la base de données. En effet, les premières méthodes statistiques usuelles ont été conçues pour traiter des bases de données contenant au plus quelques dizaines de variables. Certaines de ces méthodes sont donc dans l'incapacité d'imputer les problèmes de plus grandes dimensions (Bzdok, Altman et Krzywinski, 2018).

Selon leur approche d'imputation (approches à imputation unique ou multiple et de type paramétrique ou non paramétrique), les méthodes ne sont pas impactées de la même manière par les caractéristiques structurelles d'une base de données. En raison des hypothèses faites sur la distribution des données, ce sont les méthodes paramétriques qui sont les plus affectées. Par exemple, à cause de l'hypothèse de linéarité faite par certaines de ces méthodes (comme la régression linéaire), une non-linéarité dans le système peut induire un biais dans les résultats. Par ailleurs, ces méthodes peuvent rencontrer des difficultés à traiter les bases de données dans lesquelles deux variables sont hautement corrélées (Shah *et al.*, 2014). Ainsi, une colinéarité peut empêcher une méthode paramétrique d'imputer des données manquantes. Cela est d'autant plus problématique dans la pratique, car il est impossible d'obtenir les différents paramètres structurels exacts d'une base de données sans la totalité des données.

1.1.2.2 Type de manque de données

Il existe plusieurs types de manque de données différents et ces derniers peuvent affecter la qualité d'une imputation (Misztal, 2013). De plus, selon le manque rencontré, certaines méthodes peuvent ne pas fonctionner. Une classification récente, majoritairement employée actuellement, en différencie trois types (Little et Rubin, 2002). Pour des fins de compréhension, les termes suivants, associés à chacun des types, sont proposés :

Manque aléatoire – MA (*missing completely at random)* : cas dans lequel n'importe quel élément d'une base de données peut être manquant. Les manques de données se présentent indépendamment les uns des autres et sont répartis uniformément. Ils ne dépendent donc pas des paramètres et variables du système. En d'autres termes, les données manquantes constituent un sous-ensemble aléatoire de l'ensemble des données. La distribution des données manquantes est supposément identique à celle des données observées. C'est néanmoins une hypothèse stricte.

Manque dépendant des observations – MDO (*missing at random)* : dans le cas MDO, le manque de données n'est pas aléatoire. La probabilité qu'un élément soit manquant dépend des autres variables observables du système. Afin d'illustrer ce type de données manquantes, Baradi et Enders (2010) présentent l'exemple suivant : une école fait passer un test d'aptitudes en mathématiques et les étudiants ayant obtenu les meilleures notes participent par la suite à un cours avancé. Les notes qu'obtiendront les élèves au cours avancé de mathématiques sont MDO car le manque de données dépend des notes obtenues au test d'aptitudes, les élèves qui ont échoué n'auront pas de notes pour le cours avancé (Baraldi et Enders, 2010).

Manque non aléatoire – MNA (*not missing at random)* : les données MNA sont les données qui ne sont ni MA, ni MDO. Ici la probabilité qu'un élément soit manquant dépend directement de sa propre valeur (Liu et Lei, 2006). Afin d'illustrer ce type de données manquantes, Baradi et Enders (2010) donnent l'exemple suivant : un lycée fait passer un questionnaire d'auto-évaluation à ses étudiants pour déterminer leur consommation d'alcool. Les résultats de ce

questionnaire seront MNA si les gros buveurs décident de ne pas répondre à certaines questions pour éviter des ennuis vis-à-vis du lycée. Dans ce cas, la probabilité qu'une note soit manquante dépend de la consommation d'alcool de l'élève.

Dans la pratique, de ces trois types de manque différents, seules les données MA sont empiriquement vérifiables par des tests statistiques. À l'inverse, les données MDO et MNA ne peuvent être vérifiées ou différenciées car elles dépendent de données non observables. En effet, pour comprendre le lien entre la probabilité qu'une donnée soit manquante et les valeurs sous-jacentes d'une variable incomplète, il est nécessaire d'avoir des connaissances sur les données manquantes du système (Baraldi et Enders, 2010). De plus, contrairement aux données MA et MDO, dans les cas MNA rien n'indique que la distribution des données manquantes d'une variable coïncide avec celle des données observables. Les données MNA nécessitent d'expliciter le mécanisme du manque de données sans quoi l'imputation n'est pas possible (Vaquero, 2018). Il est donc souvent nécessaire d'avoir des connaissances sur la raison du manque afin de les traiter.

Pour ces raisons, certaines méthodes d'imputation ne sont pas en mesure de traiter les problèmes MNA. C'est notamment le cas des méthodes à imputation multiple qui supposent des données MA ou MDO. De nombreuses méthodes ne fournissent d'estimateurs que dans le cas de données MA et un biais peut résulter d'un non-respect de cette supposition de départ (Baraldi et Enders, 2010). C'est pourquoi la plupart des études de ce domaine ont été effectuées sur des données MA. Il a par ailleurs été avancé (Farhangfar, Kurgan et Dy, 2008; Matsubara *et al.*, 2008) que seules les analyses portant sur des données MA peuvent donner des résultats non-biaisés car elles seules permettent de faire des inférences à partir des données observables grâce à l'hypothèse faite sur la distribution.

1.1.3 Études de performance

Avec l'essor du domaine de l'imputation de données manquantes, un grand nombre d'études comparatives visant à évaluer les performances de méthodes d'imputation ont vu le jour. Dans

la plupart des études menées récemment, les performances de quelques méthodes d'imputation sont comparées sur plusieurs bases de données en faisant varier le pourcentage de données manquantes. L'objectif est souvent de mettre en avant les performances et particularités d'une ou plusieurs de ces méthodes.

Cette section présente chronologiquement quelques publications scientifiques clefs dont l'objectif est d'évaluer les performances de différentes méthodes d'imputation. Dans la littérature, trois méthodes ont particulièrement suscité l'intérêt des chercheurs : la méthode des *K-nearest neighbors* – KNN (la méthode des *K*-plus proches voisins) (Troyanskaya *et al.*, 2001), *multivariate imputation by chained equations* – MICE (imputation multivariée par équations en chaîne) (Buuren et Oudshoorn, 1999) et missForest (Stekhoven et Bühlmann, 2012).

La première étude comparative qui a été menée (Grzymala-Busse et Hu, 2001) s'intéresse à neuf méthodes d'imputation différentes dont se démarque C4.5, une méthode basée sur l'apprentissage automatique. Ce travail a été complété en introduisant la méthode d'imputation basée sur la recherche des plus proches voisins (Batista et Monard, 2003). Ces études avancent que l'imputation effectuée avec KNN est plus précise, bien que cela devienne moins vrai en présence de bases de données complexes.

Farhangfar *et al*. (2008) comparent quant à eux des méthodes d'imputation classiques à imputation unique (comme l'imputation par la moyenne) avec une méthode d'imputation multiple basée sur la régression multinomiale. Cependant, les résultats obtenus ne suffisent pas à départager ces méthodes (Farhangfar, Kurgan et Dy, 2008). Dans la même lignée, Garcia-Laencina *et al*. (2010) considèrent quatre méthodes d'imputation, dont KNN, sur trois bases de données (deux réelles et une artificielle). Les résultats de leur étude suggèrent que chaque cas réel mérite une analyse précise pour orienter le choix de la méthode d'imputation à utiliser (García-Laencina, Sancho-Gómez et Figueiras-Vidal, 2010).

En 2012, Stekhoven *et al*. (2012) introduisent missForest (Stekhoven et Bühlmann, 2012). Ils évaluent ses performances en la comparant aux méthodes KNN, MICE et missPALasso sur onze bases de données réelles. Les résultats montrent que missForest est plus performante que les autres méthodes dans pratiquement tous les cas. Une étude encore plus récente (Mandel, 2015) compare six méthodes d'imputation sur quatre bases de données réelles. Parmi les méthodes étudiées figurent MICE et une méthode basée sur KNN, et bien que leurs performances soient du même ordre, l'avantage semble en faveur de la méthode des *K* plus proches voisins.

Plus récemment, une étude a évalué la méthode KNN avec quatre autres méthodes en utilisant cinq bases de données réelles (Aljuaid et Sasi, 2017). À nouveau, la méthode KNN est l'approche la plus efficace et la plus polyvalente, même s'il ne s'agissait pas de la plus performante en termes de temps d'imputation, en particulier face à des bases de données de grandes dimensions.

Bien que les études citées précédemment ont toutes contribué à apporter des connaissances sur les méthodes d'imputations, elles procèdent de manières différentes et elles présentent leurs propres limites. Par exemple, certaines études ne s'intéressent qu'à une seule catégorie de méthodes d'imputation (Zhang, 2016; Vaquero, 2018; Le et Tan, 2018; Wu, Jia et Enders, 2015). De plus, plusieurs travaux ont étudié l'imputation de données manquantes dans un seul domaine en particulier. C'est principalement le cas pour les domaines de la médecine (Dávila, 2015; Waljee *et al.*, 2013) et de la biologie (Celton *et al.*, 2010; Liao *et al.*, 2014; Gromski *et al.*, 2014), où les chercheurs ont été confrontés à des problèmes à haute dimension. C'est également le cas des travaux de Stekhoven (2012) introduisant la méthode missForest. D'autres études, quant à elles, ne s'intéressent qu'à un seul type de données, qu'elles soient quantitatives (Solaro *et al.*, 2014), qualitatives (Wu, Jia et Enders, 2015), ou même binaires (Ghorbani et Desmarais, 2017). Toutefois, pour déterminer si une méthode d'imputation est apte à traiter la grande variété des problèmes rencontrés en environnement, il apparaît nécessaire de la confronter à des bases de données variant en termes de dimension et de la nature de ses variables (Aljuaid et Sasi, 2017; Mandel, 2015).

Par ailleurs, malgré la multitude d'études citées précédemment et les approches différentes au problème de l'imputation de données manquantes, il apparaît que peu d'attention ait été portée aux raisons qui justifient la qualité d'une imputation effectuée par une méthode. Quelques études (Luengo, García et Herrera, 2012; Brock *et al.*, 2008) proposent des approches ayant pour but de sélectionner la méthode optimale. Les travaux de Solaro *et al.* (2014) mettent spécifiquement de l'avant les liens qui existent entre la précision d'une imputation et la structure de la base de données traitée. Au cours de leurs travaux étude, Solaro *et al*. (2014) ont abordé les limites de missForest comparativement à d'autres méthodes d'imputations spécifiquement mises au point pour traiter les problèmes quantitatifs. Leurs travaux ont également montré que lors d'une imputation, il était nécessaire de s'intéresser à la nature de la distribution des variables et aux interactions entre ces dernières pour identifier les méthodes susceptibles d'être performantes. Le critère principal mis en avant par Solaro *et al*. (2014) est la corrélation entre les variables ainsi que l'effet de ces interactions qui peut être amplifié par une possible asymétrie dans la distribution des variables. Le principe des distributions asymétriques est présenté à la Figure 1.2. La littérature semble donc indiquer que, selon la nature du problème rencontré, pour identifier la méthode d'imputation la plus performante, il est nécessaire de comprendre les paramètres qui ont un impact sur les performances des méthodes d'imputation. En l'occurrence, il s'agit de la nature des interactions entre les variables des bases de données étudiées.

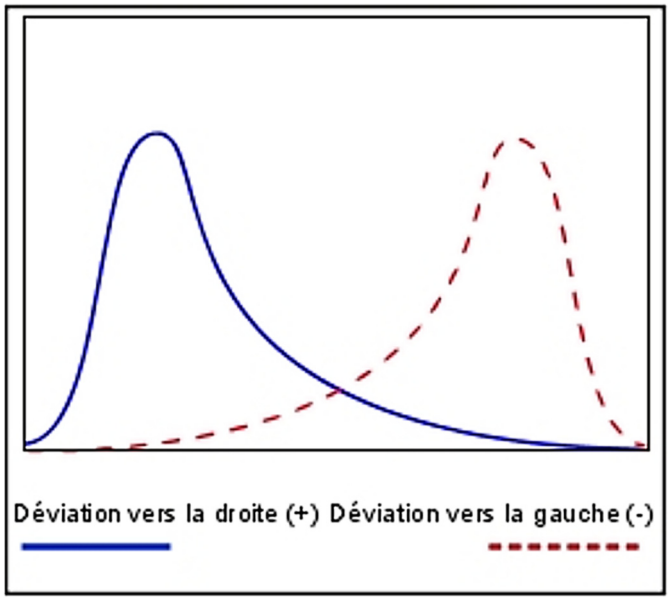

Figure 1.2 Diagramme illustrant les asymétries positives et négatives
Tirée de IBM

Dans une étude plus récente, Solaro *et al*. (2015) a fourni des indicateurs permettant de décrire la structure des bases de données. Cependant, leur approche ne prend en compte que les variables quantitatives et il existe des cas de figure où il est nécessaire d'étudier la corrélation entre deux variables de nature différente (quantitative et qualitative). Les premiers travaux portant sur cette problématique proposent de convertir les variables quantitatives d'une base de données en variables qualitatives puis d'étudier leur corrélation (Johansson, Jern et Johansson, 2008). Cependant, cette conversion induit un certain degré de discrétisation. Une autre approche propose la conversion des variables qualitatives en variables quantitatives (Zhang *et al.*, 2015) où la valeur numérique attribuée à une classe de la variable qualitative est la moyenne des valeurs correspondantes dans la variable quantitative concernée. Les valeurs des classes dépendent donc de la variable quantitative, ce qui impose d'effectuer une nouvelle conversion pour chaque calcul de coefficient de corrélation.

1.2 Caractéristiques des principales méthodes d'imputation

Cette section présente et décrit le fonctionnement des principales méthodes d'imputation identifiées à partir de la littérature : KNN, MICE et missForest. Comme l'illustre le Tableau 1.1, ces méthodes diffèrent dans leurs caractéristiques. Elles offrent donc un spectre étendu de ce qu'il est possible d'accomplir grâce à l'imputation de données manquantes. Cependant, deux caractéristiques les réunissent : leur capacité à traiter les bases de données mixtes et le fait qu'elles soient en mesure d'imputer, peu importe la dimension du problème rencontré.

Tableau 1.1 Caractéristiques des méthodes d'imputation de données manquantes utilisées

Méthodes d'imputation	Paramétrique	Itérative	Imputation unique	Apprentissage automatique
KNN	Non	Non	Oui	Oui
MICE	Oui	Oui	Non	Non
missForest	Non	Oui	Oui	Oui

1.2.1 KNN

KNN est une méthode d'imputation basée sur la recherche des *K* plus proches voisins. Elle a été initialement introduite (Troyanskaya *et al.*, 2001) pour l'étude de l'expression des gènes. Soit une base de données avec *n* enregistrements et *p* variables et en son sein, une donnée manquante à l'enregistrement *i*. L'imputation de cette donnée s'effectue en deux étapes :

- Étape 1 : les distances entre les *n-1* paires d'enregistrements (qui contiennent l'enregistrement *i*) sont calculées à partir des variables complètes afin d'identifier les *K* plus proches voisins de l'enregistrement *i*. Pour être capable de prendre en compte les données quantitatives et qualitatives, la distance de Gower est utilisée (Gower, 1971).
- Étape 2 : la donnée manquante de l'enregistrement *i* est calculée à partir des *K* plus proches voisins qui possèdent la donnée associée à la variable concernée. Pour les données quantitatives, il s'agit de faire une moyenne des données des *K* plus proches voisins. Pour les données qualitatives, c'est un vote sur la majorité sur les *K* données.

Même si cela peut paraître contre-intuitif, la méthode KNN n'est pas paramétrique. C'est en raison du fait que le nombre de paramètres n'augmente pas avec le nombre de dimensions du système. De plus, *K* est un hyper paramètre; il ne peut donc pas être déterminé par l'algorithme KNN et est à fixer par l'utilisateur.

1.2.2 MICE

MICE (Buuren et Oudshoorn, 1999) est une méthode paramétrique qui attribue un modèle d'imputation différent pour chaque variable dont les prédicteurs sont les autres variables du système. C'est une méthode itérative, c'est-à-dire que MICE impute les données manquantes une nouvelle fois à chaque itération en se basant sur les résultats obtenus à l'itération précédente jusqu'à ce que le critère d'arrêt soit atteint. L'imputation avec MICE s'effectue en cinq étapes :

- Étape 1 : une première imputation temporaire est effectuée pour imputer les variables incomplètes du système. En général c'est une imputation par la moyenne.
- Étape 2 : les données précédemment imputées sont réinitialisées à « manquantes » pour une seule des variables du système.
- Étape 3 : les données manquantes de cette variable sont ensuite imputées à nouveau grâce aux autres variables (complètes) du système qui servent de prédicteurs à son modèle d'imputation spécifique. Ces données nouvellement imputées seront utilisées lorsque cette variable servira de prédicteur aux autres variables du système.
- Étape 4 : les étapes 2 et 3 sont répétées pour toutes les variables du système jusqu'à ce que chaque donnée manquante soit imputée. À cet instant, une itération est terminée.
- Étape 5 : les étapes 2 jusqu'à 4 sont répétées le nombre de fois précisé par l'utilisateur et les données imputées sont mises à jour à chaque itération. Les prédicteurs sont donc de plus en plus précis et les paramètres responsables de l'imputation sont de plus en plus stables.

MICE étant une méthode à imputations multiple, la procédure d'imputation est répétée *m* fois pour créer *m* bases de données imputées qui seront mises en commun après avoir été analysées

(Azur *et al.*, 2011). Par ailleurs, en raison des modèles d'imputation qu'elle propose (Shah *et al.*, 2014), elle définit les variables quantitatives seulement par l'intermédiaire de termes linéaires et sans interaction.

1.2.3 MissForest

Cette troisième méthode d'imputation est la plus récente des trois abordées (Stekhoven et Bühlmann, 2012). C'est une méthode basée sur le principe des forêts aléatoires (Breiman, 2001). Le choix de Stekhoven et Bühlmann s'est porté sur les forêts aléatoires de Breiman en raison de leur polyvalence : elles peuvent traiter les problèmes à haute dimension, contenants des données de type mixte et avec des structures complexes. Ces forêts sont une combinaison de deux idées : les arbres de décisions et le *bootstrap aggregating* – bagging (l'agrégation de modèles).

1.2.3.1 Arbres de décision

Les arbres de classification et de régression sont des modèles de prédiction (Loh, 2011). Ils sont bâtis à partir des données observables et leur structure est similaire à celle d'un organigramme. À chaque nœud de l'arbre, un test est réalisé sur une ou plusieurs variables. Chaque branche représente le résultat du test correspondant au nœud qui précède la branche et chaque feuille (extrémité de l'arbre) est un résultat possible. Pour le cas d'un arbre de régression, le résultat final est une valeur numérique et pour un arbre de classification il s'agit d'une classe. Les tests binaires effectués au niveau des nœuds ne sont pas arbitraires. Ils sont faits en sorte d'optimiser le gain d'information tiré du résultat du test (Mingers *et al.*, 2007). La Figure 1.3 illustre un exemple d'arbre de classification dans le cas d'un problème à deux dimensions. Considérant que la classe du point (2,4) est recherchée, il s'agit de parcourir l'arbre en répondant aux conditions pour trouver que la classe de ce point est « 1 ».

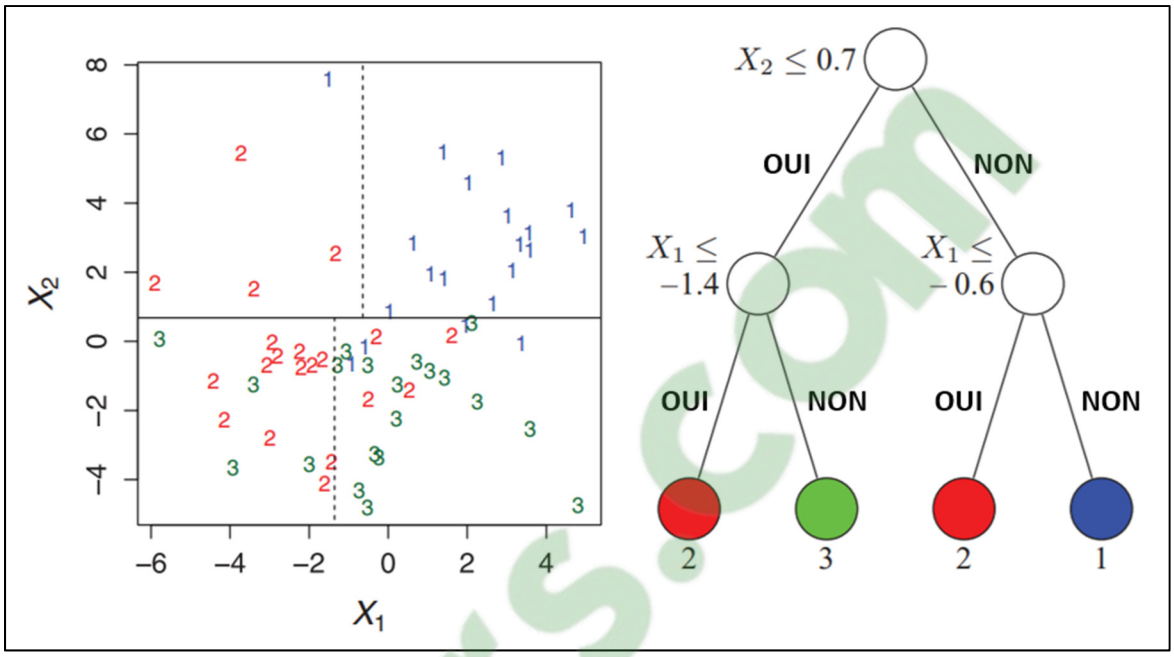

Figure 1.3 Arbre de classification pour un problème à 2 dimensions
Adaptée de Loh (2011)

1.2.3.2 Bagging

La particularité de l'approche des forêts aléatoires de Breiman repose sur le *bagging*. Au lieu de générer un seul arbre de décision à partir de l'ensemble des données observables, *t* arbres de décisions sont générés à partir de *t* sous-ensembles aléatoires de l'ensemble des données. Le résultat recherché est ensuite obtenu en effectuant une moyenne sur tous les arbres de décisions. Le fait d'agréger les arbres de décisions augmente la stabilité et la précision du modèle de prédiction (Breiman, 1996a; Domingos, 1997; Grandvalet, 2004). En effet, le bagging équilibre l'influence des points dans l'ensemble des données et réduit donc le biais qui pourrait être induit par l'un d'eux.

1.2.3.3 Algorithme de missForest

En utilisant les travaux de Breiman, Stekhoven et Bühlmann introduisent missForest en 2012. Son fonctionnement itératif n'est pas sans rappeler celui de MICE et peut le diviser en six étapes :

- Étape 1 : une première imputation temporaire est effectuée pour imputer les variables incomplètes du système. En général c'est une imputation par la moyenne.
- Étape 2 : les variables sont classées par ordre croissant du pourcentage de données manquantes.
- Étape 3 : les données précédemment imputées sont réinitialisées à « manquantes » pour une seule des variables du système.
- Étape 4 : les données manquantes de cette variable sont ensuite imputées à nouveau grâce à une forêt aléatoire bâtie à partir des autres variables du système.
- Étape 5 : les étapes 3 et 4 sont répétées pour toutes les variables du système jusqu'à ce que chaque donnée manquante soit imputée. À cet instant, une itération est terminée.
- Étape 6 : les étapes 3 jusqu'à 5 sont répétées jusqu'à ce que la différence entre la matrice dernièrement imputée et la matrice imputée à l'itération précédente cesse de diminuer.

1.2.3.4 Estimateur de l'erreur d'imputation

Une des particularités de missForest est sa capacité à fournir une estimation de l'erreur d'imputation (Breiman, 1996b). Cette estimation est appelée *Out-of-bag error* – OOB error (erreur OOB). En général, pour estimer la précision d'un modèle d'exploration de données, il est nécessaire d'effectuer un test par validation croisée (Refaeilzadeh, Tang et Liu., 2009) qui consiste à séparer l'ensemble des données observables en plusieurs sous-ensembles, les données d'apprentissage et les données de test. Le modèle est élaboré à partir de l'échantillon d'apprentissage et est validé par l'échantillon de test. Cela signifie néanmoins que la totalité des données ne sert pas à l'élaboration du modèle.

Or, l'erreur OOB n'impose pas une telle découpe de l'ensemble des données. C'est grâce au bagging que cela est possible. En effet, les arbres de décisions étant construits sur des sous-ensembles aléatoires, ils n'utilisent pas toutes les données observables simultanément. La procédure de l'erreur OOB pour évaluer la précision du modèle après son élaboration est divisée en cinq étapes :

- Étape 1 : une des données de l'ensemble des données observables est retirée. Elle est donc considérée comme manquante.
- Étape 2 : la donnée précédemment retirée est imputée à partir des arbres de décisions du modèle d'imputation qui ont été construit sur des sous-ensembles aléatoires qui ne la contenaient pas.
- Étape 3 : les étapes 1 et 2 sont répétées pour toutes les données de l'ensemble des données observables.
- Étape 4 : pour chacune des données, la différence entre la donnée observée originellement et la donnée imputée à l'étape 2 est calculée.
- Étape 5 : l'erreur d'imputation réelle est estimée en calculant la moyenne des différences obtenues à l'étape précédente.

Dans une étude sur l'estimation de l'erreur d'imputation de classificateurs agrégés, Breiman (1996b) affirme qu'en termes de précision, utiliser l'erreur OOB équivaut à utiliser un jeu de données de test de la même taille que l'ensemble des données observables.

CHAPITRE 2

MÉTHODOLOGIE

Ce deuxième chapitre aborde la méthodologie suivie dans le cadre de la présente étude. Il se subdivise en deux volets principaux. Le premier volet traite de l'approche utilisée pour analyser la performance des trois méthodes retenues soit : missForest, MICE et KNN. Dans un deuxième temps, le chapitre présente les caractéristiques de la base données des paramètres de traitement enregistrés des stations d'épuration au Québec ainsi que le protocole d'imputation qui y sera appliqué.

2.1 Analyse de la performance des méthodes d'imputation

Cette première partie présente d'une part les dix bases de données sélectionnées pour l'étude comparative et, d'autre part, les paramètres propres à chacune des trois méthodes d'imputation retenues ainsi que les indicateurs associés à l'évaluation de la performance des imputations. L'algorithme utilisé pour l'étude comparative des trois méthodes sélectionnées est présenté à la Figure 2.1 en identifiant les quatre sections présentées dans cette première partie du chapitre (2.1.1 à 2.1.4).

Tout au cours de l'analyse qui a été menée, le facteur aléatoire est constamment présent. Il a donc été décidé sur la base de la littérature (Mandel, 2015; Waljee *et al.*, 2013; Solaro *et al.*, 2017) de répéter la procédure d'imputation sur 1000 itérations, puis d'effectuer une moyenne sur les résultats obtenus.

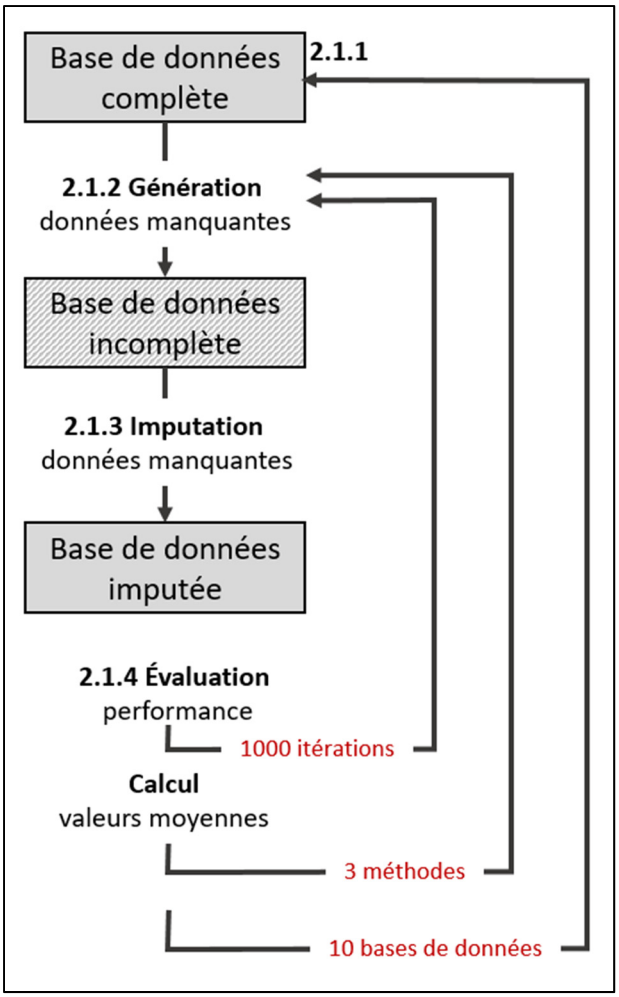

Figure 2.1 Algorithme de l'analyse de la performance des méthodes d'imputation

2.1.1 Description des bases de données

Parce que les problèmes auxquels s'intéresse la science environnementale sont divers et variés, les bases de données à traiter dans cette discipline peuvent être considérablement différentes en termes de dimension, de la nature de leurs variables et de la structure de leur distribution. Le choix des bases de données utilisées pour cette étude a donc été fait en sorte de refléter cette réalité et ainsi permettre d'évaluer les performances des trois méthodes d'imputation retenues en regard de cette diversité.

Dix bases de données de sources différentes et de cas réels ont été sélectionnées en fonction de : la nature des données (quantitatives, qualitatives, mixtes), du nombre d'enregistrements et du nombre de variables qui définissent chaque enregistrement. Le Tableau 2.1 présente les caractéristiques des différentes bases de données. Les bases de données sont décrites avec d'avantage de détails par la suite.

Deux de ces bases de données proviennent de la version de base de R, comme « Iris », qui réapparaît fréquemment dans ce type d'étude (Misztal, 2013; Stekhoven et Bühlmann, 2012; Mandel, 2015; Aljuaid et Sasi, 2017), et sept proviennent du référentiel de données UCI (Bache et Lichman, 2013), où il est possible de trouver plus de 440 bases de données. Les données des « fromageries mexicaines » ont été obtenues auprès des bases de données gouvernementales mexicaines (MEX) et sur le terrain.

Tableau 2.1 Caractéristiques des bases de données utilisées dans l'étude des performances des méthodes d'imputation de données manquantes

Bases de données	Nombre d'enregistrements (lignes)	Nombre de variables (colonnes)	Nature des variables	Source
Fromageries mexicaines	37	4	Qualitative	MEX
Hayes-Roth	132	5	Qualitative	UCI
Tic-Tac-Toe Endgame	958	10	Qualitative	UCI
Rock	48	4	Quantitative	R
Concrete Slump Test	103	10	Quantitative	UCI
Wine Quality	122	12	Quantitative	UCI
Parkinsons	195	22	Quantitative	UCI
Iris	150	5	Mixte	R
Contraceptive Method Choice	313	10	Mixte	UCI
Musk	476	167	Mixte	UCI

2.1.1.1 Données qualitatives

Fromageries mexicaines : base de données répertoriant les fromageries de la région de Xalapa (Mexique). Elle est composée de 37 fromageries qui sont caractérisées selon 4 variables qualitatives : taille de la fromagerie (4 classes), quantité d'eaux usées rejetée annuellement (9 classes), type de production fromagère (industrielle ou artisanale) et localisation (9 municipalités différentes).

Hayes-Roth : base de données contenant les résultats d'une étude sociologique effectuée sur 132 personnes (Bache et Lichman, 2013). L'objectif de cette étude est d'attribuer une classe sociale (1, 2 ou 3) à partir des informations fournies par les individus interrogés : âge (4 classes), statut conjugal (4 classes), niveau d'étude (4 classes) et loisir (3 classes).

Tic-Tac-Toe Endgame : base de données qui encode l'ensemble des 958 configurations possibles à la fin d'une partie de Morpion où « x » est le premier joueur (Bache et Lichman, 2013). Elle contient donc 9 variables représentant chacune une case et pouvant prendre trois valeurs : « x », « o » ou « b ». La dixième et dernière variable qualitative détermine si « x » a été victorieux ou non.

2.1.1.2 Données quantitatives

Rock : description de mesures effectuées sur 48 échantillons de roches issues d'un réservoir à pétrole (R Development Core Team, 2016). Les échantillons sont caractérisés selon 4 variables : périmètre (pixel), aire (pixel²), forme (rapport du périmètre sur la racine carrée de l'aire) et perméabilité (mD).

Concrete Slump Test : tests effectués sur 103 échantillons de béton afin de déterminer si leur composition peut avoir un impact sur leur affaissement (Yeh, 2007). La composition de ces échantillons est caractérisée selon 7 variables : ciment ($kg \cdot m^{-3}$), scorie ($kg \cdot m^{-3}$), cendres volantes ($kg \cdot m^{-3}$), eau, superplastifiant ($kg \cdot m^{-3}$), granulats fins et grossiers ($kg \cdot m^{-3}$). Les

variables testées sont : l'affaissement (cm), l'écoulement (cm) et la résistance à la compression sur 28 jours (MPa).

Wine Quality : mesures faites sur 122 échantillons de vin avec pour objectif de modéliser la qualité du vin à partir de tests physicochimiques (Cortez *et al.*, 2009). Les tests portent sur 12 attributs du vin : 3 liés à l'acidité ($g \cdot dm^{-3}$), 2 liés au dioxyde de soufre ($mg \cdot dm^{-3}$), quantité de sucre ($g \cdot dm^{-3}$), de sel ($g \cdot dm^{-3}$), de sulfates ($g \cdot dm^{-3}$), densité ($g \cdot cm^{-3}$), pH, teneur en alcool (%) et qualité (valeur réelle entre 1 et 10).

Parkinsons : base de données de 196 enregistrements vocaux provenant de 31 personnes, dont 23 atteintes de la maladie de Parkinson (Little *et al.*, 2007). Ces enregistrements sont caractérisés selon 22 mesures : 3 liées à la fréquence fondamentale (Hz), 8 liées aux variations de la fréquence fondamentale (%), 6 liées aux variations de l'amplitude (dB), 2 liées au ratio entre bruit et composantes tonales, 2 liées à la complexité dynamique et l'exposant de mise à l'échelle du signal fractal.

2.1.1.3 Données mixtes

Iris : base de données qui décrit la forme de 150 fleurs et les classe en trois classes différentes, soit : Setosa, Versicolour et Virginica (Fisher, 1936). La forme de ces fleurs est décrite en fonction de la largeur et de la longueur des pétales et des sépales (cm).

Contraceptive Method Choice : base de données (Bache et Lichman, 2013) décrivant la méthode contraceptive (aucune, long ou court terme) de 313 femmes en fonction de 9 caractéristiques démographiques et socioéconomiques, soit : l'âge, le nombre d'enfants, l'éducation de la femme et du mari (4 classes), le statut religieux (oui ou non), l'emploi (oui ou non), l'occupation du mari (4 classes), la qualité de vie (4 classes) et l'exposition médiatique (oui ou non).

Musk : base de données qui caractérise un ensemble de 102 molécules selon 166 variables qui décrivent quantitativement la forme et la conformation des molécules. Le but est de prédire la classe de nouvelles molécules (muscs ou non) en fonction de leurs caractéristiques (Bache et Lichman, 2013).

2.1.2 Génération de données manquantes

Afin d'analyser la performance des méthodes retenues, un manque de données a été artificiellement généré sur les dix bases de données sélectionnées avant de les confronter à missForest, MICE et KNN. Les données imputées ont par la suite été comparées aux données originelles.

Les manques de données figurant dans la suite de l'étude ont été générés de manière à ce que les données soient manquantes aléatoirement – MA. Ils ont été générés avec une fonction aléatoire présente dans le pack informatique de la méthode missForest. La comparaison des méthodes d'imputation a été effectuée pour un pourcentage de données manquantes variant de 10 à 50 % avec un pas de 10 %. Le seuil supérieur de 50 % a été retenu sur la base de tests préliminaires qui ont montré que sur certaines bases de données, les méthodes testées ont été dans l'incapacité d'imputer les données manquantes.

2.1.3 Méthodes d'imputation

Les paramètres de chacune des méthodes et la valeur associée sont présentés au Tableau 2.2. Bien que chacune des méthodes utilisées possède plusieurs paramètres pouvant faire l'objet d'un réglage spécifique, ils ont été fixés aux valeurs par défaut pour reconstruire les bases de données. Aucun test tel que ceux réalisables par validation croisée (Refaeilzadeh, Tang et Liu, 2009) n'a été effectué. Ainsi, les performances des méthodes ont été évaluées dans leur configuration de base. L'étude comparative effectuée dans cette étude ne requiert donc pas une connaissance détaillée des différents paramètres ou de compétences avancées en programmation informatique.

Tableau 2.2 Paramètres à considérer pour l'imputation de données manquantes via l'une des méthodes sélectionnées

Méthode d'imputation	Paramètre	Valeur	Remarque
missForest	maxiter	10	Limite le nombre d'itérations et permet donc de diminuer le temps de calcul lorsque l'erreur d'imputation optimale est atteinte avant le critère d'arrêt.
	ntree	100	Nombre d'arbres générés aléatoirement à chaque itération par missForest. Le diminuer a pour effet de réduire linéairement le temps de calcul au détriment de l'erreur d'imputation.
	mtry	$\sqrt{p}$	Nombre de variables choisies pour effectuer des tests binaires à chaque nœud d'un arbre de décision. Le diminuer engendre une diminution du temps de calcul lorsque le nombre de variables p est supérieur au nombre de lignes n.
MICE	m	5	Nombre d'imputations multiples. Le diminuer a pour effet de réduire le temps de calcul au détriment de l'erreur d'imputation.
	maxit	5	Limite le nombre d'itérations. Cela a pour effet de diminuer le temps de calcul au détriment de la stabilité des paramètres régissant les imputations.
KNN	K	5	Nombre de plus proches voisins pris en compte. Le diminuer a pour effet de réduire le temps de calcul au détriment de l'erreur d'imputation.

Une des particularités de MICE est qu'il est possible de choisir la méthode d'imputation associée à chaque variable imputée. Dans cette étude, toutes les variables de même type ont été imputées avec la même méthode d'imputation. La liste des méthodes utilisées, selon le type de la variable imputée est la suivante :

- *Predictive mean matching – PMM*, pour les données numériques;
- *Logistic regression imputation – logreg*, pour les données binaires;
- *Polytomous regression imputation – polyreg*, pour les données qualitatives non ordonnées;
- *Proportional odds model – polr*, pour les données qualitatives ordonnées.

Le traitement des bases de données et les calculs d'implantation des données manquante par les trois méthodes ont été réalisés avec la version 3.5.1 du logiciel R (R Development Core Team, 2016). Pour les méthodes d'implémentation, les modules suivants ont été utilisés dans R :

- « VIM » pour la méthode KNN (Kowarik et Templ, 2016);
- « missForest » (Stekhoven, 2011);
- « MICE » (Buuren et Groothuis-Oudshoorn, 2011).

2.1.4 Évaluation de la performance des méthodes

Les indicateurs d'évaluation de la performance des méthodes d'imputation sélectionnées sont associés à trois catégories. Les deux premières concernent les erreurs d'imputations soit réelles, soit estimées et la troisième catégorie porte sur la structure des bases de données.

2.1.4.1 Erreurs d'imputation réelles

Pour les erreurs d'imputation associées aux variables quantitatives, l'indicateur utilisé est le *normalized root mean squared error* – NRMSE (Oba *et al.*, 2003). Cet indicateur calcule la différence entre la base de données réelle et la base de données imputée. Sa valeur peut être supérieure à 100 %. Il s'appuie sur le principe de normalisation qui permet la comparaison entre des bases de données de tailles différentes. En considérant que la moyenne et la variance sont calculées sur les données manquantes uniquement, le NRMSE s'exprime de la manière suivante :

$$NRMSE = \sqrt{\frac{moy[(X^{Complète} - X^{Imputée})^2]}{var(X^{Complète})}} \tag{2.1}$$

Où :

$X^{Complète}$: base de données complète;

$X^{Imputée}$: base de données imputée;

moy : moyenne;

var : variance.

Pour les variables qualitatives, l'erreur est calculée par l'indicateur *proportion of falsely classified entries* – PFC (proportion de données mal classées) tel exprimé par la formule suivante :

$$PFC = \frac{nombre\ de\ données\ mal\ classées}{nombre\ de\ données\ classées} \tag{2.2}$$

Dans le cas de bases de données mixtes, ces deux indicateurs sont utilisés simultanément pour évaluer la performance des méthodes d'imputation.

2.1.4.2 Erreurs d'imputation estimées

L'estimation de la qualité d'imputation peut être estimées par la méthode de l'*Out-of-bag error* – erreur OOB fournie par la méthode missForest. Cette estimation de l'erreur prend tout son sens lorsque la base de données originale est incomplète et qu'il n'est ainsi pas possible de comparer la base de données imputée avec la base de données complète d'origine.

Dans le cas des tests sur les bases de données complètes, l'erreur d'imputation estimée selon la méthode OOB est réalisée sur neuf bases de données. Par la suite, les résultats obtenus sont comparés aux valeurs des indicateurs de l'erreur réelle précédemment calculées.

2.1.4.3 Caractérisation de la structure des bases de données

L'identification d'une relation potentielle entre la performance spécifique et générale des méthodes d'imputation et les caractéristiques de la structure des bases de données repose sur le calcul des trois indicateurs associés à la méthode des moments – *moment-based indices* (Solaro, 2015). Ces indicateurs ont été calculés à partir du coefficient de corrélation linéaire de Bravais-Pearson – ρ_{XY}. Leur coefficient de corrélation se formule selon la relation suivante :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.3}$$

Où :

X et *Y :* variables ;

$Cov(X,Y)$: covariance ;

σ_X (resp. σ_Y) : écart type de la variable *X* (resp. *Y*).

Le coefficient de corrélation linéaire de Bravais-Pearson n'est applicable qu'aux couples de variables quantitatives. Le calcul de la corrélation entre deux variables de nature différente repose sur la conversion des variables qualitatives en variables quantitatives (Zhang *et al.*, 2015). Cette conversion attribue une valeur numérique aux classes de la variable qualitative à convertir. C'est alors la moyenne des valeurs des enregistrements de la variable quantitative qui appartiennent à la classe correspondante.

Le coefficient de corrélation renseigne sur l'intensité de la liaison qui peut exister entre deux variables. Il est compris entre -1 et 1 et les valeurs seuils (Wassertheil et Cohen, 1970) à utiliser pour interpréter ce coefficient sont définies au Tableau 2.3.

Tableau 2.3 Valeurs seuils du coefficient de corrélation linéaire de Bravais-Pearson

Valeurs absolues de ρ_{XY}	**Intensité de la liaison**
Autour de 0,8	Forte
Autour de 0,5	Modérée
Autour de 0,2	Faible

Le premier indicateur est la corrélation absolue moyennée – ρ_{abs} qui correspond à la moyenne, en valeur absolue, de tous les coefficients de corrélation des variables de la base de données (équation 2.4). Il permet d'exprimer l'intensité des corrélations tant positives que négatives existantes au sein de la base de données. La corrélation absolue moyennée est exprimée par la formule suivante :

$$\rho_{abs} = \frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} |\rho_{jl}| \tag{2.4}$$

Où :

p : nombre de variables dans la base de données;

ρ_{jl} : coefficient de corrélation entre deux variables d'indice j et l.

Le deuxième indicateur correspond à l'écart-type absolu – sd_{abs} et indique s'il y a un déséquilibre au sein des corrélations (équation 2.5). L'indice est nul quand les corrélations sont tout à fait équilibrées. L'écart-type absolu se formule par la relation suivante :

$$sd_{abs} = \sqrt{\frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} \left(|\rho_{jl}| - \rho_{abs}\right)^2} \tag{2.5}$$

Le troisième indicateur correspond à l'indice absolu d'asymétrie – $skew_{abs}$ qui renseigne sur la nature de l'asymétrie au sein des corrélations (équation 2.6). Un coefficient positif implique une déviation de la distribution vers la gauche de la médiane et inversement si le coefficient est négatif. L'indice absolu d'asymétrie s'exprime de la manière suivante :

$$skew_{abs} = \frac{\frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{l>j} \left(|\rho_{jl}| - \rho_{abs}\right)^3}{sd_{abs}} \tag{2.6}$$

2.2 Cas d'application : stations d'épuration du Québec

La base de données de la performance de traitement des stations d'épuration du Québec est issue du Ministère de l'Environnement et de la Lutte contre les changements climatiques (MELCC) qui assure la cohérence et la validité des données contenues dans la base jusqu'en 2013 (https://pce.eauquebec.gouv.qc.ca). La version de la base de données utilisée comprend le suivi de 32 paramètres pour 811 stations d'épuration pour l'année 2013.

En ce qui concerne les 32 paramètres de performance, leur fréquence de mesure (fixée par le MELCC) est variable. Ainsi, dans un premier temps une moyenne annuelle a été calculée par paramètre et par station d'épuration. L'analyse préliminaire a cependant mis en évidence un pourcentage de données manquantes pouvant atteindre 90 % et plus pour certains paramètres. En fixant le seuil de données manquantes acceptable à 50 %, cela a conduit à retenir sept paramètres. De plus, l'analyse a révélé que certaines stations ne possédaient aucun des 32 paramètres. Après extraction de ces dernières, 657 stations d'épuration ont été retenues. Le tableau 2.4 présente les caractéristiques générales de la base de données originales et de la base de données résultante de la phase de prétraitement.

Tableau 2.4 Caractéristiques des bases de données originale et prétraitée

Caractéristiques	Base de données originale	Base de données prétraitée
Nombre d'enregistrements	294041	657
Nombre de paramètres	32	7
Pourcentage de données manquantes	95 %	4 %

Les 7 paramètres de performance retenus sont les suivants :

- type de traitement – TRT (9 catégories);
- demande chimique en oxygène – DCO ($mg \cdot L^{-1}$);
- demande biochimique en oxygène sur 5 jours – DBO_5 ($mg \cdot L^{-1}$);
- matière en suspension – MES ($mg \cdot L^{-1}$);
- phosphore total – Ptot ($mg \cdot L^{-1}$);
- pH (sans unité);
- Dépassement du débit de conception – DDC (sans unité).

Le dépassement du débit de conception est l'écart relatif entre le débit journalier moyen et le débit de conception. Le pourcentage de données manquantes dans cette base de données prétraitée de 4 % est principalement dû à la variable qui décrit l'enlèvement du phosphore total dans laquelle 30 % des données sont manquantes.

L’applicabilité de la méthode missForest à cette base de données prétraitée a été évaluée selon deux critères, soit : l’erreur d’imputation estimée par l’erreur OOB ainsi que le temps d’exécution.

CHAPITRE 3

RÉSULTATS

Ce troisième chapitre présente, dans un premier temps, les résultats de l'étude comparative des trois méthodes d'imputations retenues ainsi que de l'estimation de l'erreur d'imputation spécifiquement fournie par la méthode missForest. Dans un deuxième temps, le chapitre présente les résultats associés à l'application de la méthode d'imputation missForest au contexte des données de performance des stations d'épurations du Québec.

3.1 Performances comparées des trois méthodes d'imputation

Les performances de trois méthodes d'imputation (missForest, MICE et KNN) ont été comparées sur les dix bases de données complètes sélectionnées en considérant systématiquement cinq pourcentages de données manquantes générées (10, 20, 30, 40 et 50 %) et en s'appuyant sur les indicateurs : *proportion of falsely classified entries* – PFC (%), *normalized root mean squared error* – NRMSE (%), et le temps d'exécution (s). Plus spécifiquement, les performances des trois méthodes sont respectivement présentées pour : (i) les trois bases de données qualitatives (indicateurs : PFC et temps d'exécution); (ii) les quatre bases de données quantitatives (indicateurs : NRMSE et temps d'exécution) et (iii) les trois bases de données mixtes (indicateurs : PFC, NRMSE et temps d'exécution). Chacune des valeurs obtenues pour les différents cas étudiés correspond à la moyenne de 1000 simulations (génération de données manquantes suivie de l'imputation des données manquantes).

3.1.1 Bases de données qualitatives

La Figure 3.1 présente les erreurs d'imputation (exprimées par le PFC) et les temps d'exécution moyennés sur les 1000 imputations effectuées sur les trois bases de données qualitatives en fonction du pourcentage de données manquantes.

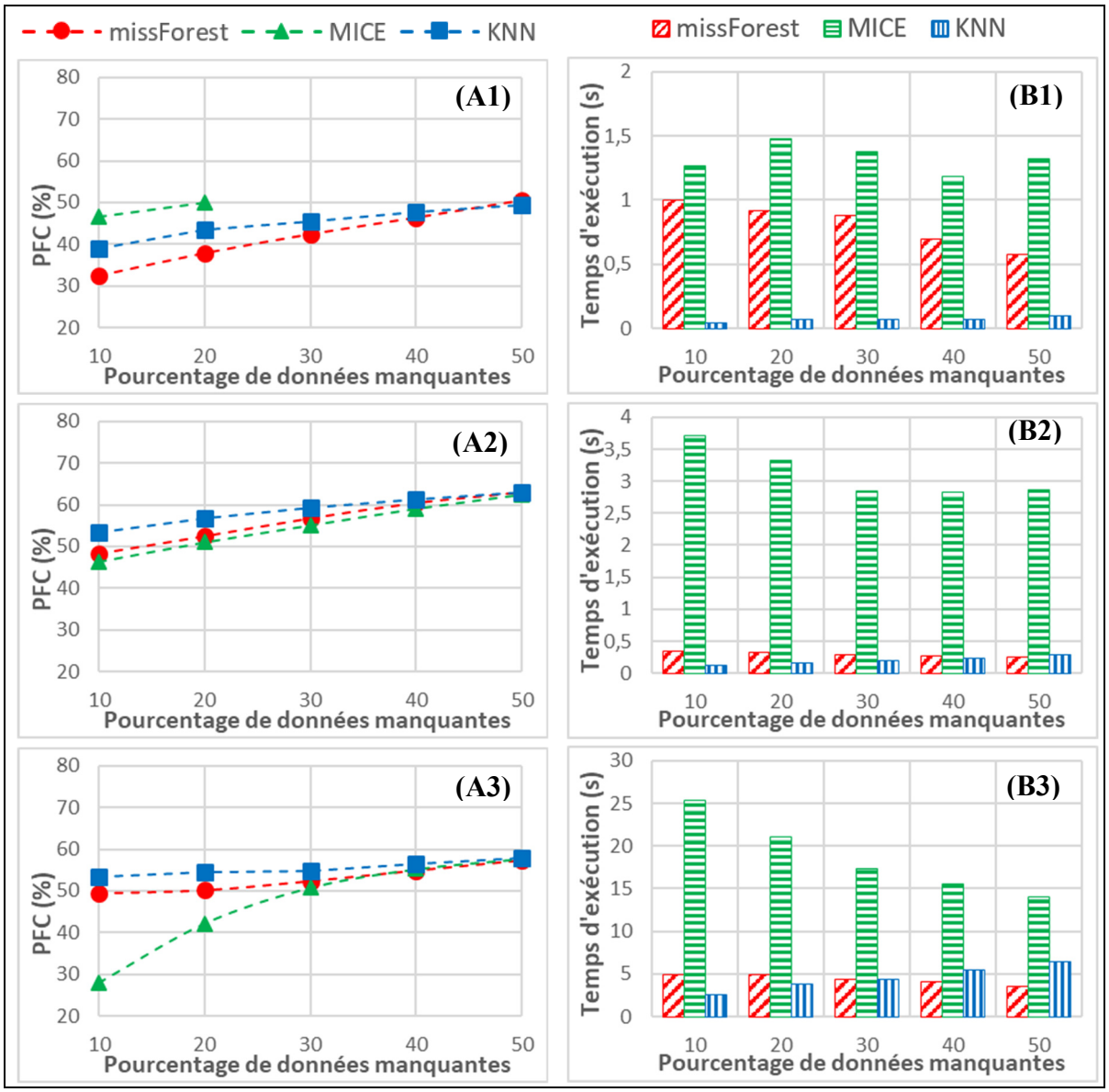

Figure 3.1 Moyenne des valeurs: (A) PFC (%) et (B) temps d'exécution (s) en fonction du pourcentage de données manquantes pour les trois méthodes d'imputation et sur les trois bases de données qualitatives : (1) « Fromageries mexicaines »; (2) « Hayes-Roth » et (3) « Tic-Tac-Toe Endgame »

Les trois graphiques (A) de la Figure 3.1 mettent en évidence que, quelle que soit la méthode d'imputation, les erreurs d'imputation (PFC) augmentent avec l'accroissement du pourcentage de données manquantes. Cette relation n'est cependant pas systématiquement observée en termes de temps d'exécution. En effet, dans le cas des trois graphiques (B) de la Figure 3.1, il ressort des comportements différents suivants les méthodes. Les temps d'exécution de KNN

augmentent systématiquement avec l'accroissement du pourcentage de données manquantes, ceux de missForest tendent à diminuer et ceux de MICE fluctuent, notamment pour le cas des « fromageries mexicaines ». Ces différences s'expliquent par les algorithmes de ces méthodes. Pour imputer une donnée manquante, KNN doit calculer la distance entre l'élément concerné et tous les autres du système afin d'identifier les *K* plus proches voisins. Donc plus il y a de données manquantes, plus KNN doit répéter ce processus. À l'inverse, plus le pourcentage de données manquantes est grand, moins missForest a de données à tester pour construire ses arbres de décision. L'élaboration du modèle d'imputation demande donc moins de temps de calcul. Quant à la nature fluctuante des temps d'exécution de MICE, elle peut s'expliquer par la nature paramétrique de son algorithme. En effet, pour imputer des données manquantes, les méthodes paramétriques définissent les variables du système comme des fonctions linéaires, l'élaboration du modèle dépend ensuite de l'estimation des paramètres de distribution de ces variables et n'est donc pas affectée par le pourcentage de données manquantes.

En considérant que les trois bases de données qualitatives de la Figure 3.1 possèdent dans l'ordre de présentation un nombre croissant d'enregistrements et de variables correspondant respectivement à 37, 132 et 958 lignes et 4, 5 et 10 colonnes, il apparaît que les erreurs d'imputation (PFC) ne semblent pas être affectées par la taille des données. Dans le cas de la méthode MICE appliquée à la troisième base de données présentée, soit « Tic-Tac-Toe Endgame » qui comporte le nombre d'enregistrements et de variables le plus élevé, l'erreur d'imputation (PFC) est significativement plus faible à bas pourcentage de données manquantes et croit de façon non linéaire avec l'augmentation du pourcentage de données manquantes. Ce comportement s'explique par la particularité de MICE qui est d'attribuer un modèle d'imputation spécifique selon le type de variable rencontré. Contrairement aux deux premières bases de données, la classification pour « Tic-Tac-Toe Endgame » porte sur une variable binaire. Le modèle d'imputation utilisé est donc différent.

Sur le plan du temps d'exécution, il apparaît que la dimension des bases de données n'affecte pas les méthodes d'imputation de la même manière. Alors que les temps associés aux méthodes KNN et MICE tendent à augmenter avec la taille de la base de données, la méthode missForest

est plus de deux fois plus rapide sur « Hayes-Roth » qu'elle ne l'est sur le cas des « fromageries mexicaines » qui comportent respectivement 132 et 37 enregistrements et 5 et 4 variables. En effet, plus il y a d'enregistrements, plus KNN a de distances à calculer. De même, plus il y a de variables, plus MICE à de paramètres de distribution à estimer. Le temps nécessaire à l'élaboration modèle de missForest dépend davantage de la structure de la base de données et de la facilité d'y identifier des schémas de données récurrents plutôt que de sa dimension.

Il ressort néanmoins quelques tendances. En raison des multiples imputations (5 dans cette étude) que MICE doit effectuer, c'est systématiquement la méthode la plus lente. C'est la méthode KNN qui est la méthode la plus rapide, notamment à bas pourcentage de données manquantes. En effet, étant donné que le temps d'exécution de la méthode missForest diminue et que celui de KNN augmente lorsque le pourcentage de données manquantes augmente, c'est missForest qui devient la méthode la plus rapide à partir d'un certain seuil du pourcentage de données manquantes pour les deux bases de données ayant le nombre d'enregistrements le plus élevé. Cependant, après comparaison des erreurs d'imputation (PFC), aucune des trois méthodes d'imputation ne semble se démarquer des autres.

À noter que dans le cas de la méthode MICE appliquée à la base de données des « fromageries mexicaines », les valeurs de PFC pour des pourcentages de données manquantes de 30, 40 et 50 % n'ont pas pu être calculées. Cette impossibilité s'explique par la particularité des méthodes paramétriques qui ne parviennent plus à effectuer une imputation lorsque le pourcentage de données manquantes augmente et qu'elles appartiennent à des variables qui sont quasiment colinéaires. C'est le cas de la base de données sur les fromageries mexicaines entre les deux variables décrivant la quantité de rejets annuels et la taille de la fromagerie.

3.1.2 Bases de données quantitatives

La Figure 3.2 présente les erreurs d'imputation (exprimées par le NRMSE) et les temps d'exécution moyennés sur les 1000 imputations effectuées sur les quatre bases de données quantitatives en fonction du pourcentage de données manquantes.

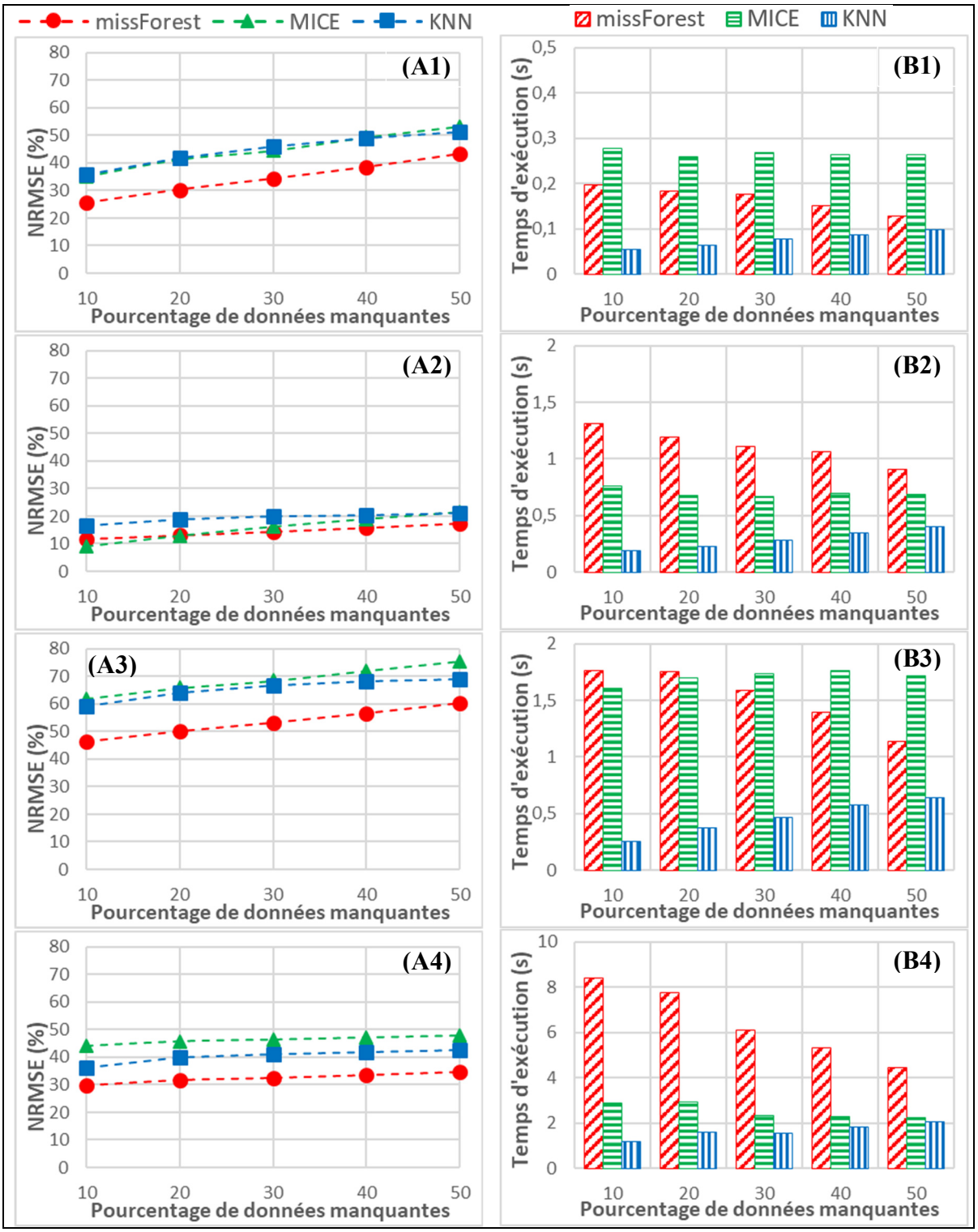

Figure 3.2 Moyenne des valeurs: (A) NRMSE (%) et (B) temps d'exécution (s) en fonction du pourcentage de données manquantes pour les trois méthodes d'imputation et sur les quatre bases de données quantitatives : (1) « Rock »; (2) « Concrete Slump Test »; (3) « Wine Quality » et (4) « Parkinsons »

À l'instar des résultats obtenus sur les données qualitatives, les quatre graphiques (A) montrent que les erreurs d'imputation (NRMSE) augmentent avec l'accroissement du pourcentage de données manquantes et ceci, pour les trois méthodes d'imputation. Les quatre graphiques (B) mettent également en avant le comportement différent des temps d'exécution des méthodes vis à vis du pourcentage de données manquantes.

La Figure 3.2 met davantage en évidence le fait que la dimension du problème traité n'a pas d'impact apparent sur l'erreur d'imputation (NRMSE). En effet, c'est sur la base de données « Wine Quality », qui possède 103 enregistrements et 12 variables, que les méthodes ont été les moins performantes. Les erreurs d'imputations sont systématiquement supérieures à celles obtenues pour « Parkinsons », qui possède pourtant le nombre d'enregistrements et de variables le plus élevé. De même, les erreurs obtenues pour les données « Rock » sont plus de deux fois supérieures à celles de « Concrete Slump Test » qui possède pourtant deux fois plus d'enregistrements et de variables. En considérant l'hétérogénéité des échelles des axes des ordonnées des graphiques (B), il apparaît que les temps d'exécution de toutes les méthodes augmentent avec la dimension des bases de données.

En ce qui concerne les performances respectives des trois méthodes d'imputation, les graphiques (B) de la Figure 3.2 révèlent que, contrairement au cas des données qualitatives et à l'exception de la base de données « Rock » pour laquelle MICE est sensiblement plus lente, missForest met généralement le plus de temps pour imputer. C'est KNN qui est systématiquement la méthode la plus rapide. Cependant, la tendance tend à s'inverser car les temps d'exécution de missForest diminuent avec l'accroissement du pourcentage de données manquantes alors que ceux de KNN augmentent. À l'inverse, les temps d'exécution de MICE restent peu affectés. En conséquence, c'est MICE qui devient la méthode la plus lente pour la base de données « Wine Quality » à partir de 30 % de données manquantes. Sur le plan des erreurs d'imputation (NRMSE), c'est ici missForest qui est la méthode la plus performante. La seule exception est pour la base de données « Concrete Slump Test » où les erreurs d'imputations fournies par les trois méthodes sont pratiquement équivalentes.

3.1.3 Bases de données mixtes

La Figure 3.3 présente les erreurs d'imputation et les temps d'exécution moyennés sur les 1000 imputations effectuées sur les trois bases de données mixtes en fonction du pourcentage de données manquantes (à l'exception du cas « Musk » pour lequel 500 imputations ont été effectuées en raison des temps d'exécution nécessaires pour imputer). Parce que ces bases de données contiennent des variables de type qualitatif et quantitatif, les deux indicateurs PFC et NRMSE sont utilisés simultanément pour évaluer la performance des méthodes d'imputation.

L'ensemble des graphiques de la Figure 3.3 montre que l'évolution des erreurs d'imputation (PFC et NRMSE) et des temps d'exécution est la même vis-à-vis de l'accroissement du pourcentage de données manquantes que pour les bases de données non mixtes. Autrement dit, les erreurs d'imputation des trois méthodes augmentent systématiquement, tandis que le comportement des temps d'exécution dépend de l'algorithme employé. L'efficacité de calcul de missForest tend à diminuer, celle de KNN augmente et celle de MICE fluctue.

En considérant que les trois bases de données mixtes de la Figure 3.3 possèdent dans l'ordre de présentation un nombre croissant d'enregistrements et de variables correspondant respectivement à 150, 313 et 476 lignes et 5, 10 et 167 colonnes, il apparaît que la dimension des bases de données ne semble pas affecter les erreurs d'imputation (PFC et NRMSE). En effet, bien que les NRMSE des méthodes MICE et KNN augmentent avec l'accroissement de la dimension des données. Ce n'est pas le cas pour la méthode missForest. De plus, sur le plan du PFC, c'est pour le cas « Contraceptive Method Choice » que les erreurs d'imputation obtenues sont les plus grandes alors que cette base de données est significativement plus petite que « Musk ». En ce qui concerne l'efficacité de calcul, les trois graphiques (B) montrent une tendance d'augmentation des temps d'exécution de chaque méthode avec l'accroissement de la dimension des bases de données. Cette augmentation est notable dans le cas de « Musk » où les temps d'exécution de l'ordre de plusieurs minutes sont observés.

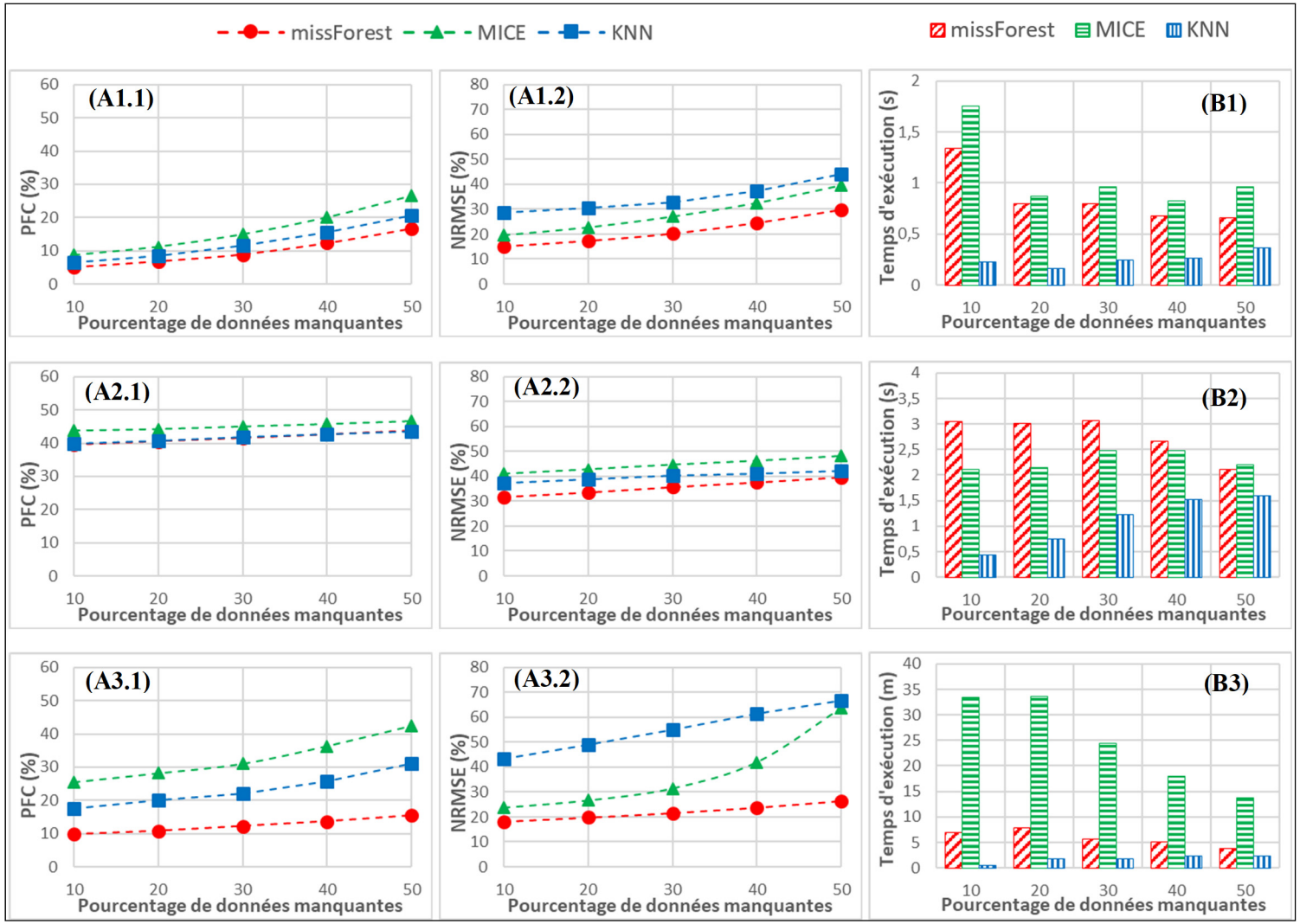

Figure 3.3 Moyenne des valeurs: (A.1) PFC (%); (A.2) NRMSE (%) et (B) temps d'exécution (s) en fonction du pourcentage de données manquantes pour les trois méthodes d'imputation et sur les trois bases de données mixtes : (1) « Iris »; (2) « Contraception Method Choice » et (3) « Musk »

La comparaison des performances des trois méthodes à partir des six graphiques (A) montre que missForest surpasse MICE et KNN dans tous les cas traités, et ceci, pour les deux indicateurs PFC et NRMSE. L'écart est toutefois moins significatif pour les données qualitatives en général notamment pour le cas « Contraceptive Method Choice » où les erreurs d'imputation sont pratiquement similaires. C'est pour la base de données « Musk » que missForest offre les meilleures performances en diminuant les PFC et NRMSE de moitié par rapport à MICE et KNN lorsque 50 % des données sont manquantes. En ce qui concerne l'efficacité de calcul, c'est KNN qui est la méthode la plus rapide malgré sa tendance à ralentir avec l'accroissement du pourcentage de données manquantes. À l'exception du cas « Musk », les temps d'exécution de missForest et MICE sont comparables.

3.1.4 Synthèse de l'étude comparative

La comparaison générale de la performance des trois méthodes d'imputation étudiées est illustrée dans la Figure 3.4. Elle présente la diminution des erreurs d'imputation de MICE et KNN par missForest en fonction du pourcentage de données manquantes moyennée sur les bases de données qualitatives, quantitatives et mixtes. Cette diminution est calculée à partir de l'écart relatif moyen entre les PFC et les NRMSE de chaque méthode.

Les quatre graphiques de la Figure 3.4 montrent que, comparativement à MICE et KNN, c'est missForest qui a généralement démontré les meilleures aptitudes à imputer des données manquantes, et ce, grâce aux principes sur lesquels repose son algorithme. En effet, contrairement à MICE, c'est une méthode non paramétrique. Cela signifie qu'elle prend en compte les interactions entre les variables au sein d'une base de données et qu'elle n'omet pas les structures de données non linéaires. L'algorithme de KNN se rapproche d'une certaine manière de celui de missForest dans le sens où c'est une méthode non paramétrique qui cherche à identifier des schémas de structures similaires afin de déterminer le résultat le plus probable. KNN ne cherche que les K enregistrements qui sont les plus proches de celui dont on veut apprendre quelque chose, tandis que missForest effectue des tests sur la totalité des données observables.

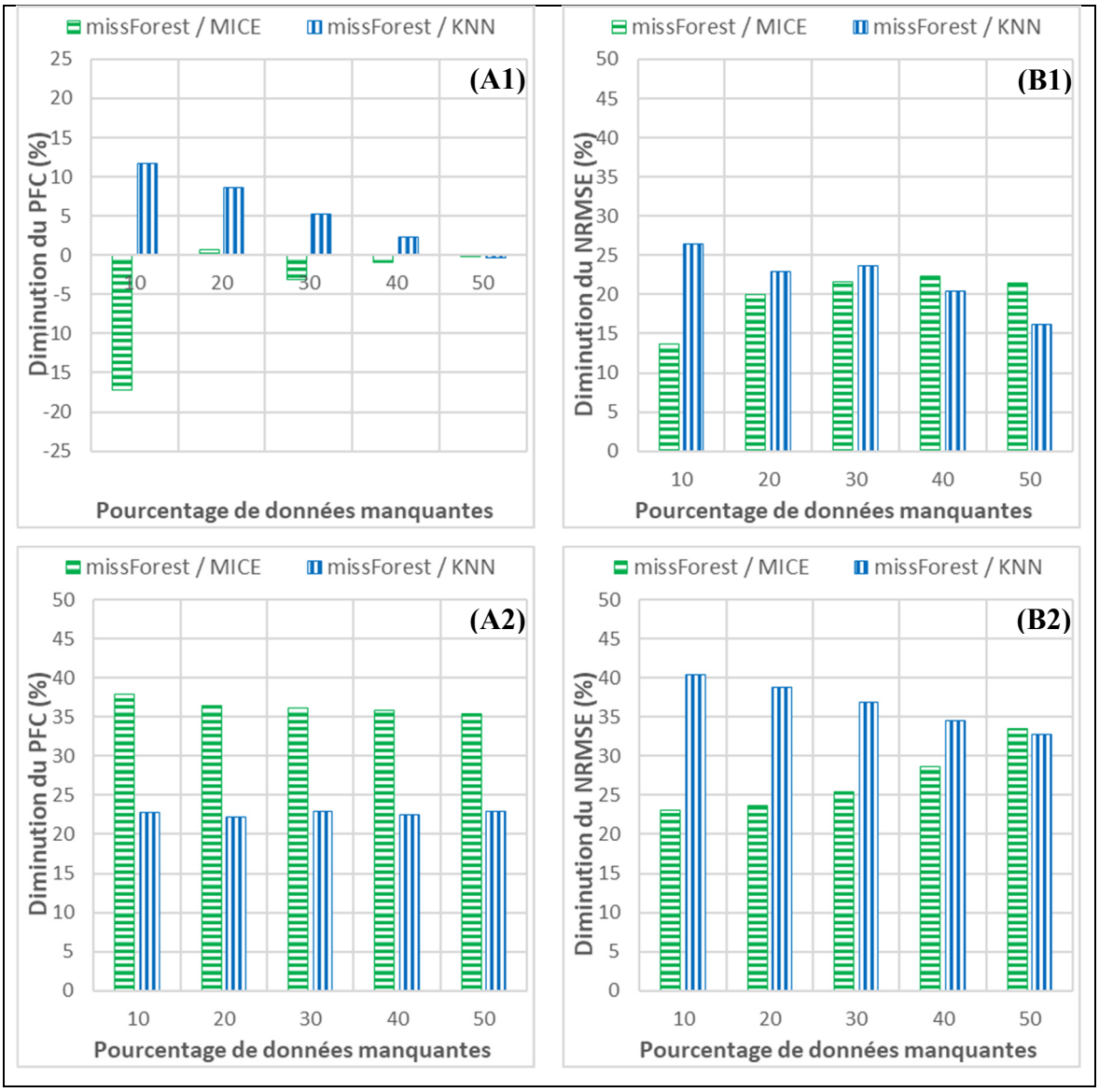

Figure 3.4 Diminution des PFC (%) et NRMSE (%) moyennée sur les données (A) de type qualitatif et (B) de type quantitatif sur les bases de données (1) non mixtes et (2) mixtes

Grâce au bagging, missForest diminue l'influence des éléments qui pourraient introduire un biais dans le modèle d'imputation. Ces différences expliquent l'écart observé entre les performances de KNN et de missForest, écart qui est d'autant plus important pour les bases de données mixtes. Cela est en partie dû au fait qu'une des faiblesses de KNN est que la qualité d'une imputation dépend de la fonction utilisée pour calculer la distance entre l'enregistrement concerné et ses voisins (Yao et Ruzzo, 2006). Cela est problématique avec à des bases de

données hétérogènes qui nécessitent un type de distance spécifique à chaque type de données qu'elle contient, ce qui n'est pas possible avec KNN.

C'est sur les bases de données mixtes que missForest a montré les meilleures différences de performances. En effet, pour ces bases de données, la diminution des erreurs d'imputation par missForest est systématiquement supérieure à 20 % et ceci, même pour les erreurs associées aux variables qualitatives (exprimées par le PFC). À l'inverse, le graphique (A1) montre que les bases de données exclusivement qualitatives ont été les seuls cas où la méthode missForest n'a pas offert la meilleure performance. Cela provient du fait que cette méthode a été développée pour traiter des problèmes hétérogènes contenant des variables de tous types. Une de ses particularités par rapport à MICE et KNN est qu'elle traite les variables quantitatives et qualitatives simultanément, d'où un écart de performance plus important avec les bases de données mixtes.

Afin d'identifier un éventuel lien entre la qualité des imputations effectuées par missForest et la complexité de la structure d'une base de données traitée, les interactions qui existent entre les variables des bases de données quantitatives ont été caractérisées. La caractérisation de ces interactions permet également d'identifier une relation potentielle entre la variabilité des erreurs d'imputations (exprimées par le NRMSE) obtenues et les caractéristiques de la structure des bases de données étudiées. En effet, au sein d'un même type de base de données, l'ensemble des méthodes d'imputations ont démontré des performances significativement différentes. Le Tableau 3.1 présente les indices de structure des bases de données quantitatives, les erreurs d'imputation moyennes qui leur sont associées ainsi que les diminutions des erreurs d'imputation de MICE et KNN par missForest. Les trois indicateurs utilisés sont : la corrélation absolue moyennée (ρ_{abs}), l'écart type absolu (sd_{abs}) et l'indice absolu d'asymétrie ($skew_{abs}$). Ces indices renseignent respectivement sur l'intensité des corrélations entre les variables, sur le déséquilibre entre les corrélations et sur le sens de ce déséquilibre.

Tableau 3.1 Indices de structures des bases de données qualitatives mis en relation avec les NRMSE moyens et la diminution des NRMSE par missForest

Base de données (lignes × colonnes)	ρ_{abs}	sd_{abs}	$skew_{abs}$	**NRMSE moyens**	**Diminution moyenne du NRMSE missForest/MICE**	**Diminution moyenne du NRMSE missForest/KNN**
Rock (48 × 4)	0,52	0,22	-0,07	41 %	24 %	24 %
Concrete Slump Test (103 × 10)	0,25	0,17	1,52	16 %	3 %	26 %
Wine Quality (122 × 12)	0,21	0,18	1,18	62 %	23 %	19 %
Parkinsons (195 × 22)	0,49	0,29	0,02	40 %	30 %	19 %

Les indices de structure présentés dans le Tableau 3.1 révèlent que l'avantage de missForest par rapport à MICE semble moins prononcé pour les bases de données dont les variables sont peu corrélées. En effet, les ρ_{abs} des bases de données « Concrete Slump Test » et « Wine Quality » sont proches de 0,2 (valeurs retenues par Wassertheil et Cohen, 1970). Leurs variables sont donc en moyenne faiblement corrélées. La corrélation est plus grande (proche de 0,5) pour les deux autres bases de données et la diminution de l'erreur atteint 30 % pour le cas « Parkinsons ». À l'inverse, l'avantage de missForest par rapport à KNN ne semble pas affecté par l'intensité des corrélations des bases de données. En effet, la diminution moyenne des NRMSE de KNN par missForest est la même entre les cas « Wine Quality » et « Parkinsons » et entre les cas « Rock » et « Concrete Slump Test ». Il n'est pas possible d'affirmer que missForest est significativement plus performante que MICE et KNN sur les bases de données dont les variables sont davantage corrélées à partir des indices utilisés. Cela peut être dû au fait que la structure des bases de données « Rock » et « Parkinsons » n'est pas suffisamment complexe pour que l'avantage de l'algorithme de missForest ait un réel impact sur les erreurs d'imputation. Par ailleurs, aucun lien n'a été fait entre les performances des méthodes et le déséquilibre des distributions des corrélations car les indices respectifs des méthodes sont du même ordre de grandeur. Il ne semble pas non plus y avoir de lien entre ces indices et la qualité des imputations fournie par les méthodes. En effet, les meilleurs et les pires

erreurs d'imputation pour les trois méthodes missForest, MICE et KNN ont été obtenus sur les bases de données « Concrete Slump Test » et « Wine Quality » (respectivement 16 et 62 % en moyenne) qui ont pourtant des indices de structures quasiment similaires.

3.2 Évaluation de la précision de l'estimateur de l'erreur d'imputation fourni par la méthode missForest

La méthode missForest fournit une estimation de l'erreur d'imputation, l'*Out-of-bag error* – erreur OOB. La précision de cet estimateur a été évaluée sur neuf des dix bases de données complètes en comparant l'erreur OOB (%) avec l'erreur d'imputation réelle (%) et ceci en considérant cinq pourcentages de données manquantes générées (10, 20, 30, 40 et 50 %). Les indicateurs utilisés sont : PFC pour les données qualitatives et NRMSE pour les données quantitatives. Chaque valeur obtenue pour les différents cas étudiés correspond à la moyenne de 1000 simulations (génération de données manquantes suivie de l'imputation des données).

La Figure 3.5 présente les courbes associées aux différences (%) entre l'erreur d'imputation réelle et l'erreur OOB en fonction du pourcentage de données manquantes pour les données (A) qualitatives et (B) quantitatives pour les bases de données (1) non mixtes et (2) mixtes. Les valeurs numériques figurant en marge de chaque point correspondent aux moyennes des erreurs d'imputations réelles.

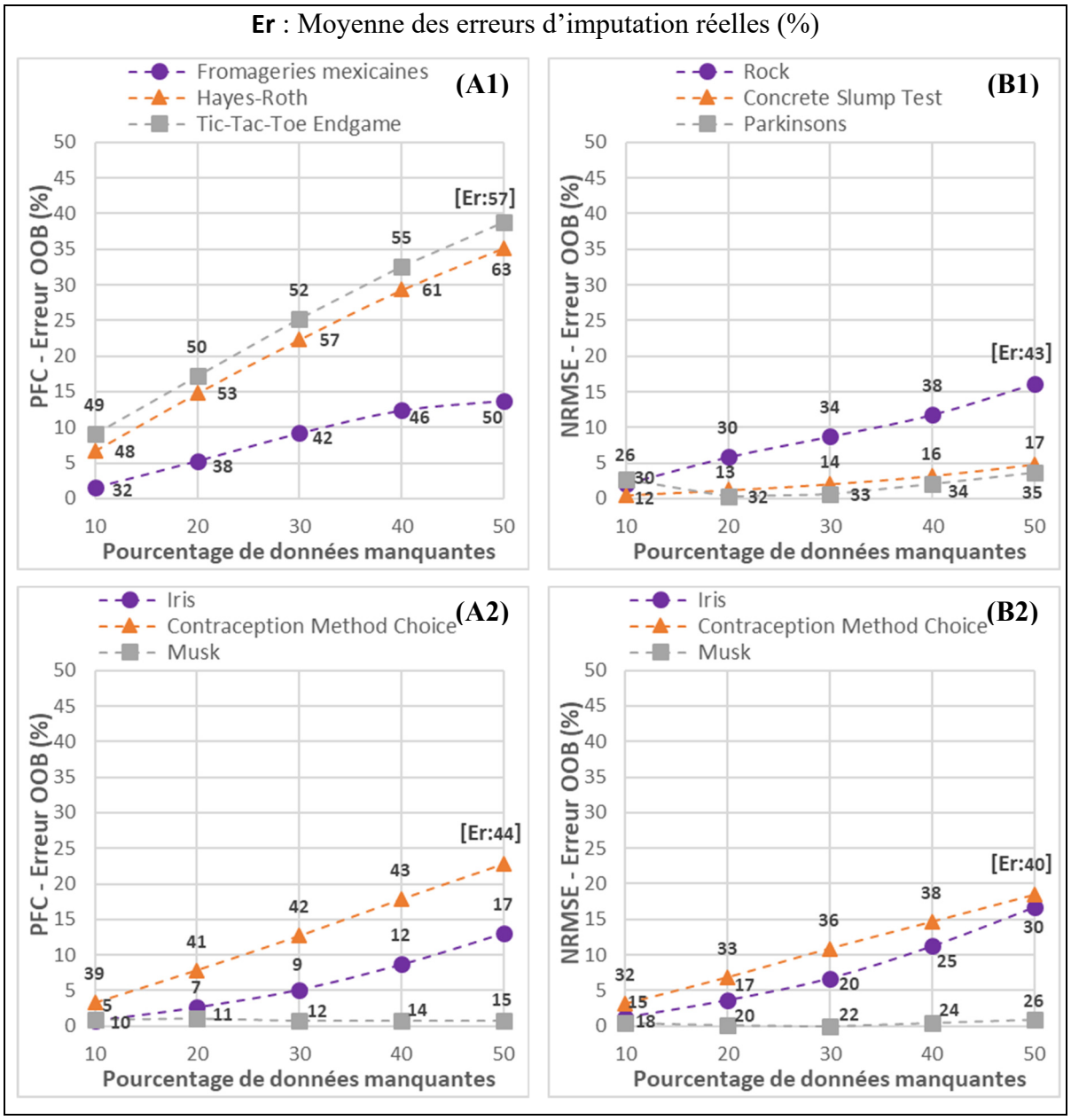

Figure 3.5 Moyenne des valeurs : (A) différences entre le PFC et l'erreur OOB (%), (B) différences entre le NRMSE et l'erreur OOB (%) et les erreurs d'imputation réelles correspondantes (%) en fonction du pourcentage de données manquantes pour les bases de données (1) non mixtes et (2) mixtes

Les deux graphiques (A) de la Figure 3.5 mettent en évidence que l'erreur OOB est plus précise pour les bases de données quantitatives qu'elle ne l'est pour les données qualitatives. En effet, à l'exception du cas des Fromageries mexicaines, la différence entre l'erreur réelle (exprimée

par le PFC) et l'erreur OOB avoisine les 10 % pour des pourcentages de données manquantes proches de 10 %. De plus, la précision de l'estimateur décroît avec le pourcentage de données manquantes, allant jusqu'à atteindre 40 %. La différence entre l'erreur OOB et l'erreur réelle est donc quatre fois plus grande pour 50 % de données manquantes que pour 10 %. À l'inverse, les différences obtenues pour les bases de données quantitatives sont proches de zéro pour de faibles pourcentages de données manquantes. En outre, la baisse de précision provoquée par l'augmentation du pourcentage de données manquantes n'est pas aussi prononcée, car à l'exception de la base de données « Rock », les différences ne dépassent pas 5 % lorsque la moitié des données sont manquantes.

Cette différence de précision de l'estimateur entre les données qualitatives et quantitatives n'apparaît pas sur les deux graphiques (2) associés aux bases de données mixtes. Cependant, tous les graphiques de la Figure 3.5 montrent que la précision de l'estimation diminue systématiquement avec l'augmentation du pourcentage de données manquantes. Cela est dû à la procédure de l'erreur OOB. En effet, la méthode missForest calcule l'erreur OOB à partir de la capacité du modèle d'imputation à retrouver les données observables qui ont été préalablement retirées une par une (Breiman, 1996b). Or, plus il y a de données manquantes, moins missForest a de données observables sur lesquelles tester son modèle d'imputation et, par conséquent, moins l'estimation de l'erreur d'imputation réelle est précise.

En considérant que, par graphique, les bases de données sont présentées par ordre croissant de dimensions, le graphique (A1) semble indiquer que la précision de l'erreur OOB diminue avec la taille des données. Elle est en effet quatre fois plus précise sur la plus petite base de données « Fromageries mexicaines », qu'elle ne l'est sur la plus grande. L'écart est toutefois moins notable entre « Hayes-Roth » et « Tic-tac-Toe Endgame » alors que cette dernière contient deux fois plus de variables et dix fois plus d'enregistrements. Cependant, le graphique (B1) invalide cette hypothèse, car l'estimation de l'erreur est la moins bonne pour la base de données la plus petite (« Rock »). Les deux graphiques (2) mettent également en évidence que la dimension des données n'a pas d'impact sur la précision de l'estimation. En effet, c'est la base de données la plus imposante « Musk » qui fournit la meilleure estimation avec des différences

stables en deçà des 2 %. De même, la dimension n'influence pas l'erreur d'imputation réelle et elle n'influence pas l'erreur OOB.

Finalement, l'ensemble des graphiques de la Figure 3.5 met en évidence le fait qu'il n'y a pas de lien apparent entre la taille de l'erreur réelle et la précision de l'erreur OOB. Pour le graphique (A1) par exemple, c'est pour la base de données « Hayes-Roth » que les erreurs réelles sont les plus grandes tandis que la précision de l'erreur OOB pour ces données se situe entre celles obtenues pour les « Fromageries mexicaines » et « Tic-Tac-Toe Endgame ». De même, pour le graphique (B1), les erreurs réelles associées au cas « Parkinsons » sont deux fois plus grandes que pour les données « Concrete Slump Test » alors que les différences obtenues pour ces deux bases de données sont pratiquement similaires. L'inexistence d'un lien entre la taille de l'erreur d'imputation réelle et la précision de l'erreur OOB s'explique à nouveau par la procédure de cette dernière. En effet, la précision du modèle d'imputation influe de la même manière sur l'erreur d'imputation réelle que sur la capacité du modèle à retrouver les données observables.

3.3 Imputation de données manquantes appliquée à la base de données des stations d'épuration du Québec

La performance de la méthode missForest a été évaluée sur un cas d'application associé aux données journalières enregistrées sur les paramètres de traitement des eaux usées des stations d'épuration du Québec pour l'année 2013. Pour rendre possible l'imputation des données manquantes, un prétraitement a été effectué sur ces données. Les caractéristiques de la base de données résultante sont décrites au Tableau 3.2.

Tableau 3.2 Caractéristiques de la base de données prétraitée des stations d'épuration du Québec pour l'année 2013

	Caractéristiques de la base de données		
Enregistrements	657 Stations		
Variables	**Nom**	**Unité / Nombre de classes**	**Pourcentage de données manquantes**
	TRT	9	0
	DCO	$mg \cdot L^{-1}$	0
	DBO_5	$mg \cdot L^{-1}$	0,15
	MES	$mg \cdot L^{-1}$	0
	Ptot	$mg \cdot L^{-1}$	21,6
	pH	sans unité	0,46
	DDC	sans unité	0
Indices de structure	$\boldsymbol{\rho_{abs}}$=0,253	$\boldsymbol{sd_{abs}}$=0,231	$\boldsymbol{skew_{abs}}$=1,06

La qualité de l'imputation de la base de données des stations d'épuration du Québec effectuée par missForest a été évaluée en utilisant deux indicateurs : le temps d'exécution (s) et l'estimation de l'erreur d'imputation réelle fournie par missForest, l'erreur OOB (%). Afin de prendre en compte l'impact de l'aléatoire sur le fonctionnement de l'algorithme de missForest, les données manquantes ont été imputées 1000 fois. Les erreurs d'imputation et temps d'exécution obtenus sont présentés à la Figure 3.6.

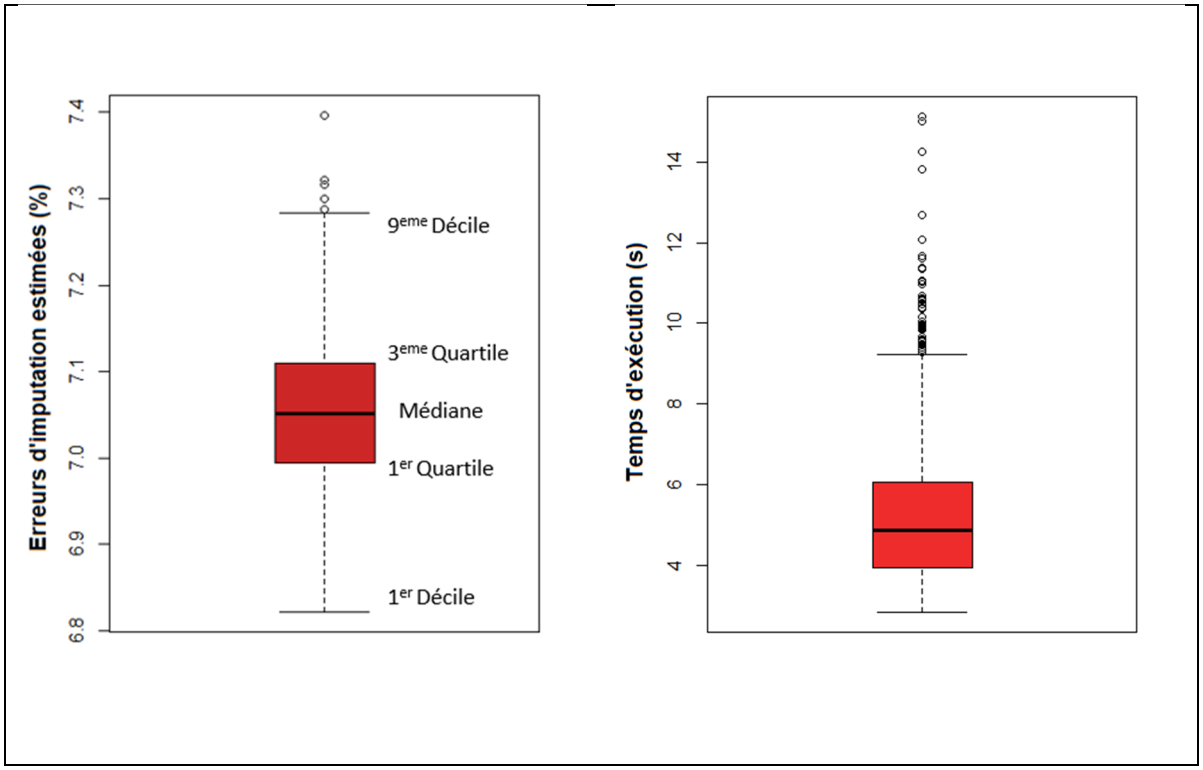

Figure 3.6 Erreurs d'imputation estimées OOB (%) et temps d'exécution (s) pour les 1000 imputations effectuées sur la base de données des stations d'épuration au Québec pour l'année 2013

La Figure 3.6 montre que le caractère aléatoire de l'élaboration du modèle d'imputation par missForest ne semble pas avoir d'impact sur l'erreur d'imputation. En effet, l'erreur d'imputation estimée varie aux alentours de 7 % (entre 6,8 et 7,4 %) avec une variance proche de zéro. Cela signifie que le nombre de sous-ensembles aléatoires utilisés pour générer les arbres de décisions (100 dans cette étude) est suffisant pour assurer la stabilité du modèle d'imputation. Bien que la variance obtenue pour les temps d'exécution soit plus élevée (3,65), ils sont systématiquement inférieurs à 15 secondes. Le temps nécessaire pour imputer les données ne pose donc aucune contrainte. Comparativement aux résultats obtenus sur la base de données « Iris » qui présente des similitudes avec celle des stations d'épuration en raison des variables de type mixte qu'elle possède (dont une seule est de type qualitatif), les erreurs d'imputation obtenues sont faibles.

CHAPITRE 4

DISCUSSION

Ce dernier chapitre traite, dans un premier temps, de la portée des résultats générés au cours de cette étude. Dans un deuxième temps, quelques perspectives et recommandations en regard de travaux de recherche futurs sont abordées.

4.1 Portées des résultats

Afin d'évaluer l'applicabilité de la méthode missForest aux données issues des problématiques environnementales, les performances de trois méthodes d'imputation (missForest, MICE et KNN) ont été comparées sur 10 bases de données complètes. Par la suite, la méthode missForest a été appliquée aux données enregistrées concernant les paramètres de traitement des stations d'épuration des eaux usées du Québec. En ce qui concerne l'évaluation comparative des trois méthodes sur les 10 bases de données, le Tableau 4.1 synthétise le classement des méthodes pour chacune des bases de données sur les deux critères retenus, soit l'erreur d'imputation exprimée par le PFC (données qualitatives) ou le NRMSE (données quantitatives) ainsi que le temps d'exécution. Le classement des méthodes est établi sur une comparaison des classements obtenus pour chacun des cinq pourcentages de données manquantes testés. Si la différence de performance entre deux méthodes est inférieure à 10 %, leur classement est identique. Les valeurs seules indiquées au Tableau 4.1 correspondent aux classements globaux tandis que les valeurs entre crochets sont les classements respectivement obtenus pour chacun des pourcentages de données manquantes soit : 10, 20, 30, 40 et 50 %.

Tableau 4.1 Bilan de l'évaluation des performances des méthodes missForest, MICE et KNN selon deux indicateurs : erreur d'imputation et temps d'exécution. Les chiffres renseignent sur la performance de la méthode d'imputation, de (1) la plus performante à (3) la moins performante

		Erreur d'imputation			**Temps d'exécution**		
	Bases de données	**missForest**	**MICE**	**KNN**	**missForest**	**MICE**	**KNN**
Qualitative	**Fromageries mexicaines**	**1** [1;1;1;1;1]	3 [3;3;3;3;3]	2 [2;2;1;1;1]	2 [2;2;2;2;2]	3 [3;3;3;3;3]	**1** [1;1;1;1;1]
	Hayes-Roth	**1** [1;1;1;1;1]	**1** [1;1;1;1;1]	**1** [1;1;1;1;1]	2 [2;2;2;2;1]	3 [3;3;3;3;3]	**1** [1;1;1;1;2]
	Tic-Tac-Toe Endgame	2 [2;2;1;1;1]	**1** [1;1;1;1;1]	2 [2;2;1;1;1]	1 [2;2;2;1;1]	3 [3;3;3;3;3]	**1** [1;1;1;2;2]
Quantitative	**Rock**	**1** [1;1;1;1;1]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	3 [3;3;3;3;3]	**1** [1;1;1;1;1]
	Concrete Slump Test	**1** [2;1;1;1;1]	2 [1;1;2;2;2]	3 [3;3;3;2;2]	3 [3;3;3;3;3]	2 [2;2;2;2;2]	**1** [1;1;1;1;1]
	Wine Quality	**1** [1;1;1;1;1]	3 [2;2;2;2;3]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	3 [2;2;2;3;3]	**1** [1;1;1;1;1]
	Parkinsons	**1** [1;1;1;1;1]	3 [3;3;3;3;3]	2 [2;2;2;2;2]	3 [3;3;3;3;3]	2 [2;2;2;2;1]	**1** [1;1;1;1;1]
Mixte	**Iris**	**1** [1;1;1;1;1]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	3 [3;3;3;3;3]	**1** [1;1;1;1;1]
	Contraceptive Method Choice	**1** [1;1;1;1;1]	2 [2;2;2;2;2]	**1** [1;1;1;1;1]	3 [3;3;3;2;2]	2 [2;2;2;2;2]	**1** [1;1;1;1;1]
	Musk	**1** [1;1;1;1;1]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	2 [2;2;2;2;2]	3 [3;3;3;3;3]	**1** [1;1;1;1;1]
	Global	**1**	2	3	2	3	**1**

Le classement des méthodes (Tableau 4.1) révèle que, sur le plan des erreurs d'imputation, missForest est généralement la méthode d'imputation la plus performante. Ce constat s'explique par les principes sur lesquels repose son algorithme, notamment par sa nature non paramétrique et par son aptitude à prendre en compte la totalité des données observables du système dans son modèle d'imputation, tout en minimisant l'influence des enregistrements susceptibles d'introduire un biais. Elle a ainsi fait preuve de robustesse face à dix bases de données différentes en termes de dimension, de la nature de leurs variables et de la complexité

de leur structure. En ce qui concerne les temps d'exécution, bien que missForest est généralement plus lente que la méthode KNN, cette tendance s'inverse car son efficacité de calcul tend à s'améliorer avec l'augmentation du pourcentage de données manquantes, la rendant en général plus rapide que MICE. Ces observations sont corroborées par les affirmations faites par plusieurs travaux récents portant sur cette problématique (Dávila, 2015; Misztal, 2013; Stekhoven et Bühlmann, 2012; Gromski *et al.*, 2014; Waljee *et al.*, 2013). Cependant, toutes les études comparatives existantes ne sont pas catégoriques quant à l'avantage de l'algorithme de missForest et à la supériorité de ses performances (Solaro *et al.*, 2017; Ghorbani et Desmarais, 2017). En effet, Ghorbani et Desmarais (2017) et Solaro *et al*. (2017) ne sont pas parvenus à départager les méthodes d'imputation que leur étude aborde malgré une compétitivité certaine de missForest. Néanmoins, ces travaux se sont limités à l'étude d'un seul type de données et, comme cela a été mis en avant par Solaro *et al*. (2017) dans l'analyse de leurs résultats, missForest constitue une méthode qui a été spécifiquement conçue pour imputer les bases de données contenant des variables de type mixte. C'est également cette particularité qui explique la diminution de la performance de missForest avec les bases de données exclusivement qualitatives (les trois premières bases de données présentées au Tableau 4.1). Malgré cette nuance, il apparaît que, comparativement à MICE et KNN, missForest est davantage apte à traiter des bases de données hétérogènes dans leurs caractéristiques et que ce ne sont pas nécessairement les méthodes d'imputation les plus utilisées qui imputent avec le plus de précision. Ce résultat est d'ailleurs en accord avec les conclusions des travaux de Celton *et al*. (2010).

En ce qui concerne les caractéristiques des bases de données à imputer pouvant potentiellement affecter la qualité des imputations, plusieurs observations ont été faites au cours de cette étude et sont résumées au Tableau 4.2. Si une relation a été observée entre la précision des imputations des trois méthodes étudiées et une des caractéristiques des bases de données imputées et ceci, pour tous les pourcentages de données manquantes testées, la mention « OUI » est indiquée. La caractéristique de corrélation entre les variables fait référence à l'indicateur de corrélation absolue moyennée – ρ_{abs} qui renseigne sur l'intensité des corrélations au sein des bases de données étudiées.

Tableau 4.2 Synthèse des relations entre les caractéristiques des bases de données imputées et la précision des imputations

Caractéristique de la base de données incomplète	Relation observée
Dimension	NON
Type de données	OUI
Pourcentage de données manquantes	OUI
Corrélation entre les variables	NON

Les résultats présentés au Tableau 4.2 mettent en évidence que trois caractéristiques sur les cinq évaluées influencent la capacité des méthodes à imputer les données manquantes d'une base de données. Le pourcentage de données manquantes et le type de données à imputer sont deux caractéristiques pour lesquelles l'effet est intuitivement compréhensible. Ainsi, une diminution systématique de la précision des imputations effectuées par les trois méthodes d'imputation sur les dix bases de données a été observée lorsque le pourcentage de données manquantes augmente. Ces résultats sont soutenus par les études comparatives citées précédemment. Pour le type de données à imputer (qualitative, quantitative, mixte), la précision des trois méthodes étudiées n'a pas été affectée de la même manière par cette caractéristique de la base de données. La méthode MICE est particulièrement impactée par le type de données en raison de son algorithme qui associe un modèle d'imputation spécifique à chaque variable selon son type. Cette particularité est également mise en avant par Azur *et al*. (2011). En revanche, la méthode KNN traite les variables quantitatives et qualitatives avec la même mesure de distance. Ainsi, la différence de performance observée n'est pas aussi importante que pour les deux autres méthodes. Bien que peu d'attention ait été portée dans la littérature aux caractéristiques structurelles des bases de données à imputer, les travaux de Solaro *et al*. (2014) semblent indiquer que d'autres caractéristiques que les deux abordées précédemment sont susceptibles d'affecter la précision d'une imputation. Ces caractéristiques sont notamment la dimension de la base de données à imputer et la complexité de la distribution de ses variables, en particulier les paramètres structurels de corrélation et de symétrie. Cependant, malgré l'utilisation des indicateurs fournis par ces auteurs pour la caractérisation

de la structure des bases de données sélectionnées (Solaro *et al.*, 2015), aucun lien n'a été observé entre l'intensité des corrélations au sein des bases de données et la précision des imputations des trois méthodes étudiées. Ceci vient contredire ce qu'affirmaient Solaro *et al*. (2015). En effet, leurs résultats ont montré une diminution générale de la précision des méthodes d'imputation lorsque la dimension augmentait et quand les coefficients de corrélation entre les variables diminuaient et que cet effet pouvait être amplifié par la présence d'une asymétrie au sein des corrélations. Cette divergence dans les résultats obtenus indique que ces deux caractéristiques ne déterminent pas à elles seules la précision d'une imputation. Par ailleurs, le fait que l'importance du paramètre de corrélation soit nuancée peut expliquer pourquoi l'avantage de missForest n'a pas été plus manifeste avec les bases de données dont les variables étaient plus corrélées.

Les résultats de l'étude démontrent que missForest constitue une méthode d'imputation robuste dont l'algorithme rend possible la reconstruction des bases de données diverses auxquelles donnent lieu les problématiques environnementales. Bien que cette affirmation soit nuancée par une diminution de performance avec des bases de données non mixtes, cette méthode a systématiquement démontré des performances compétitives comparativement à deux des méthodes d'imputation parmi les plus utilisées actuellement. De plus, le paramètre d'erreur OOB constitue un avantage certain de missForest dans la pratiue de par le fait qu'elle ne nécessite pas d'extraire des données de l'ensemble des données observables pour estimer la précision du modèle d'imputation. Ainsi, excepté les cas exclusivement qualitatifs pour lesquels la méthode MICE peut surpasser la méthode missForest, les résultats de cette étude préconisent de privilégier cette dernière si une démarche d'imputation est entreprise. Cependant, en raison de la multitude de paramètres qui peuvent affecter la performance d'une méthode d'imputation, il est délicat de s'assurer de la précision d'une imputation ou de la supériorité de missForest au travers des caractéristiques de la base de données à imputer. Par ailleurs, l'estimation des paramètres structurels d'une base de données incomplète peut donner lieu à des résultats biaisés (Leite et Beretvas, 2017).

Dans l'optique d'une mise utilisation, une approche d'optimisation possible concerne tous les paramètres à régler pour l'application des trois méthodes d'imputation qui ont été étudiées. En effet, une des conditions initiales de cette étude a été de les fixer à des valeurs « par défaut ». Cette décision a été prise afin d'étudier le comportement standard de ces méthodes et pour permettre à n'importe quel utilisateur d'appliquer la méthodologie utilisée dans ce mémoire sans que cela exige des compétences en programmation informatique. Cependant, tous ces paramètres peuvent individuellement faire l'objet d'une optimisation, ce qui pourrait modifier les résultats obtenus dans cette étude, sans nécessairement changer l'issue de leur interprétation. Par exemple, en ayant recours au paramètre « ntree » fourni par missForest, il est possible dans la pratique de diminuer significativement les temps d'exécution. En effet, en divisant par quatre le nombre d'arbres de décisions générés (soit en passant de 100 à 25 arbres) sur la base de données « Musk », l'imputation a été rendue quatre fois plus rapide et ceci n'a affecté l'erreur d'imputation que de quelques pour cent.

4.2 Perspectives et recommandations

Cette section aborde quelques pistes de recherche ayant pour but d'approfondir les connaissances apportées par la présente étude. Ces recommandations concernent la dimension des bases de données à imputer et les types de manque de données qu'elles peuvent contenir.

4.2.1 Effet seuil pour des mégadonnées

Au cours de cette étude, les bases de données ont été sélectionnées de manière à être variables en termes du nombre de lignes et de colonnes qu'elles contiennent. Leur taille est toutefois relative et les problèmes rencontrés dans la réalité peuvent être de dimension supérieure. Bien que la base de données « Musk » soit considérée comme de grande dimension (Mandel, 2015), il serait intéressant d'évaluer le traitement d'ensembles de données suffisamment grands pour nécessiter des superordinateurs. Au regard des particularités de missForest, une étude effectuée sur des bases de données de dimensions supérieures devrait corroborer les résultats obtenus au cours de cette étude. De plus, l'étude de mégadonnées pourrait mettre en avant l'existence d'un

seuil de dimension à partir duquel certaines méthodes d'imputation seraient sujettes à une diminution significative de performance.

4.2.2 Impact du type de manque de données sur la qualité d'une imputation

Tous les manques de données auxquels ont été confrontées les méthodes d'imputation au cours de l'étude comparative ont été générés de manière à ce que les données soient manquantes aléatoirement – MA. Cependant, comme cela a été mis en avant par Misztal (2013) dans ses travaux sur la méthode missForest, le type de manque de données est non seulement susceptible d'avoir un impact sur la précision générale des méthodes d'imputation, mais également sur l'avantage de l'algorithme de missForest. Il serait donc intéressant d'évaluer la performance des méthodes étudiées lorsque les données à imputer sont MDO ou MNA car ces types de données peuvent être rencontrés en pratique. L'évaluation de la précision de l'erreur OOB a donné des résultats indiquant qu'il s'agissait d'un estimateur précis face à de faibles pourcentages de données manquantes et bien que ces résultats soient corroborés par Stekhoven et Bühlmann (2012) et Breiman (1996b), la procédure de cet estimateur repose sur l'hypothèse que les données manquantes ont la même distribution que les données observables. Ainsi, dépendamment du type de manque de données, cette hypothèse peut ne pas être satisfaite. Or, il n'est pas possible d'affirmer avec certitude que les données des stations d'épuration des eaux usées du Québec soient MA. Le manque de données de ce cas d'application pourrait donc avoir un effet sur les résultats et il serait également judicieux d'étudier l'impact du type de manque de données sur la précision de l'erreur OOB.

CONCLUSION

À l'instar de la plupart des domaines d'intérêts actuels, l'environnement est confronté à un problème de manque d'information. Cela est dû au fait que, en raison de dysfonctionnements dans les processus d'acquisitions de données, les bases de données destinées à être analysées en vue d'en extraire des connaissances sont souvent incomplètes. Ce manque de données complique l'analyse des bases de données et est susceptible de biaiser la prise de décision qui s'en suit. Dans ce contexte, plusieurs approches de traitement de données manquantes ont émergé. Cependant, le domaine d'application restreint de certaines de ces approches et l'échelle grandissante des problématiques de ces dernières années a précipité l'émergence des méthodes d'imputation de données manquantes. Parmi ces méthodes figure missForest, une méthode récente, réputée pour sa robustesse face aux données de type mixte, aux systèmes de grandes dimensions et aux structures de données complexes.

L'objectif principal de la présente étude visait à évaluer l'applicabilité de la méthode d'imputation missForest à la diversité des bases de données issues des problématiques environnementales. Dix bases de données différentes en termes de dimension, de la nature de ses variables et de la structure de leur distribution ont donc été considérées afin d'évaluer la robustesse de missForest comparativement à deux des méthodes d'imputation les plus utilisées aujourd'hui : MICE et KNN. Suite à cette étude comparative, missForest a été appliquée aux données de suivi des effluents des stations d'épuration du Québec.

Les résultats de l'étude ont mis en avant que, comparativement à MICE et KNN, missForest a démontré les meilleures aptitudes à imputer des données manquantes en offrant les erreurs d'imputation les plus faibles sur la plupart des bases de données traitées avec une efficacité de calcul satisfaisante. Parmi les onze bases de données étudiées, missForest ne s'est pas révélée comme la méthode la plus performante que sur une seule d'entre elles. L'analyse de ce résultat en regard de la nature des données à imputer ont montré que la méthode missForest se démarque particulièrement lorsque les données sont mixtes, ce qui est attendu en terme de données environnementales. Par ailleurs, l'imputation des données de suivi des effluents des

stations d'épuration a été estimée à environ 7 % par l'indicateur OOB fourni par la méthode missForest dont la précision avait elle-même été préalablement évaluée.

En ce qui concerne les temps d'imputation, les résultats obtenus dans le cadre de la présente étude ont mis en évidence que la méthode KNN s'est avérée comme la méthode la plus rapide pour l'ensemble des 10 bases de données testées lorsque le pourcentage de données manquantes était inférieur ou égal à 30 %. Dans le cas de la méthode missForest, l'analyse comparative du temps de traitement montre qu'elle améliore sa performance dans le cas de données hétérogènes et avec un accroissement du nombre de données manquantes.

Dans le contexte d'une démarche d'imputation de données manquantes issues de problématiques environnementales, les résultats de cette étude montrent que, comparativement à MICE à KNN, missForest est la méthode d'imputation à privilégier. Bien qu'il n'est pas possible d'affirmer que missForest sera la méthode la plus performante indépendamment de la nature du problème rencontré, les principes sur lesquels repose son algorithme rendent possible la reconstruction de bases de données diverses et l'estimateur de l'erreur d'imputation qu'elle fournit est un avantage certain dans la pratique.

BIBLIOGRAPHIE

Afifi, A. A. et R. M. Elashoff. 1966. « Missing Observations in Multivariate Statistics I. Review of the Literature ». *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.1966.10480891>.

Aljuaid, Tahani et Sreela Sasi. 2017. « Proper imputation techniques for missing values in data sets ». *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016*. <https://doi.org/10.1109/ICDSE.2016.7823957>.

Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis et Philip J. Leaf. 2011. « Multiple imputation by chained equations: What is it and how does it work? » *International Journal of Methods in Psychiatric Research*. <https://doi.org/10.1002/mpr.329>.

Bache, K. et M. Lichman. 2013. *UCI Machine Learning Repository*. *Univ. Calif. Irvine Sch. Inf.* <https://doi.org/University of California, Irvine, School of Information and Computer Sciences>.

Baraldi, Amanda N. et Craig K. Enders. 2010. « An introduction to modern missing data analyses ». *Journal of School Psychology*, vol. 48, nº 1, p. 5-37. <https://doi.org/10.1016/j.jsp.2009.10.001>.

Batista, Gustavo E. A. P. A. et Maria Carolina Monard. 2003. « An analysis of four missing data treatment methods for supervised learning ». *Applied Artificial Intelligence*. <https://doi.org/10.1080/713827181>.

Breiman, Leo. 1996a. « Bagging predictors ». *Machine Learning*, vol. 24, nº 2, p. 123-140. <https://doi.org/10.1007/BF00058655>.

Breiman, Leo. 1996b. « Out-of-Bag Estimation ». *Technical Report*. <https://doi.org/10.1016/j.patcog.2009.05.010>.

Breiman, Leo. 2001. « Random forests ». *Machine Learning*. <https://doi.org/10.1023/A:1010933404324>.

Brock, Guy N., John R. Shaffer, Richard E. Blakesley, Meredith J. Lotz et George C. Tseng. 2008. « Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes ». *BMC Bioinformatics*, vol. 9, p. 1-12. <https://doi.org/10.1186/1471-2105-9-12>.

Buuren, Stef van et Karin Groothuis-Oudshoorn. 2011. « mice : Multivariate Imputation by Chained Equations in *R* ». *Journal of Statistical Software*, vol. 45, nº 3. <https://doi.org/10.18637/jss.v045.i03>.

Buuren, Stef van et Karin Oudshoorn. 1999. « Flexible multivariate imputation ». p. 1-20.

Bzdok, Danilo, Naomi Altman et Martin Krzywinski. 2018. « Points of Significance: Statistics versus machine learning ». *Nature Methods*, vol. 15, nº 4, p. 233-234. <https://doi.org/10.1038/nmeth.4642>.

Celton, Magalie, Alain Malpertuy, Gaelle Lelandais *et al*exandre G. de Brevern. 2010. « Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments ». *BMC Genomics*, vol. 11, nº 1, p. 15. <https://doi.org/10.1186/1471-2164-11-15>.

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos et José Reis. 2009. « Modeling wine preferences by data mining from physicochemical properties ». *Decision Support Systems*. <https://doi.org/10.1016/j.dss.2009.05.016>.

Dávila, Saylisse. 2015. « Performance of Missing Value Imputation Schemes in Women's Health Data ». p. 1-20.

Domingos, Pedro. 1997. « Why does bagging work? A Bayesian Account and its implications ». *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1.1.40.1298>.

Farhangfar, A., L. A. Kurgan et W. Pedrycz. 2007. « A Novel Framework for Imputation of Missing Values in Databases ». *IEEE Transactions on Systems Man and Cybernetics Part A Systems and Humans*. <https://doi.org/10.1109/TSMCA.2007.902631>.

Farhangfar, Alireza, Lukasz Kurgan et Jennifer Dy. 2008. « Impact of imputation of missing values on classification error for discrete data ». *Pattern Recognition*. <https://doi.org/10.1016/j.patcog.2008.05.019>.

Fisher, R. A. 1936. « Has Mendel's work been rediscovered? » *Annals of Science*. <https://doi.org/10.1080/00033793600200111>.

García-Laencina, Pedro J., José-Luis Sancho-Gómez et Aníbal R. Figueiras-Vidal. 2010. « Pattern classification with missing data: a review ». *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-009-0295-6>.

Ghorbani, Soroosh et Michel C. Desmarais. 2017. « Performance Comparison of Recent Imputation Methods for Classification Tasks over Binary Data ». *Applied Artificial Intelligence*, vol. 31, nº 1, p. 1-22. <https://doi.org/10.1080/08839514.2017.1279046>.

Gower, J. C. 1971. « A General Coefficient of Similarity and Some of Its Properties ». *Biometrics*. <https://doi.org/10.2307/2528823>.

Grandvalet, Yves. 2004. « Bagging equalizes influence ». *Machine Learning*, vol. 55, nº 3, p.

251-270. <https://doi.org/10.1023/B:MACH.0000027783.34431.42>.

Gromski, Piotr, Yun Xu, Helen Kotze, Elon Correa, David Ellis, Emily Armitage, Michael Turner et Royston Goodacre. 2014. « Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data ». *Metabolites*, vol. 4, nº 2, p. 433-452. <https://doi.org/10.3390/metabo4020433>.

Grzymala-Busse, Jerzy W. et Ming Hu. 2001. « A Comparison of Several Approaches to Missing Attribute Values in Data Mining BT - Rough sets and current trends in computing ». *Rough sets and current trends in computing*. <https://doi.org/10.1007/3-540-45554-X_46>.

Johansson, Sara, Mikael Jern et Jimmy Johansson. 2008. « Interactive quantification of categorical variables in mixed data sets ». *Proceedings of the International Conference on Information Visualisation*, p. 3-10. <https://doi.org/10.1109/IV.2008.33>.

Kowarik, Alexander et Matthias Templ. 2016. « Imputation with the *R* Package VIM ». *Journal of Statistical Software*, vol. 74, nº 7. <https://doi.org/10.18637/jss.v074.i07>.

Le, Tan Duy et Yasuo Tan. 2018. « Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare ». *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, p. 247-251.

Leite, Walter et S. Natasha Beretvas. 2017. « The Performance of Multiple Imputation for Likert-type Items with Missing Data ». *Journal of Modern Applied Statistical Methods*. <https://doi.org/10.22237/jmasm/1272686820>.

Liao, Serena G., Yan Lin, Dongwan D. Kang, Divay Chandra, Jessica Bon, Naftali Kaminski, Frank C. Sciurba et George C. Tseng. 2014. « Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? » *BMC Bioinformatics*, vol. 15, nº 1, p. 1-12. <https://doi.org/10.1186/s12859-014-0346-6>.

Little, Max A., Patrick E. McSharry, Stephen J. Roberts, Declan A.E. Costello et Irene M. Moroz. 2007. « Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection ». *BioMedical Engineering Online*. <https://doi.org/10.1186/1475-925X-6-23>.

Little, Roderick J.A. et Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. <https://doi.org/10.2307/1533221>.

Liu, Peng Liu Peng et Lei Lei Lei Lei. 2006. « Missing Data Treatment Methods and NBI Model ». *Sixth International Conference on Intelligent Systems Design and Applications*. <https://doi.org/10.1109/ISDA.2006.194>.

Loh, Wei-Yin. 2011. « Classification and regression trees ». *WIREs Data Mining Knowl*

Discov. <https://doi.org/10.1016/0169-7439(91)80113-5>.

Luengo, Julián, Salvador García et Francisco Herrera. 2012. *On the choice of the best imputation methods for missing values considering three groups of classification methods*. 77-108 p. <https://doi.org/10.1007/s10115-011-0424-2>.

Mandel, Schmitt P. 2015. « A Comparison of Six Methods for Missing Data Imputation ». *Journal of Biometrics & Biostatistics*, vol. 06, nº 01, p. 1-6. <https://doi.org/10.4172/2155-6180.1000224>.

Matsubara, Edson T., Ronaldo C. Prati, Gustavo E.A.P.A. Batista et Maria C. Monard. 2008. « Missing value imputation using a semi-supervised rank aggregation approach ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (2008). <https://doi.org/10.1007/978-3-540-88190-2-27>.

Mingers, John, Bayesian Networks, F. Faltin et R. Kenett. 2007. « Bayesian Networks ». *Machine Learning*, vol. 1, nº 1, p. 319-342. <https://doi.org/10.1002/wics.48>.

Misztal, Małgorzata. 2013. « Some remarks on the data imputation “MissForest” Method ». *Acta Universitatis Lodziensis, Folia Oeconomica*.

Oba, S., M.-A. Sato, I. Takemasa, M. Monden, K.-I. Matsubara et S. Ishii. 2003. « A Bayesian missing value estimation method for gene expression profile data ». *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btg287>.

Pigott, Therese D. 2001. « A Review of Methods for Missing Data ». *Educational Research and Evaluation*. <https://doi.org/10.1076/edre.7.4.353.8937>.

Pyle, Dorian, Senior Editor et Diane D. Cerra. 1999. *Data Preparation for Data Mining*. <https://doi.org/10.1080/713827180>.

R Development Core Team. 2016. « R: A Language and Environment for Statistical Computing ». *R Foundation for Statistical Computing Vienna Austria*. <https://doi.org/10.1038/sj.hdy.6800737>.

Refaeilzadeh, Payam, Lei Tang et Huan Liu. 2009. « “Cross-Validation.” » In *Encyclopedia of database systems*. <https://doi.org/10.1007/978-0-387-39940-9_565>.

Seaman, Shaun R., Jonathan W. Bartlett et Ian R. White. 2012. « Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods ». *BMC Medical Research Methodology*. <https://doi.org/10.1186/1471-2288-12-46>.

Sessa, Jadran et Dabeeruddin Syed. 2017. « Techniques to deal with missing data ».

International Conference on Electronic Devices, Systems, and Applications, p. 1-4. <https://doi.org/10.1109/ICEDSA.2016.7818486>.

Shah, Anoop D., Jonathan W. Bartlett, James Carpenter, Owen Nicholas et Harry Hemingway. 2014. « Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study ». *American Journal of Epidemiology*, vol. 179, nº 6, p. 764-774. <https://doi.org/10.1093/aje/kwt312>.

Solaro, Nadia, Alessandro Barbiero, Giancarlo Manzi et Pier Alda Ferrari. 2014. « Algorithmic-Type Imputation Techniques with Different Data Structures: Alternative Approaches in Comparison ». In *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*. (2014), p. 253-261. Springer International Publishing.

Solaro, Nadia, Alessandro Barbiero, Giancarlo Manzi et Pier Alda Ferrari. 2017. « A sequential distance-based approach for imputing missing data: Forward Imputation ». *Advances in Data Analysis and Classification*, vol. 11, nº 2, p. 395-414. <https://doi.org/10.1007/s11634-016-0243-0>.

Solaro, Nadia, Alessandro Barbiero, Giancarlo Manzi et Pier Alda Ferrari. 2015. *A Comprehensive Simulation Study on the Forward Imputation*. <https://ideas.repec.org/p/mil/wpdepa/2015-04.html>.

Stekhoven, Daniel J. et Peter Bühlmann. 2012. « Missforest-Non-parametric missing value imputation for mixed-type data ». *Bioinformatics*, vol. 28, nº 1, p. 112-118. <https://doi.org/10.1093/bioinformatics/btr597>.

Stekhoven, Daniel J. 2011. « Using the missForest Package ». p. 1-11.

Troyanskaya, O., M Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein et R. B. Altman. 2001. « Missing value estimation methods for DNA microarrays ». *Bioinformatics. 2001 Jun;* <https://doi.org/10.1093/bioinformatics/17.6.520>.

Vaquero, Daniel Salfrán. 2018. « Multiple Imputation for Complex Data Sets ».

Waljee, Akbar K., Ashin Mukherjee, Amit G. Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu et Peter D.R. Higgins. 2013. « Comparison of imputation methods for missing laboratory data in medicine ». *BMJ Open*, vol. 3, nº 8, p. 1-8. <https://doi.org/10.1136/bmjopen-2013-002847>.

Wang, Hai et Shouhong Wang. 2010. « Mining incomplete survey data through classification ». *Knowledge and Information Systems*. <https://doi.org/10.1007/s10115-009-0245-8>.

Wassertheil, Sylvia et Jacob Cohen. 1970. « Statistical Power Analysis for the Behavioral Sciences ». *Biometrics*. <https://doi.org/10.2307/2529115>.

Wu, Wei, Fan Jia et Craig Enders. 2015. « A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables ». *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2015.1022644>.

Yao, Zizhen et Walter L. Ruzzo. 2006. « A regression-based K nearest neighbor algorithm for gene function prediction from heterogenous data ». *BMC Bioinformatics*, vol. 7, nº SUPPL.1, p. 1-11. <https://doi.org/10.1186/1471-2105-7-s1-s11>.

Yeh, I. Cheng. 2007. « Modeling slump flow of concrete using second-order regressions and artificial neural networks ». *Cement and Concrete Composites*. <https://doi.org/10.1016/j.cemconcomp.2007.02.001>.

Zhang, Zhiyuan, Kevin T. McDonnell, Erez Zadok et Klaus Mueller. 2015. « Visual correlation analysis of numerical and categorical data on the correlation map ». *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, nº 2, p. 289-303. <https://doi.org/10.1109/TVCG.2014.2350494>.

Zhang, Zhongheng. 2016. « Missing data imputation: focusing on single imputation. » *Annals of translational medicine*, vol. 4, nº 1, p. 9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38>.