HEC MONTRÉAL

Comparaison de méthodes d'imputation des données manquantes appliquées à la base nationale sur les collisions

par

Mohamed Jabir

Science de la gestion

Option Intelligence d'affaire

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(Msc)

Août 2018

Résumé

En statistique, on parle de données manquantes (DM) quand il s'agit d'absence de valeurs pour une observation d'une variable donné. Ce sujet représente aujourd'hui un problème fréquent et a fait l'objet de nombreuses études. Il suffit d'un faible taux de DM, par variable du modèle d'intérêt, pour voir le jeu de données se réduire drastiquement. Le fait de n'utiliser que des observations complètes peut mener à une perte de précision, engendrer des biais et induire à de fausses conclusions analytiques. En effet, si on limite l'analyse uniquement aux cas complets, on risque de se retrouver avec un échantillon de données qui n'est pas vraiment représentatif de la population étudiée et par conséquent pas valide pour en déduire de l'inférence statistique.

On peut traiter les nuisances des données manquantes à l'aide de deux alternatives. La première, consiste à remplacer les données manquantes par une ou plusieurs valeurs plausibles ; dans ce cas on parle d'imputation. La deuxième option, consiste quant à elle, à utiliser le maximum de vraisemblance pour estimer les modèles de l'étude. Ces deux approches sont ce qu'on peut qualifier de « State of the Art », et sont les plus adéquates pour remédier au problème des Données Manquantes.

Les recherches dans le domaine des DM ont mis à la disposition des analystes de nombreux outils pour faire de l'imputation on peut citer plusieurs exemples :

Pour une imputation simple on peut faire appel à :

- MissMDA (Josse et Husson, 2016),
- YaImpute (Crookston and Finley-2008),
- CoImp (Di Lascio, Giannerini et Reale, 2014),
- MissForest (Stekhoven et Bühlmann, 2011).

Pour une imputation multiple on peut user de :

- Mice (van Buuren et Groothuis-Oudshoorn-2011),
- Hmisc (Harrell Jr et Dupont, 2015),
- Mi (Su *et al.*, 2011).

Ces méthodes sont aujourd'hui fréquemment utilisées, et font leurs preuves en donnant d'excellents résultats dans de nombreuses applications.

La base de données dont nous disposons est multidimensionnelle et contient à la fois des variables continues, binaires et multinomiales. La vérification de la performance des méthodes

d’imputation, sur des données réelles, est une opération délicate à faire avec le plus grand soin, puisque les valeurs que peuvent prendre les données manquantes sont inexistantes, ainsi, on ne pourra pas faire de comparaison avec les valeurs plausibles générées par le processus d’imputation. Afin de contourner ce problème nous avons éliminé les observations avec des données manquantes et restreindre l’échantillon de départ aux cas complets. Nous avons, par la suite, introduit artificiellement des taux de données manquantes, selon la typologie MAR. Nous disposons alors de jeux de données réels avec aucune donnée manquante et d’autres qui en contiennent à des taux bien précis.

Après avoir testé quelques méthodes d’imputation et choisi quelques-unes parmi les mieux adaptées à la nature de notre base de données. Nous avons ensuite procédé à de multiples diagnostics graphiques et numériques dans le but de valider la qualité des jeux de données imputés en comparant les valeurs de remplacement aux valeurs initiales. Nous avons aussi étudié les conséquences de l’imputation sur un modèle d’analyse et utilisé un modèle proche du modèle Fredette *et al.* (2008) à cet effet.

Mots-clés : données manquantes, imputation multiple, imputation simple, modèles d’imputation, sélection de variables, données mixtes, diagnostic post-imputation

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

DM	Données manquantes
FCS	Fully Conditional Specification
IM	Imputation multiple
MAR	Missing At Random
MCAR	Missing Competly At Random
MNAR	Missing Not At Random
KNN	K-Nearest Neighbor
NCDB	National Collision Database

À mes enfants adorés, Yassine, Fatima et Ziad pour l'espoir que vous représentez.

Remerciements

Je souhaite avant tout adresser mes profonds remerciements et mes sincères reconnaissances à mon Directeur de mémoire François Bellavance pour l'orientation, la confiance, le soutien, la disponibilité et la patience qui ont constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené à bon port.

Je tiens aussi à remercier ma famille et mes amis qui par leurs encouragements, j'ai pu m'accrocher et surmonter tous les obstacles.

Introduction

Les données manquantes (DM) sont un problème général qui affecte les analyses statistiques. Ainsi, la présence de données manquantes est plus souvent la règle qu'une exception. L'abondance de la littérature de 1930 à nos jours traduit l'importance de ce sujet.

De nombreuses causes sont à l'origine des données manquantes. Dans le cas d'une enquête d'opinion, il est probable qu'une personne faisant partie de l'échantillon refuse tout simplement de répondre à certaines questions. Il s'agit de *non-réponse partielle*. La *non-réponse totale* caractérise une situation où le questionnaire complet est manquant; ou encore lorsque le sujet refuse de participer à l'enquête. D'autre part, une mauvaise retranscription de l'information générera des données aberrantes qu'il convient de considérer et de traiter comme des données manquantes. Une très bonne connaissance des causes qui ont conduit à l'absence de données est nécessaire avant de prendre toute action pour remédier à leur nuisance.

Selon Little et Rubin (2016) , les méthodes statistiques standards ont été développées dans le but d'analyser des données rectangulaires. Les lignes représentent des unités, observations ou sujets, selon le contexte, et les colonnes représentant les variables mesurées. De plus, la plupart des logiciels statistiques excluent, par défaut, les observations qui présentent des données manquantes pour les variables considérées dans l'analyse. Cette méthode communément appelée « Complete-Case Analysis » est généralement inappropriée, puisque l'analyste cherche à avoir l'inférence statistique relative à toute la population et non pas à la portion des observations qui ne présentent aucune donnée manquante. Elle séduit par sa simplicité, mais réduit de façon drastique la taille de l'échantillon. Dans leur ouvrage « *Statistical Analysis with missing data* » Little et Rubin (2016) estiment que, pour un ensemble de données de 20 variables dont

10% des valeurs sont manquantes pour chaque variable, la portion de données complètes qui sera conservée par cette technique ne représente que 13% de l'ensemble des données.

Durant les dernières années, de nombreuses techniques ont été développées pour tenter d'inclure les observations qui présentent des valeurs manquantes. L'imputation multiple (IM) et la méthode du maximum de vraisemblance (ML) représentent les deux méthodes les plus élaborées pour remédier au problème des données manquantes. Les autres méthodes, que nous allons décrire plus loin en détail, notamment les méthodes « parwise deletion » qui afin d'éviter de supprimer trop de données font des suppressions par paires, imputation par la moyenne, et la prédiction à l'aide de régressions ont prouvé leurs limites sauf dans des conditions très restrictives.

La décision de remplacer ou non les valeurs manquantes est une question délicate. Par exemple, lors d'une enquête, un sujet peut refuser de révéler son revenu ou un équipement technique de mesure peut tomber en panne. Dans ces cas, il est naturel d'essayer d'imputer ces valeurs qui auraient dû être renseignées si l'équipement n'était pas défaillant et si les méthodes d'enquêtes étaient plus appropriées. Dans le cas où un sujet serait incapable d'exprimer son opinion, il est moins naturel de traiter ce cas comme une donnée manquante et d'essayer de l'imputer puisque la non-réponse apporte une information additionnelle qu'il faudra considérer. Il est plus approprié dans ce cas de créer deux strates en utilisant une variable indicatrice de données manquantes (Little et Rubin, 2016). Par ailleurs, même si la méthode d'imputation choisie peut préserver la distribution des données, il devient erroné d'analyser les données qui ont été complétées en remplaçant les données manquantes par des mesures estimées comme si elles étaient des données complètes . Les données manquantes représentent un défi à l'analyste qu'il devra traiter afin de préserver le nombre d'observations disponibles et d'obtenir des résultats fiables.

Le but de ce travail est de comparer les méthodes existantes qui prennent en considération les données manquantes et de les valider empiriquement sur une base de données réelle. Nous allons utiliser des données qui présentent un taux assez élevé de valeurs manquantes pour plusieurs variables. Pour atteindre cet objectif, nous allons passer en revue les méthodes d'imputation disponibles à ce jour puis les comparer pour désigner celles qui performent le mieux dans notre contexte. Nous allons dédier un premier chapitre de ce mémoire à présenter ces méthodes de traitement des données manquantes en relevant les avantages et les inconvénients de chaque méthode. Un deuxième chapitre sera consacré au fondement théorique de l'imputation multiple. L'imputation multiple est recommandée pour traiter les données manquantes. De plus son utilisation comporte de nombreuses hypothèses et règles notamment pour regrouper les résultats obtenus. Nous avons jugé bon de lui consacrer un chapitre au complet. Nous allons décrire la méthodologie et l'approche utilisé pour imputer les données dans le chapitre 3. Nous allons enchainer et dédier le chapitre 4 à la présentation des résultats. Nous allons conclure et partager avec le lecteur les enseignements acquis de cette étude. Nous disposons d'une base de données sur les accidents de la route de 2000 à 2008. La NCDB (National Collision Database) contient l'ensemble des accidents de la route qui ont été rapportés à Transport Canada par les provinces. Cette base comporte de nombreuses données manquantes.

Chapitre 1 : Revue de la littérature

Dans ce chapitre, nous passerons en revue les techniques de traitement des données manquantes. Nous commencerons par un survol historique des développements qu'a connu ce domaine de recherche. Nous présenterons ensuite les principales méthodes existantes, tout en relevant les avantages et inconvénients de chaque technique.

1.1) Historique des techniques de traitement des données manquantes

Bien que les chercheurs aient étudié le problème des données manquantes depuis presqu'un siècle, les percées majeures n'ont été réalisées que durant les années 1970 (Enders, 2010). En effet, c'est à Allan et Wishart (1930) qu'on doit la première publication qui traite du problème des données manquantes. À l'époque, les publications faisant référence à l'imputation des données manquantes étaient très rares et l'utilisation de l'analyse complète, qui ignore les données manquantes, était la technique la plus répandue. Allan et Wishart (1930) ont présenté une méthode pour imputer une seule valeur manquante. Elle consiste à estimer la valeur manquante en minimisant la somme des moindres carrés. Quelques années plus tard Yates (1933), démontre que la méthode peut être étendue au cas de plusieurs valeurs manquantes. Kramer et Glass (1960) suggèrent par la suite de procéder de manière itérative.

Le premier article théorique important traitant du problème des données manquantes date de 1932. Wilks (1932) le consacre à l'estimation des paramètres des échantillons incomplets constitués de données normales bivariées. La méthode qu'il propose vise à estimer simultanément les moyennes, les variances et la corrélation par la méthode du maximum de vraisemblance. Matthai (1951)

l'étendit au cas général d'une distribution normale multivariée et publia les résultats pour le cas trivarié.

Dear (1959), proposa d'utiliser le modèle d'analyse en composante principale pour imputer les données manquantes. Selon ce même auteur, cette méthode est très sensible au nombre de composantes principales retenues.

C'est à Buck (1960) qu'on doit l'idée d'estimer la valeur manquante d'un individu à partir d'autres informations disponibles pour ce même individu. L'approche préconisée était d'imputer la donnée manquante par une fonction des autres valeurs présentes. Il démontra qu'après avoir imputé les valeurs manquantes à l'aide d'un modèle de régression, l'estimation de la covariance est non biaisée, cependant la variance doit être corrigée.

Par ailleurs, jusqu'en 1976, les auteurs se contentaient d'hypothèses floues sur la nature des données manquantes. Il a fallu attendre l'article de Rubin (1976) pour disposer d'un modèle probabiliste formel des données manquantes en général. Les concepts apportés par et Little et Rubin (1987) sont très utilisés de nos jours par les chercheurs, formant ainsi les principes de base de la théorie sur les données manquantes. Dans ce qui suit, nous allons les aborder en détails étant donné leur importance dans la recherche sur les données manquantes.

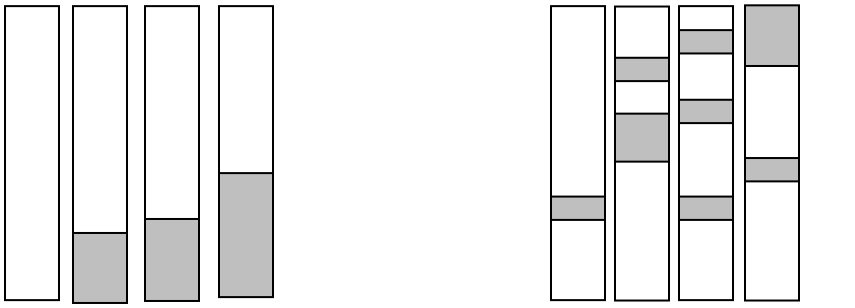
1.2) Configuration des données manquantes

Il est important de comprendre la nature des données manquantes afin de définir la méthode adéquate qui sera utilisée pour résoudre le problème. Selon Kenward et Molenberghs (1998) , pour inclure les données incomplètes dans le processus de modélisation, nous devons réfléchir sur la nature du mécanisme des données manquantes et ses conséquences sur l'inférence statistique.

Il convient de distinguer deux caractéristiques qui permettent de décrire la nature des données : l'une est appelée configuration « *pattern* », et l'autre a trait au

processus ou mécanisme à l'origine de l'apparition des données manquantes. La première caractéristique désigne l'emplacement des données manquantes tandis que le mécanisme des données manquantes décrit la relation entre les variables mesurées et la probabilité de non-réponse. Ce dernier joue un rôle essentiel dans la théorie de Rubin (1976) sur les données manquantes. On parle de configuration *monotone* quand, en données longitudinales par exemple, un individu quitte la cohorte qui est suivie. Selon Schafer (1997) ce type de configuration a reçu l'attention des chercheurs, car il réduit la complexité mathématique du maximum de vraisemblance. La configuration générale *non monotone* est le schéma des données manquantes le plus répandu. Dans ce cas particulier, les vides laissés par les DM sont répartis de façon non organisée ou aléatoire. Ces deux configurations sont illustrées par la figure 1 qui est inspiré d'Enders (2010). Historiquement, les chercheurs avaient développé des techniques qui étaient dédiées à un type particulier de DM (Enders, 2010).

Figure 1 : Représentation schématique des configurations de DM



Configuration monotone Configuration générale

1.3) Classification des données manquantes

Rubin (1976) a été le premier à étudier les mécanismes derrière la présence des DM. Le concept théorique apporté par Rubin implique deux types de paramètres; les paramètres à estimer et les paramètres décrivant la probabilité des DM. Les travaux de Rubin (1976) sont importants parce qu'ils clarifient les conditions

requises pour estimer les paramètres sans avoir à connaitre les paramètres relatifs à la distribution des DM (Enders, 2010).

1.3.1) Concepts fondamentaux

Nous reprenons dans cette partie la théorie de base développée par Little (1987). Pour mieux la comprendre, nous devons adopter quelques notations et terminologies. Pour commencer, nous allons considérer que l'ensemble des données est composé de deux parties une portion de données observées qui sera notée Y^{Obs} et une portion manquante qui représente les non-réponses et sera notée Y^{M}. Ce partitionnement des données en deux joue un rôle central dans la théorie de Rubin (1976). Ce dernier a défini une variable binaire R qui prend la valeur 0 si une mesure est observée pour la variable et prend la valeur 1 si aucune valeur n'est observée. Ainsi, pour chaque individu, nous aurons une mesure observée ou non, et une variable indicatrice de DM qui est R d'où l'existence d'une loi de probabilité qui détermine la valeur prise par R. Cette probabilité peut être reliée ou non à d'autres variables de notre ensemble de données. Dans la théorie de (Rubin, 1976), c'est cette relation entre R et les autres variables qui permettent de différencier les processus des données manquantes.

1.3.2) Taxonomie des données manquantes

Rubin (1976) et Little et Rubin (2002) ont établi une méthode de classification qui est largement reprise dans la littérature statistique traitant du problème des données manquantes. Nous allons l'exposer dans ce qui suit pour comprendre les processus qui sont derrière les données manquantes.

1.3.2.1) Données manquantes au hasard (MAR : *missing at random)*

Le processus des données manquantes est dit *MAR* si la probabilité que la valeur de Y soit manquante est reliée à d'autres mesures de notre ensemble des données, mais cette probabilité ne dépend pas de la valeur de Y (c'est-à-dire la valeur hypothétique qu'elle aurait dû prendre si elle n'était pas manquante). En d'autres termes, il n y a aucune relation entre la probabilité que Y soit manquante et les valeurs que prendrait Y. Cependant, une relation existe entre quelques variables et la probabilité que Y soit manquante. A titre d'exemple supposons qu'une école donnée a conduit un test d'aptitude en Math et que les étudiants qui ont un certain niveau vont suivre un cours avancé. Les DM des notes de ce cours avancé vont être MAR car elles sont déterminées par les notes du test d'aptitude.

Malgré son nom qui prête à confusion, le processus MAR ne signifie pas que les données constituent un simple échantillon aléatoire de toutes les données Schafer (1997). Le type MAR est le moins restrictif des processus, car il requiert seulement que les données constituent un échantillon représentatif de l'ensemble des sous-classes définies par les données observées (Schafer, 1997). Il n'existe aucun moyen de confirmer que le processus est MAR. La probabilité que Y soit manquante peut dépendre de $Y^{Obs,}$ mais ne dépend pas de Y^{M}, d'où la représentation suivante

$$P(R \mid Y^{Obs}, \Phi)$$

Où l'ensemble des paramètres qui décrit la relation entre R et les données est représenté par Φ.

1.3.2.2) DM complétement au hasard (MCAR) missing completely at random

Dans le cas où la probabilité que la valeur de Y soit manquante est indépendante des valeurs de Y et des autres variables mesurées on parle de processus MCAR. A titre d'exemple considérant une cohorte d'élève d'une école qui sont suivies pour les besoins d'une étude donnée; si un élève change d'école ou de zone scolaire durant ce suivi sans que ce changement soit relié aux autres variables collectées (Statut sociodémographiques, problème disciplinaire, …), alors on est en présence

de DM complétement au hasard. Dans ce cas précis, les données observées représentent un échantillon aléatoire de la population (Enders, 2010). Dans la pratique, il est possible de vérifier si la probabilité est MCAR à l'aide de tests de moyennes. Selon Little et Rubin (2002) la condition MCAR s'applique s'il n'y a aucune différence entre les individus ayant répondu et ceux qui ont refusé de répondre. La méthode la plus simple pour tester MCAR est d'utiliser une série de tests de moyennes pour comparer les sous-groupes qui ont des DM avec ceux qui n'en présentent pas. Cette vérification peut se faire en séparant les observations dont les données sont complètes Y^{Obs} et les observations qui ont des DM en deux groupes puis réaliser des tests d'égalité de moyenne sur les autres variables qui présentent des données.

De plus, Y^{Obs} et Y^{M} n'ont aucune relation avec R et la distribution de la probabilité que Y soit manquante est représentée par :

$$P(R \mid \Phi)$$

L'équation ci-dessus signifie que des paramètres continuent à avoir de l'influence sur les valeurs que prend R, mais cela n'a aucune relation avec le processus qui génère les DM.

1.3.2.3) Données manquantes non au hasard (MNAR : *missing not at random*)

Les données sont dites manquantes non au hasard MNAR si la probabilité que la valeur de Y soit manquante dépend des valeurs prises par Y. Il n'existe aucun test pour vérifier si le processus qui génère les données manquantes est MNAR.

Afin d'illustrer ça, supposons qu'une firme a embauché 20 nouveaux employés qui seront suivis selon leurs performances pendant une période de six mois, puis ils seront notés et évalués à la fin des six mois. Les responsables de cette firme ont jugé qu'ils ne peuvent pas attendre la fin des six mois, et jugent que certains employés doivent être licenciés à cause de leur trop faible score. À la fin de la période des six mois, l'évaluation finale de ces employés qui ont quitté l'emploi

est manquante et la probabilité qui a induit ces DM dépend de la valeur de la performance elle-même. C'est un cas typique d'un scénario de DM qui sont MNAR.

1.4) Méthodes de traitement des données manquantes

Jusqu'à présent, plusieurs techniques ont été développées pour résoudre le problème généré par les DM. Certaines méthodes ignorent toutes les observations avec une ou plusieurs DM. Cette approche, qui conduit à une perte d'information et potentiellement à l'introduction de biais dans les paramètres estimés, est celle adoptée par défaut dans les logiciels statistiques. Elle est inappropriée puisque l'analyste cherche à estimer les paramètres pour l'ensemble des données et non pas seulement pour la portion ne présentant pas de donnée manquante (Rubin, 1976). D'autres stratégies, dites inclusives, tiennent compte des données manquantes et les incluent dans le processus d'estimation des paramètres. Les techniques d'imputation produisent des résultats valides sans compliquer les analyses une fois que l'imputation a été conduite (Donders *et al.*, 2006). Nous allons décrire ces différentes méthodes avec leurs forces et faiblesses.

1.4.1) Méthodes qui ignorent les observations avec DM

1.4.1.1) Analyse des cas complets seulement:

Cette méthode, utilisée par défaut par les logiciels statistiques, ignore les observations qui contiennent des données manquantes. En plus de la perte d'information, cette approche risque aussi d'introduire un biais dans les paramètres estimés. Dans le cas où plusieurs variables sont incluses dans le modèle, cette méthode a pour conséquence de réduire énormément le nombre d'observations disponibles. Selon (Schafer et Graham, 2002) , il n'est pas exclu de trouver des publications qui font référence à des analyses où jusqu'à 73% des observations ont été exclues à cause des données manquantes. La méthode

« Complete Case Analysis » est non seulement inefficace, mais réduit aussi l'analyse à une portion de données qui n'est plus représentative de la population à cause de l'éventuelle différence entre les observations qui présentent des données manquantes et celles qui sont complètes. Par conséquent, les paramètres estimés seront potentiellement biaisés à cause des non-réponses (Little et Rubin, 2002).

Si le processus générateur des données manquantes est MCAR, les observations complètes constituent effectivement un sous-échantillon aléatoire de l'ensemble des données. Dans ce cas précis, en ignorant les données incomplètes, les estimations ne seront pas biaisées. Dans certains cas précis où le pourcentage de données manquantes n'excède pas 5% et que le processus générateur des données manquantes est MCAR, l'approche « Complete Case Analysis » peut être acceptable (Graham, 2009; Schafer, 1997). En réalité plus souvent qu'autrement, les données complètes diffèrent largement des données de l'échantillon original (Little et Rubin, 1987).

Malgré le fait que l'analyse des cas complets réduit énormément le nombre d'observations et peut conduire à des estimations biaisées, elle séduit par sa simplicité et reste très utilisée. Selon Peugh et Enders (2004), la méthode du « Complete Case Analysis » est la méthode la plus répandue pour traiter les données manquantes. Par rapport à d'autres approches , elle a l'avantage d'utiliser le même jeu de données pour calculer par exemple la moyenne de X et la moyenne de Y car ce sont les observations qui n'ont pas des DM à la fois pour X et Y qui seront utilisées. Cependant, ses inconvénients dépassent les bénéfices que l'analyste peut en tirer. D'après Little et Rubin (2002) cette technique, très implantée dans les logiciels statistiques, n'est pas recommandée; sauf dans le cas où la portion de données manquantes est très faible.

1.4.2) Méthodes qui tiennent compte des données manquantes

1.4.2.1) Analyse par paires (pairwise deletion)

La méthode consiste à analyser chaque aspect du problème à l'aide d'une partie des données disponibles. Les estimations sont générées à l'aide des variables d'intérêt présentes seulement. S'il s'agit, par exemple, d'estimer la corrélation pour un ensemble de variables X_1, X_2 et X_3. La corrélation entre X_1 etX_2 sera calculée en se servant des unités disponibles pour X_1 et X_2. La corrélation entre X_1 et X_3 sera elle aussi réalisé avec les cas disponibles pour X_1 et X_3.Idem pour X_2 et X_3. Ceci a pour conséquence que les corrélations sont conduites avec différents sous-ensembles de données qui ne sont pas nécessairement similaires et comparables entre eux. Cette méthode conserve plus d'observations que la « Complete Case Analysis ». De plus, le nombre d'observations variera d'une analyse à l'autre. Comme pour le « Complete Case Analysis », si les données manquantes diffèrent des données complètes, cette méthode conduira à des estimations biaisées.

1.4.2.2) Méthodes d'imputation simples

a) Imputation par la moyenne

Cette méthode est aussi appelée imputation par la moyenne non conditionnelle, car aucune autre information auxiliaire (condition) n'est utilisée pour imputer cette valeur aux données manquantes. Cette approche que la littérature attribue à wilks (1932) est pratique, car elle permet d'avoir un ensemble de données complet (Enders, 2010). Elle consiste à remplacer les valeurs manquantes ou les non-réponses par la moyenne des valeurs observées. Elle est très attrayante par sa simplicité et les données imputées peuvent être analysées comme si on disposait des données complètes dès le départ. Cependant, l'approche conduit à une

estimation biaisée des paramètres de la distribution. L'imputation par la moyenne réduit les associations entre les variables, car elle injecte des valeurs qui n'ont pas de relation avec les autres variables (Enders, 2010). En effet, les variances et les covariances entre variables imputées seront diminuées. Si l'on observe l'équation (1) on peut déduire que, lors de l'utilisation de cette technique, le numérateur est augmenté de 0 alors que l'échantillon augmente par la réintégration des unités exclues à cause des DM. Cette augmentation du dénominateur fait baisser l'estimation de la covariance de X. Ceci est valable aussi pour la corrélation de Pearson ($r= cov_{xy}/\sigma_x\sigma_y$). Mis à part la moyenne, les autres paramètres de la distribution qui seront produits seront aussi biaisés si l'imputation par la moyenne est utilisée.

$$\widehat{Cov_{xy}} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)} \qquad (1)$$

(Little et Rubin, 2002) ont proposé des ajustements pour obtenir des covariances et des corrélations non biaisées sous l'hypothèse MCAR. Les résultats de simulations montrent que l'imputation par la moyenne est la plus mauvaise approche à adopter. Elle ne peut généralement pas être défendue et doit être évitée (Enders, 2010).

Une amélioration à cette méthode consiste à imputer, non pas par la moyenne de l'ensemble des données complètes, mais par la moyenne de la classe à laquelle appartient l'unité. Les observations qui contiennent des DM et celle qui n'en contiennent pas sont regroupées dans des classes ou cellules d'ajustement selon les critères communs disponibles puis des moyennes sont calculées pour chaque classe observée et les DM sont remplacées par la moyenne de la classe à laquelle appartient l'observation (Little et Rubin, 2002).

b) Imputation par la médiane

L'imputation par la moyenne est très sensible aux valeurs aberrantes. L'imputation par la médiane apporte plus de robustesse à l'estimation. Dans cette approche, la médiane des données complètes est utilisée pour remplacer les données manquantes. S'il s'agit de données catégorielles, la valeur modale est utilisée. Elle est largement utilisée quand les données déviaient de la distribution normale. Le remplacement des DM par la médiane des valeurs complètes a un effet nuisible sur les estimations des paramètres. Cette approche a tendance à produire des variances trop larges qui ne sont pas optimales pour éviter l'erreur de type I et rejeter l'hypothèse nulle. ¸

c) Imputation par la moyenne conditionnelle

Remplacer les valeurs manquantes par la moyenne des données complètes est la plus simple méthode pour estimer les DM. Cependant, cette stratégie peut s'avérer problématique si d'autres variables auxiliaires nous permettent de juger que la moyenne est une estimation non plausible de la mesure manquante. De meilleures estimations peuvent être obtenues à l'aide d'une régression (Frane, 1976). L'approche, proposée au départ par Buck (1960), estime en premier lieu les paramètres de la régression à partir des données complètes Y^{Obs} de l'échantillon. Ces estimations sont ensuite utilisées pour prédire les valeurs manquantes à l'aide des autres variables auxiliaires présentes (Little et Rubin, 1987). Les variables ont tendance à être corrélées. Il est donc naturel d'utiliser l'information que fournissent les mesures disponibles pour estimer les valeurs manquantes. L'imputation par la moyenne conditionnelle est meilleure que l'utilisation de la moyenne non conditionnelle, mais induit également des biais aux paramètres estimés. En effet, les valeurs imputées seront fortement corrélées avec les autres variables (Enders, 2010). Par conséquent, l'imputation par régression surestime le coefficient R^2 et la corrélation entre les variables et sous-estime, à un niveau

moindre que l'imputation par la moyenne non conditionnelle, la variance et la covariance des variables (Enders, 2010). Le même auteur déconseille l'utilisation de cette méthode si les données imputées sont prédestinées à être analysée par un modèle de régression.

d) Imputation en utilisant la régression stochastique

L'imputation par régression présente de nombreux défauts et les paramètres estimés sont biaisés. En effet, l'imputation par les méthodes déterministes déforme la distribution des variables imputées et la variance est généralement sous-estimée (Allison, 2001; Enders, 2010; Little et Rubin, 2002; Rubin, 1996) . Pour remédier à ces biais, la régression stochastique peut être utilisée pour prédire des valeurs de remplacement pour les données manquantes à partir des variables complètes et en incorporant à la valeur prédite un résidu additionnel. Ce résidu est sélectionné de façon aléatoire à partir d'une distribution normale $N(0,\sigma 2)$. La variance $\sigma 2$ étant la variance résiduelle de la régression ou la variance des résidus. L'imputation par régression stochastique est semblable à l'imputation par régression sauf que les valeurs proposées pour remplacer les DM sont augmentées d'un nombre aléatoire tiré d'une distribution dont la moyenne est égale à 0 et dont la variance est celle générée par la régression des cas disponibles.

Le résidu ajouté rétablit la variabilité qui a été perdue et remédie au problème des biais (Enders, 2010). L'imputation par régression stochastique est la seule technique classique qui donne des résultats exempts de biais quand le processus qui génère les DM est MAR (Enders, 2010).

Comme pour la moyenne conditionnelle, cela se fait en deux étapes. Dans un premier temps, les paramètres de la régression sont estimés en utilisant les données complètes uniquement, puis les valeurs manquantes sont prédites en utilisant les paramètres déjà estimés. Finalement, un résidu normalement distribué est rajouté à chaque valeur prédite pour rétablir la variabilité (Enders,

2010). Little et Rubin (2002) ont démontré que cette approche produit des paramètres non biaisés.

e) Imputation par la dernière valeur observée (Last Observed Carried Forward).

Le problème des données manquantes est très fréquent dans les données longitudinales. Selon (Roth, 1994), ce problème serait plus sévère avec ce type de données et les techniques de traitement connues sont inadéquates. Par exemple, « Listewise deletion » peut conduire à une perte énorme de données parce que les sujets sont observés plusieurs fois (Roth, 1994). En effet, il n'est pas exclu d'avoir une observation manquante et toutes les données concernant le sujet sont éliminées dans ce cas. La technique « *Last Observed Carried Forward* » est spécifique aux données longitudinales. Elle consiste à remplacer la valeur manquante par la dernière valeur connue. Cette approche suppose que les valeurs ne changent pas dans le temps et ne tient pas des changements par exemple un revenu qui augmente dans le temps.

Cette stratégie s'applique à ceux qui quittent l'étude de façon permanente comme à ceux qui le font temporairement (Enders, 2010). D'autres approches peuvent être appropriées comme la moyenne des valeurs disponibles pour le même sujet, la valeur la plus proche dans le temps ou la moyenne des deux valeurs avant et après celle qui manque par exemple.

f) Méthode de l'indicatrice

La méthode est très populaire et consiste à ajouter, pour chaque variable continue qui présente des DM, une nouvelle variable binaire. Cette stratégie n'exclut pas les observations qui présentent des DM, mais rajoute au modèle à l'étude des variables binaires. Pour chaque variable incomplète une nouvelle variables binaires est ajoutée et sera codée 1 si la valeur est manquante et 0 dans le cas contraire. En effet, en plus de remplacer les valeurs manquantes par une valeur

donnée, qui peut être la moyenne par exemple pour les variables continues, une variable binaire additionnelle est rajoutée au modèle d'analyse. Parfois, la méthode de l'indicatrice est combinée à d'autres méthodes d'imputation comme la régression linéaire. Pour les variables catégorielles une nouvelle catégorie « manquante » est ajoutée. Même s'il a été relaté que cette méthode produit des estimations biaisées, elle continue à être largement utilisée (Knol *et al.*, 2010; Pedersen *et al.*, 2017)

g) Méthode intuitive ou déductive

Elle consiste à remplacer les données manquantes par des valeurs déduites à partir des données présentes des autres variables complètes de la même observation. Elle est principalement utilisée quand une seule valeur plausible de remplacement peut être déduite avec certitude à partir des autres mesures présentes. Elle utilise les relations logiques qui lient les variables pour en dériver une valeur de remplacement de la non-réponse hautement plausible. Par exemple, déduire l'âge du répondant à partir de sa date de naissance ou déduire que le répondant a terminé ses études secondaires s'il a étudié à l'université. Cette approche nécessite une très bonne connaissance des données et de l'environnement de collecte des données.

h) Méthode "hot-deck" et "cold-deck"

Dans cette méthode, la valeur manquante est remplacée par la valeur plausible observée chez un individu « donneur » qui est choisi en fonction de sa distance, par rapport au profil de l'individu « receveur ». Elle est largement utilisée dans les enquêtes d'opinion. Le principe de base est de remplacer la DM par une mesure empruntée à une unité similaire ayant participé à la même enquête. Le terme " hot " vient du fait que l'ensemble des données vient de l'enquête actuelle par opposition au " cold-deck " qui utilise les informations d'une ancienne enquête ou d'un ancien jeu de données. En d'autres termes, les DM sont remplacées par des

estimations raisonnables provenant d'une unité similaire. Les promoteurs de cette méthode avancent que l'imputation " hot-deck " tend à être plus performante que les autres techniques d'imputation, car les valeurs manquantes sont remplacées par des mesures réalistes (Peugh et Enders, 2004). De plus, les valeurs de substitution ne sont pas des moyennes qui déforment la distribution des données. Au contraire, les caractéristiques de la distribution sont préservées puisque la mesure empruntée du donneur provient d'une catégorie similaire.

Selon (Roth, 1994), la méthode "hot-deck" présente plusieurs inconvénients. Premièrement, il existe très peu de travaux théoriques et empiriques pour juger de son efficacité. De plus, le nombre de variables catégorielles nécessaires pour définir les classes peut devenir ingérable. Les variables auxiliaires continues doivent être transformées en variables catégorielles pour être utilisées par la méthode "hot-deck". Le nombre de catégories doit être suffisamment grand pour réaliser une bonne imputation à l'aide de cette méthode. Avec un nombre trop grand de catégories, la probabilité de trouver une valeur plausible provenant d'une observation similaire diminue (Nordholt et VanHuijsduijnen, 1997).

Il existe deux types d'imputation « hot-deck », l'une aléatoire et l'autre séquentielle. Dans la forme aléatoire, le donneur est choisi au hasard. Dans les deux cas, le choix du donneur est fait selon les catégories qu'il a en commun avec le receveur. Si de nombreuses catégories sont utilisées pour faire correspondre le donneur au receveur, on risque de trouver très peu de donneurs. Dans le cas où très peu de catégories sont utilisées, on risque d'imputer avec une valeur non réaliste (Sande, 1983). Comparée à d'autres méthodes d'imputation, l'approche « hot-deck » peut produire des données imputées qui préservent les paramètres de la distribution (Sande, 1983). L'approche 'Hot Deck' est capable de maintenir l'intégrité des données puisque les valeurs de replacements sont tirées à partir des données observées. Il en va de soi qu'il est impossible d'avoir des valeurs de remplacement des DM qui dépassent l'intervalle acceptable. Seules des valeurs réalisées qui ont du sens seront sélectionnées pour remplacer les DM. Elle est

intéressante par sa capacité de maintenir la nature discrète des données contrairement à certaines méthodes paramétriques qui nécessitent d'arrondir les valeurs prédites pour les rendre conformes à la nature originale des données à imputer. La variance obtenue avec la méthode « hot-deck » est plus forte que celle obtenue par la méthode d'imputation par la moyenne; elle présente donc l'avantage de ne pas déformer la distribution. Par ailleurs, l'algorithme 'Hot Deck' est intuitif et facile à comprendre.

La méthode d'imputation « cold-deck » par opposition à la méthode « hot-deck » impute aux données manquantes des valeurs tirées d'une enquête précédente. Elle est utilisée, principalement, dans les enquêtes d'opinion réalisées par les bureaux gouvernementaux de statistiques. À l'instar de l'approche « hot-deck », elle utilise également les données présentes pour créer des catégories et choisir les données d'un donneur appartenant à la même classe pour remplacer les données manquantes. Cette méthode a été implémentée dans R et un module lui est consacré.

1.4.2.2) Méthodes élaborées d'analyse en présence des données manquantes

Dans les sections précédentes, plusieurs méthodes ont été présentées. Certaines méthodes ignorent tout simplement les observations avec données manquantes et estiment les paramètres des modèles statistiques uniquement avec le reste des observations. Les Tableaux 2 et 3 synthétisent les différentes méthodes et relatent les limites de chacune d'elles. L'imputation multiple et le maximum de vraisemblance sont cités comme les deux méthodes les plus élaborées, ou «*State of the art* », pour traiter les données manquantes. Des études empiriques ont montré que ces deux méthodes produisent des estimations similaires (Schafer et Graham, 2002). De nouvelles méthodes (MissForest, MissMDA) non paramétriques qui libèrent l'analyste des hypothèses contraignantes relatives à la distribution des données ont vu le jour ces dernières années. Elles ambitionnent d'être aussi

efficaces que l'imputation multiple et maximum de vraisemblance. Nous allons détailler ces méthodes dans les sections suivantes.

a) Méthode du maximum de vraisemblance

L'approche par maximum de vraisemblance est toujours une meilleure alternative aux méthodes dites simples présentées dans le Tableau 1. Le fait qu'elle soit disponible dans les principaux logiciels statistiques la rend facile à utiliser. Rubin (1976) et Little et Rubin (1987) ont démontré que, sous l'hypothèse MAR, on n'est pas obligé de se préoccuper de la présence des données manquantes si l'on utilise le maximum de vraisemblance (MV) pour l'inférence statistique. Le mécanisme qui génère les données manquantes peut être ignoré sans remettre en cause la validité des résultats. Rubin (1976) et (Little et Rubin, 1987) réfèrent à cette fonction comme « Maximum de vraisemblance ignorant le mécanisme générant les données manquantes ». L'estimation des paramètres par maximum de vraisemblance procède de façon itérative jusqu'à l'obtention d'une solution qui donne la plus grande valeur du log-vraisemblance. Une méthode générale d'utilisation du MV en présence de données manquantes a été décrite par Dempster, Laird et Rubin (1977) dans leur article sur l'algorithme EM. L'algorithme EM est un processus itératif qui vise à obtenir une estimation des paramètres inconnus en présence de données manquantes. Chaque itération EM comprend deux étapes : une étape estimation E et une étape maximisation M.

b) Les méthodes d'imputation multiples

L'imputation multiple, développée principalement par Rubin (1976), consiste à générer plusieurs copies complètes de la base originale qui contient des DM, puis à effectuer les analyses et combiner les résultats. La figure 2 illustre ce processus. Destinée au départ à traiter les non-réponses rencontrées lors des enquêtes d'opinion, elle est très générale et peut être utilisée dans d'autres contextes (Jones,

1996). Selon Nordholt et VanHuijsduijnen (1997), l'imputation multiple ne crée pas de nouvelles informations, mais représente plutôt l'information disponible pour qu'elle puisse être analysée à l'aide des méthodes d'analyse statistique pour données complètes.

Cette approche est composée de trois étapes. Une première étape consiste à remplacer les données manquantes par des estimations: plusieurs jeux de données complets (*m* jeux) sont produits et dans chaque jeu, une valeur estimée différente remplace la valeur manquante. La valeur estimée qui remplacera la donnée manquante peut provenir d'une régression stochastique par exemple. La deuxième étape consiste à analyser les jeux de données complétés selon les méthodes standards habituelles en utilisant le même modèle. L'étape finale sert à combiner les paramètres produits par le modèle selon la méthode décrite par (Rubin, 1976).

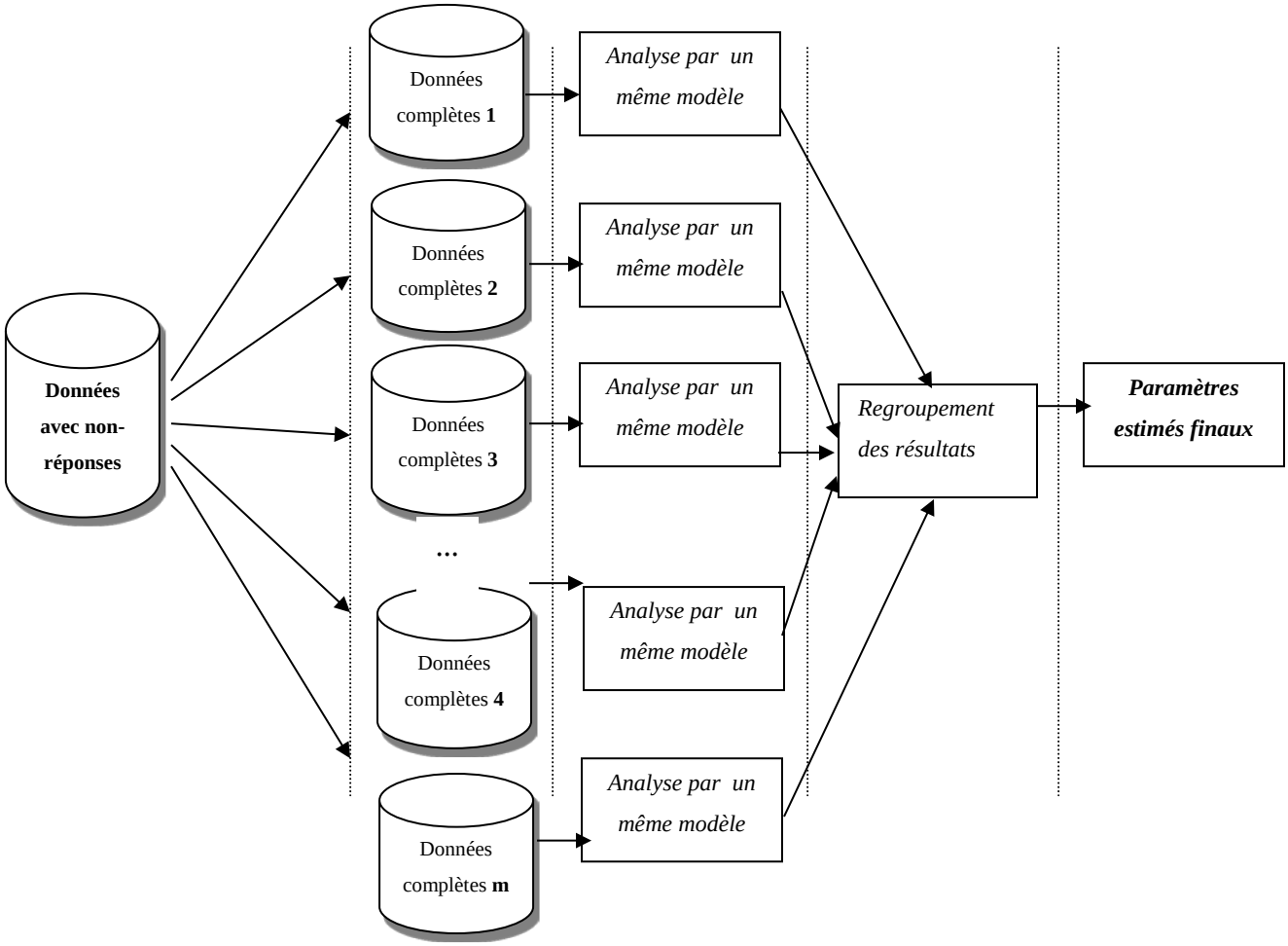
Concernant le processus qui génère les DM, bien que l'IM repose sur le mécanisme MAR de récentes études indiquent que l'IM peut être employée dans des conditions MNAR. Schafer et Graham (2002) notent que MAR n'est pas une condition à l'application de l'imputation multiple. Selon Rubin (1996), l'imputation multiple peut produire des résultats satisfaisants même si on s'éloigne légèrement de MAR.

Par ailleurs, plus le nombre de jeux de données complétés est grand, plus les paramètres estimés seront précis. Généralement, de bons résultats sont obtenus à l'aide d'un petit nombre d'imputations, soit entre 5 et 10 (Little et Rubin, 2002). En remplaçant chaque valeur manquante par plusieurs valeurs estimées plausibles, l'imputation multiple conserve l'incertitude relative à ces valeurs manquantes. Cette méthode préserve aussi la distribution des données puisque, non pas une seule et unique valeur est utilisée, mais plusieurs valeurs plausibles sont utilisées pour remplacer une seule donnée manquante.

L'IM peut être paramétrique quand un modèle d'imputation est fourni et dite non paramétriques si aucun modèle n'est spécifié. Les méthodes paramétriques se

basent sur une série d'hypothèses qui ne sont pas toujours respectées. De plus, toutes les variables qui peuvent contribuer à mieux prédire les valeurs de remplacement ainsi que leurs interactions doivent être incluses dans le modèle d'imputation. La construction d'un tel modèle est complexe et pas toujours facile à cause de la portion de données qui manquent et la non-homogénéité des variables. L'imputation non paramétrique évite ces contraintes. En effet, les méthodes non paramétriques sont moins restrictives et peuvent tenir compte des interactions qui sont parfois ignorées par les méthodes paramétriques. De nombreuses méthodes qui libèrent l'analyste des contraintes des méthodes paramétriques ont fait l'objet de développent ces dernières années. Les auteurs réclament qu'elles sont aussi bonnes que l'IM. Nous allons en présenter quelques une dans la section suivante.

Figure 2 : Représentation schématique de l'imputation multiple[1]



[1] Cette figure est inspirée de (Schafer, 1994 page 188)

L'imputation multiple a été implémentée dans de nombreux logiciels (par exemple : SAS, R, Stata, SPSS) Ces modules peuvent être classées en deux grandes classes : des techniques basées sur une distribution normale multivariée jointe pour toutes les variables et des techniques plus flexibles qui se départissent de la normalité et qui consistent à imputer chaque variable par un modèle conditionnel spécifique selon le type de la variable . Cette dernière approche est appelée imputation par equations chainées (FCS : Fully Conditionnal Specification).

Il existe différentes méthodes d'imputation multiples qui peuvent être utilisées pour traiter les données manquantes. Le choix d'utiliser l'une de ces méthodes dépend de la catégorie des données et de leur configuration « pattern ». Le Tableau 1, tiré de la documentation SAS (SAS/STAT 14.1 User's Guide), résume les différentes méthodes d'imputations disponibles dans la procédure "proc MI" de SAS et dans quel cas chacune de ces méthodes serait pertinente à utiliser.

Tableau 1 : Liste des méthodes d'IM disponibles dans SAS version 9.4.

Configuration des DM	Type de variables avec DM	Type des variables auxiliaires	Méthodes disponibles avec proc MI
Monotone	Continue	Arbitraire	Monotone: régression Monotone: PMM[2] Monotone: "Propensity score"
Monotone	Ordinale	Arbitraire	Monotone: régression logistique
Monotone	Nominale	Arbitraire	Monotone: fonction discriminante
Arbitraire	Continue	continue	MCMC[3] : imputation complète MCMC : rendre données monotones
Arbitraire	Continue	Arbitraire	FCS[4]: régression FCS : PPM
Arbitraire	Ordinale	Arbitraire	FCS : régression logistique
Arbitraire	Nominale	Arbitraire	FCS : fonction discriminante

[2] *Predictive Mean Matching*

[3] *Markov Chain Monte Carlo.*

[4] *Fully Conditionnal Specification.*

1.4.2.3) Méthode non paramétriques pour traiter les DM

Imputation à l'aide de l'algorithme d'analyses factorielles

L'analyse en composante principale (ACP) est au cœur de cette méthode (Audigier, Husson et Josse, 2012; Stacklies et al., 2007). Le but initial de la méthode est de réaliser l'analyse en composante principale en présence des données manquantes. L'imputation est réalisée durant le déroulement du processus et peut être utilisée comme méthode d'imputation.

En effet, l'approche classique pour générer les données manquantes en ACP consiste à minimiser la fonction de coût (l'erreur de reconstitution) sur tous les éléments présents. Ceci peut être effectué à travers un algorithme d'ACP itérative (aussi appelé expectation maximisation PCA, EM-PCA) (Dempster, Laird et Rubin, 1977). Celui-ci consiste à attribuer une valeur initiale aux données manquantes, effectuer l'analyse (ACP) sur le jeu rendu complet, compléter les données manquantes via la formule de reconstitution pour un nombre d'axes fixe, et recommencer ces deux étapes jusqu'à convergence (Kiers, 1997). Les paramètres (axes et composantes) ainsi que les données manquantes sont de cette manière simultanément estimés. Par conséquent cet algorithme peut être vu comme une méthode d'imputation.

L'AFDM (Analyse Factorielle des Données Mixtes) généralise l'ACP et l'ACM, elle permet de traiter à la fois des données quantitatives propres à l'ACP et des variables qualitatives propres à l'ACM. La force de l'AFDM réside donc dans la prise en compte des relations entre individus, au même titre que toutes les autres méthodes factorielles, mais aussi, et c'est là son unicité, dans les relations entre les variables quantitatives et qualitatives équilibrées, renforçant ainsi la qualité d'imputation que l'on aurait eue en utilisant séparément une imputation par ACP et une par ACM. L'équilibre entre les différents types de variables est important au risque d'altérer l'imputation.

Même si l'imputation n'est pas la fonction principale de cet algorithme il demeure un excellent moyen pour imputer les DM. En effet l'imputation est réalisée en prenant en compte les similarités entre individus et les liens entre les variables. Cette méthode peut servir comme une alternative aux autres méthodes dédiées à imputer des données de différent type (continue, catégorielles …). En présence de données quantitative et qualitatives, cette méthode génère des imputations simples, donc une seule base de données est produite.

Imputation via l'algorithme 'Baboon'

La majorité des algorithmes semi-paramétriques utilisés pour faire de l'IM se basent sur la méthode « Predictive Mean Matching » (PMM) pour prédire des valeurs de remplacement aux DM. Cependant, leur application se limite aux données continues. Le concept de base derrière PMM est d'attribuer à une DM la valeur observée la plus proche parmi les unités observées. C'est un aspect qui rappelle KNN. En remplaçant les DM par des valeurs observées, cette approche respecte la contrainte stipulant que les valeurs de remplacement doivent être plausibles. Une caractéristique intéressante de PMM est le fait qu'elle soit robuste à la mauvaise spécification du modèle (Meinfelder, Schnapp et Meinfelder, 2015). L'algorithme proposé par (Meinfelder, Schnapp et Meinfelder, 2015) élargit l'utilisation du PMM aux données catégorielles via un bootstrap bayesien. Pour pouvoir étendre l'utilisation aux variables binaires ou catégorielles cet auteur propose d'utiliser respectivement des modèles logit binaires ou des modèles logit multinomiale à la place des prédicteurs linéaires pour calculer la distance.

L'algorithme se compose de deux variantes bayésiennes pour imputer les données manquantes. La première est séquentielle et impute les variables l'une après l'autre en combinant la méthode des régressions séquentielles et la PMM pour données catégorielles. La deuxième variante qui est basée aussi sur PMM vise à

imputer plusieurs variables en même temps. Les deux variantes peuvent être utilisées aussi pour réaliser des imputations simples.

Imputation à l'aide des forêts aléatoires (Stekhoven et Bühlmann, 2011)

Cette méthode est basée sur une approche non paramétrique pour imputer les données manquantes. Elle peut traiter des données mixtes en présence de différent type de variables, des relations non linéaires et des interactions complexes. L'algorithme RandomForest (RF) sur les forêts aléatoires (Leo Breiman, 2001) et son implémentation dans R « RandomForest » est au cœur de cette approche. Les autres méthodes, MICE par exemple, dépendent du modèle d'imputation dont la construction s'avère parfois difficile en présence de données mixtes. Or la qualité de la méthode et des valeurs de remplacement dépendent de la qualité du modèle utilisé. RF est connue pour sa capacité à traiter des données mixtes (binaires, continue …), multidimensionnelles et présentant des interactions complexes.

Pour chaque variable cet algorithme construit un modèle de type forêt aléatoire basé sur les données disponibles et utilise ce modèle pour prédire des valeurs de remplacement pour les DM. Ce processus progresse jusqu'à la convergence ou l'atteinte du nombre maximal d'itérations fixé par l'utilisateur. L'algorithme s'exécute de façon itérative tout en mettant à jour la matrice des variables imputées et en vérifiant sa propre performance entre deux itérations. La vérification se fait en comparant le résultat de la valeur de remplacement actuelle avec la valeur précédente et en s'arrêtant aussitôt que la différence augmente.

Une estimation de l'erreur d'imputation est fournie à l'utilisateur en se basant sur des estimations (O.O.B) out of the bag. Stekhoven et Bühlmann (2011) ont établi que cette estimation est une bonne représentation de l'erreur d'imputation. Cette méthode génère un seul jeu de données à la fin de la phase d'imputation.

Imputation à l'aide de l'algorithme des K plus proche voisin KNN (Troyanskaya et al., 2001)

Cette approche gagne de plus en plus de popularité et elle est très utilisée pour traiter les DM. Elle consiste à choisir une valeur de remplacement parmi les unités similaires qui coexistent dans la même base. En identifiant de façon précise un ou plusieurs donneurs potentiels, la valeur de remplacement sera une valeur plausible proche de la vraie valeur. L'identification des donneurs se fait en calculant la distance entre les unités qui vont accueillir les valeurs et les donneurs potentiels. De nombreuses méthodes pour calculer cette distance ont été proposées par les auteurs (Faisal et Tutz, 2017; Liu et Zhang, 2012). (Faisal et Tutz, 2017) proposent d'utiliser une distance pondérée qui tient compte des associations entre les variables. Les variables qui ont une forte association avec la variable qu'on veut imputer vont avoir un rôle prépondérant pour estimer la distance.

Cette approche a donc l'avantage de ne pas déformer la distribution des données. Pour les données continues, la valeur de remplacement sera la moyenne calculée à partir des valeurs des k proches voisins. Pour les variables discrètes, c'est la valeur modale qui sera choisie.

Parmi les caractéristiques les plus intéressantes de cette approche : a) Les valeurs de remplacements proposées ne seront pas étrangères au jeu de données, mais bien au contraire des valeurs réalisées, b) KNN utilise les variables auxiliaires et préserve donc la structure initiale des données et c) cette approche est complètement non paramétrique et ne requière pas un effort supplémentaire pour construire un modèle d'imputation complexe.

Le nombre de voisins qui vont jouer le rôle de donneurs potentiels peut être fixé par l'analyste qui procède à l'imputation. Des simulations réalisées par Beretta et Santaniello (2016) ont démontrées qu'on peut utiliser un faible k pour préserver la structure originale des données.

De nombreuses techniques basées sur l'algorithme KNN ont été proposées par les auteurs : GKNN (Liu et Zhang, 2012), wNNSel (Faisal et Tutz, 2017), YaImpute (Crookston et Finley, 2008), KNNcatImpute (Schwender, 2012).

Tableau 2 : Synthèse des méthodes simples pour traiter les DM.

Technique	Brève description	Condition d'utilisation	Avantages	Inconvénients	Références
Utilisation des cas complets seulement (Listewise deletion).	Ignore les observations incomplètes.	Faible taux de DM (<=5%) Mécanisme MCAR.	Simple.	Perte d'informations. Perte de précision. Biais si non MCAR.	(Little et Rubin, 2002)
Analyse des cas disponibles (Pairwise deletion)	Utilise toutes les observations dont la variable d'intérêt est disponible.	Mécanisme MCAR.	Utilise toute les données par rapport à « Listwise deletion ».	Utilise des sous-ensembles du jeu de données différents pour les analyses sur différentes variables	(Little et Rubin, 2002)
Imputation par la moyenne ou la médiane	Les DM sont remplacées par la moyenne ou la médiane calculée à partir des données complètes.		Préserve toutes les observations de l'échantillon.	Sous-estimation des paramètres (covariance, corrélation..) sauf la moyenne. Déforme la distribution. Variance doit être ajustée (Little & Rubin, 2002). Sensible aux valeurs aberrantes.	(Enders, 2010; Little et Rubin, 2002)
Imputation par la moyenne conditionnelle ou régression.	Utilise les variables auxiliaires disponibles pour prédire des valeurs de remplacements des DM à l'aide de modèles de régression.	MCAR. Corrélation doit être importante entre la variable avec DM et les variables auxiliaires	Améliore la précision par rapport à l'imputation par la moyenne. Préserve les associations entre les variables. Bonne estimation de la moyenne si normalité et MCAR	Surestime R^2 et la corrélation. Sous-estime variance et covariance. Compliquée dans le cas de plusieurs configurations de DM.	(Buck, 1960; Little et Rubin, 2002; Peugh et Enders, 2004)
Imputation par régression stochastique.	Un résidu normalement distribué est ajouté à la valeur prédite par un modèle de régression.	MAR (Enders, 2010)	Paramètre estimé non biaisé	Atténue la variance ce qui peut mener à des erreurs de Type I. Peut produire des résultats légèrement différents avec les mêmes données si elle est utilisée plus qu'une fois.	(Enders, 2010; Little et Rubin, 2002)

Ajout d'une variable indicatrice	Ajout d'une variable binaire, combinée à l'imputation simple.	Seules les variables explicatives ont des DM. Modèle linéaire sans interactions.	Utilise toute l'information disponible y compris la présence ou non de mesure. Peut être utilisée avec n'importe quelle méthode d'imputation.	Paramètres sévèrement biaisés même si la condition MCAR tient.	(Buuren, 2012)
La technique « hot-deck » et « cold-deck »	La valeur d'un donneur appartenant à la même classe remplace la DM. Le donneur provient de l'enquête actuelle pour l'approche « hot-deck ».	Utilisée généralement dans les enquêtes d'opinion (Stat US, Stat Canada,..)	Préserve la distribution. Valeurs réalistes sont utilisées.	Nombre de catégorie difficile à gérer).	(Buuren, 2012; Nordholt et VanHuijsduijnen, 1997; Sande, 1983)
Last observation carried forward	Copier la valeur de la période précédente pour remplacer la DM.	Données longitudinales	Simple	Ne tient pas compte du changement de la valeur. Par exemple; un revenue n'est pas statique.	(Enders, 2010)

Tableau 3 : Synthèse des méthodes avancées pour traiter les DM.

Technique	Brève description	Condition d'utilisation	Avantages	Inconvénients	Références
Maximum de vraisemblance	Maximum de vraisemblance ignorant le mécanisme qui génère les DM. Un exemple serait l'algorithme EM itérant estimation (E) et maximisation (M) jusqu'à convergence.	MAR et MCAR	Paramètres produits sont non biaisé sous MAR Surclasse les méthodes simples.	Donne des paramètres biaisés dans le cas NMAR Temps pour convergence long dans le cas EM.	(Donzé, 2001; Enders, 2010)
Imputation multiple	Consiste à produire m copies des données originelles. Chaque DM est remplacé par plusieurs valeurs estimées. Des analyses sont faites en utilisant un même modèle sur chaque jeu de données, puis les résultats sont regroupés.	MCAR, MAR et MNAR	Incorpore l'incertitude engendrée par les DM. Utilisable avec n'importe quel type de données et n'importe quel modèle	Si elle est utilisée plus qu'une fois, peut produire des résultats légèrement différents avec les mêmes données.	(Donzé, 2001; Enders, 2010; Graham, 2009; Little et Rubin, 2002; McKnight, 2007)

Chapitre 2 : Fondement théorique de l'imputation multiple

C'est Rubin (1978) qui a proposé de traiter les données manquantes à l'aide de l'imputation multiple. Cette technique a été par la suite définie en détail par Little et Rubin (1987) et Schafer (1997). Elle consiste à remplacer chaque donnée manquante par des valeurs estimées à l'aide d'un modèle qui utilise l'information disponible dans la base de données. Elle permet ainsi de conserver les relations qui existent entre les variables, de préserver les aspects importants de la distribution et de refléter correctement l'incertitude reliée aux données manquantes. De plus, elle présente l'avantage d'être appliquée aux données indépendamment des analyses ultérieures. L'imputation multiple est fondée sur la méthode Monte Carlo par Chaîne de Markov qui sera décrite dans ce chapitre.

2.1) Méthode de Monte Carlo par Chaîne de Markov (MCMC)

La méthode MCMC a été appliquée en physique dans les années 1950 comme outil majeur d'investigation de la physique de la matière. En statistique, la principale application consiste à l'utiliser pour faire des simulations aléatoires de distributions complexes. Cette technique a de nombreuses applications dans différents domaines, notamment dans le traitement des données manquantes. Elle est utilisée lorsque la distribution à posteriori ne peut pas être évaluée directement ou lorsqu'il n'existe pas de solution analytique à des problèmes complexes de grande dimension. Parmi les applications simples, on trouve le calcul d'intégrale, l'échantillonnage de variables aléatoires et l'optimisation des fonctions. En effet, cette méthode permet de faire de l'intégration numérique en utilisant l'aléatoire lorsque l'on cherche à approcher une densité qui est incalculable explicitement à l'aide des méthodes analytiques. Supposons que l'on soit dans une situation où l'on cherche à calculer l'espérance à posteriori d'un paramètre donné. L'espérance

étant définie par une intégrale, il se peut que cette intégrale soit difficile à calculer. Les techniques d'échantillonnage sont là pour aider à contourner ce problème en simulant ces intégrations. Parmi les algorithmes les plus utilisés, on trouve l'algorithme de Metropolis-Hastings, l'échantillonneur de Gibbs et l'augmentation de données. Dans les simulations MCMC, des chaines de Markov, suffisamment longues, sont construites et vont se stabiliser en une distribution stationnaire qui est en réalité la distribution d'intérêt.

2.2) Échantillonnage de Gibbs

L'échantillonneur de Gibbs est l'application la plus simple des méthodes MCMC. Comme décrit dans l'article de Casella et George (1992), il consiste en une chaîne d'étapes ou les valeurs de l'étape actuelle dépendent de celles de l'étape précédente. En effet, supposons que l'on s'intéresse à obtenir les paramètres de la densité conjointe $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y_1}, \dots, \boldsymbol{y_p})$.

$$\boldsymbol{f}(\boldsymbol{x}) = \int \dots \int \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y_1}, \dots, \boldsymbol{y_p})\, \boldsymbol{dy_1} \dots \boldsymbol{dy_p}$$

L'approche classique préconise de calculer et d'utiliser $\boldsymbol{f}(\boldsymbol{x})$ pour obtenir les paramètres recherchés. Cependant, l'intégration de cette fonction peut s'avérer complexe voire difficilement réalisable. L'échantillonneur de Gibbs permet de contourner cette difficulté et propose d'approcher $\boldsymbol{f}(\boldsymbol{x})$ à l'aide de l'échantillonnage aléatoire. En effet, des tirages répétitifs sont effectués dans la distribution conditionnelle de chaque variable afin d'approcher la distribution conjointe $(\boldsymbol{x})$. Donc l'échantillonneur de Gibbs peut être utilisé quand la distribution conjointe n'est pas connue, mais les distributions conditionnelles de chaque variable doivent être obligatoirement connues.

Voici un exemple pour illustrer cela. Supposons que l'on dispose d'un ensemble de variables aléatoires $X_1, \dots, X_j$ et que l'on désire tirer aléatoirement un certain

nombre fini de ces j variables aléatoires que l'on note P(X_1, … , X_j). Il est très complexe d'obtenir ces valeurs à partir de la distribution conjointe. L'échantillonneur de Gibbs permet de générer des valeurs issues des distributions conditionnelles. Voici les étapes itératives qui permettent d'obtenir, indirectement, des tirages dans la distribution conjointe $\boldsymbol{P(X_1, \dots, X_j)}$:

Obtenir $\boldsymbol{x_1^{(i+1)}}$ à partir de la distribution conditionnelle $\boldsymbol{p(x_1|x_2^{(i)}, \dots, x_j^{(i)})}$

Obtenir $\boldsymbol{x_2^{(i+1)}}$ à partir de la distribution conditionnelle $\boldsymbol{p(x_2|x_1^{(i+1)}, x_3^{(i)}, \dots, x_j^{(i)})}$

.

.

.

Obtenir $\boldsymbol{x_j^{(i+1)}}$ à partir de la distribution conditionnelle $\boldsymbol{p(x_j|x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)})}$

Ce processus est répété jusqu'à ce qu'il y ait convergence. Autrement dit, ces itérations sont produites de façon successive jusqu'à l'obtention de valeurs stables. Une bonne partie des tirages initiaux est ignorée, cette période est appelée période de réchauffe ou « burn in ».

2.3) L'algorithme « Data augmentation »

Développé par Tanner et Wong (1987) cet algorithme est très similaire à l'échantillonneur de Gibbs. Il repose sur une approche qui considère les données manquantes comme des paramètres supplémentaires qui doivent être estimés. Les données manquantes et les paramètres de la distribution font l'objet alternativement de simulations qui convergent vers la distribution prédictive à

posteriori qui sera utilisée pour imputer les données. Cette méthode est décrite de façon détaillée dans Schafer (1997).

Supposons que l'on dispose d'un vecteur aléatoire z de dimension deux et partitionné en deux sous-vecteurs $z=(u,v)$. Admettons que $F(z)$, la densité conjointe de z, soit difficile à simuler alors que les deux densités conditionnelles $F(u|v)=g(u|v)$ et $F(v|u)=h(v|u)$ le sont facilement. Soit $Z^{(t)}=(z_1^{(t)}, z_2^{(t)}, \ldots, z_m^{(t)})$ un échantillon de taille m tiré d'une densité approximative de la distribution $F(z)$. Considérons :

$$Z^{(t)} =(z_1^{(t)}, z_2^{(t)}, \ldots, z_m^{(t)})$$

$$Z^{(t)} =((u_1^{(t)}, v_1^{(t)}), (u_2^{(t)}, v_2^{(t)}), \ldots, (u_m^{(t)}, v_m^{(t)}))$$

Le tirage se fait à l'aide des deux étapes suivantes.

Premièrement :

$$U^{(t+1)} =(u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_m^{(t+1)}).$$

Est tiré à partir de la distribution conditionnelle : $u_i^{(t+1)} \sim g(u|v_i^{(t)})$ de façon indépendante pour $i=1,2, \ldots ,m$.

Deuxièmement :

$$V^{(t+1)} =(v_1^{(t+1)}, v_2^{(t+1)}, \ldots, v_m^{(t+1)}).$$

Est tiré à partir de la distribution conditionnelle : $v_i^{(t+1)} \sim g(v|u_i^{(t+1)})$ de façon indépendante pour $i=1,2, \ldots ,m$.

On obtient alors :

$$Z^{(t+1)} = (z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_m^{(t+1)})$$

$$Z^{(t)} = ((u_1^{(t+1)}, v_1^{(t+1)}), (u_2^{(t+1)}, v_2^{(t+1)}), \ldots, (u_m^{(t+1)}, v_m^{(t+1)}))$$

Tanner et Wong(1987) ont démontré que la distribution de $Z^{(t)}$ tend vers $F(z)$ quand $t \to \infty$. Ce résultat ne requiert pas que m soit grand. En particulier, quand $m=1$ l'algorithme « Data augmentation » (augmentation de données) peut être considéré comme étant un cas spécial de l'échantillonneur de Gibbs avec $z=(u,v)$.

Le non « data augmentation » vient de l'application de cet algorithme en présence de données manquantes. Supposons que l'on dispose d'une base de

données $\boldsymbol{Y}$=($\boldsymbol{Y_{obs}}$, $\boldsymbol{Y_{mis}}$) qui comporte des données manquantes. La distribution à posteriori F($\boldsymbol{\theta}|\boldsymbol{Y_{obs}}$) ne peut pas être facilement manipulée ni simulée, elle est même parfois inconnue ($\boldsymbol{\theta}$ étant les paramètres de cette distribution). Par contre, lorsque $\boldsymbol{Y_{obs}}$ est augmentée par des valeurs prédites de $\boldsymbol{Y_{mis}}$, la distribution à posteriori F($\boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}$) devient plus facile à cerner. Comme $\boldsymbol{Y_{mis}}$ et $\boldsymbol{\theta}$ sont tous les deux inconnues nous allons devoir les estimer. En effet, la technique d'augmentation de données se résume en deux étapes : une première étape dite d'imputation (I-Étape) et une seconde dite d'estimation des paramètres de la distribution à postériori (P-Étape). Après avoir attribué des valeurs de départ à $\boldsymbol{\theta}$, un processus itératif I-Étape suivi de P-Étape est conduit jusqu'à la stabilisation.

I-Étape : Conditionnellement aux paramètres $\boldsymbol{\theta^t}$, estimer des valeurs de $\boldsymbol{Y_{mis}}$ à l'aide de :

$$\boldsymbol{Y_{mis}^{t+1}} \sim \boldsymbol{F}(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}}, \boldsymbol{\theta^t})$$

P-Étape : Conditionnellement aux valeurs estimées à l'étape I-Étape, faire des tirages de nouvelles valeurs des paramètres $\boldsymbol{\theta}$:

$$\boldsymbol{\theta^{t+1}} \sim \boldsymbol{F}(\boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}^{t+1}})$$

Ce processus est répété en se donnant des valeurs de départ θ^0. La séquence de valeurs $\{\theta^t , Y_{mis}^t \, ou \, t-1,2,\ldots\}$ obtenue va converger vers sa densité stable F($\theta|Y_{obs}, Y_{mis}$). Les sous-séquences $\{\theta^t : t = 1,2,\ldots\}$ et $\{ Y_{mis}^t : t = 1,2,\ldots\}$ vont converger aussi vers leurs distributions conditionnelles respectives $F(\theta|Y_{obs})$ et $F(Y_{mis}|Y_{obs})$. Les deux algorithmes (l'algorithme DA en considérant m=1 et l'échantillonneur de Gibbs en considérant (Y_{mis} , θ) composé de θ et Y_{mis}) peuvent servir à traiter les données manquantes.

2.4) **Modèle d'imputation multivarié normal** vs **IM par équations chainées**

Deux approches d'imputation multiple sont décrites dans la littérature. La première basée sur la densité conjointe et la deuxième basée sur les distributions conditionnelles de chaque variable. L'imputation multiple est reconnue, de nos jours, comme l'une des meilleures options disponibles pour traiter les Données manquantes (DM). Deux adaptations principales de la théorie de Little (1987) sont recommandées pour traiter les données manquantes. Il s'agit de l'imputation multiple basée sur la distribution jointe (JM) qui est en réalité une distribution multivariée normale Schafer (1997) et de l'imputation multiple (IM) par équations chaînées décrites par Stef Van Buuren et Oudshoorn (1999). Nous allons décrire ces deux approches.

2.4.1) Modèle d'imputation multivarié normal

Implémenté par Schafer (1997), cette technique est conseillée s'il existe une distribution multivariée plausible qui permet de décrire les données. Malheureusement, il n'est pas facile de trouver cette distribution jointe. L'IM basée sur la distribution jointe consiste à bâtir un modèle prédictif à partir des données disponibles et de sélectionner des valeurs d'imputation en utilisant la technique Markov Chain Monte Carlo (MCMC). Le modèle qui va servir à générer les imputations peut être basé sur n'importe quelle distribution conjointe multidimensionnelle. La spécification d'un modèle précis qui décrit la distribution jointe pour un ensemble très grand de variable représente un défi majeur pour l'analyste. En effet, Il est très rare que le modèle MVN soit le modèle plausible qui convient aux données. C'est pour cette raison que dans la majorité des applications de l'imputation multiple, le modèle qui sera utilisé ne sera qu'une approximation du modèle qui décrit réellement les données. La plupart des approches qui ont adopté l'approche de la distribution joint (JM), comme 'proc MI' dans SAS (Yuan,

2011) et la routine 'Norm' dans R Schafer (1997), font l'hypothèse que les données suivent une distribution normale multivariée. Quand les variables ciblées par l'imputation multiple sont binaires ou ordinales, il est acceptable de générer des valeurs de remplacements selon le modèle normal en arrondissant à la catégorie la plus proche. Des simulations ont corroboré la robustesse de l'imputation multiple au modèle d'imputation et (Schafer et Graham, 2002). En pratique, le modèle normal multivarié est utilisé non pas uniquement pour générer des valeurs plausibles pour données manquantes continues, mais il est aussi appliqué aux données catégorielles. Schafer (1997) préconise d'arrondir les valeurs vers la catégorie la plus proche lorsqu'on utilise ce modèle pour imputer des données catégorielles. Allison (2005) suggère de ne pas arrondir et de faire appel à des méthodes adaptées aux variables catégorielles comme la régression logistique ou la fonction discriminante. Dans de nombreux cas, le modèle MVN peut générer des valeurs non plausibles ou qui n'ont pas de sens (port de ceinture de sécurité pour un cycliste ou un bébé en bas âge qui est assis dans le siège du conducteur d'une automobile par exemple).

C'est pour cette raison que les chercheurs préfèrent utiliser FCS à la place. Connue aussi, sous le nom MICE (Multiple Imputation by Chained Equation), elle consiste à bâtir un modèle conditionnel pour chaque variable de façon séquentielle jusqu'à la convergence. Le modèle univarié de chaque variable tient compte de la nature de celle-ci (continues, binaires, nominales).

2.4.2) Imputation multiple par équations chainées

Cette méthode est décrite, dans la littérature, sous plusieurs dénominations; MICE (Multiple Imputation by chained Equations), régressions séquentielles et Fully Conditionnal Specification (FCS). L'imputation par équation chaînée séduit par sa flexibilité puisqu'elle ne fait pas l'hypothèse d'une distribution multivariée normale (MVN). Elle n'est pas aussi contraignante que la méthode MVN puisqu'elle ne fait pas l'hypothèse d'une distribution conjointe. De plus, Le

modèle d'imputation n'est pas une seule distribution conjointe difficile à spécifier, mais plusieurs distributions conditionnelles. En effet, pour chaque variable incomplète, un modèle spécifique sera utilisé pour imputer les données manquantes de cette même variable. Proposée par (S. van Buuren, 2007; S. van Buuren et Groothuis-Oudshoorn, 2011) est une alternative à la méthode JM (ou distribution jointe) développée par (Schaffer, 1997). L'intérêt de cette méthode est de réduire un problème complexe de dimension k en k problèmes successifs qui utilisent les données complètes et les variables précédemment imputées. Le tirage dans la distribution conditionnelle de chaque variable se fait à l'aide de l'échantillonneur de Gibbs décrit auparavant.

L'idée de base est d'estimer des valeurs de remplacements pour les données manquantes de x1 à l'aide d'une régression, des observations sans données manquantes, de x1 sur x2 , …, xk. Puis estimer des valeurs pour les données manquantes de la variable suivante x2 à l'aide d'une régression des observations de x2 sur x1 ,…, xk et répéter ce processus de régressions séquentielles pour toutes les variables qui présentent des données manquantes. Les données manquantes de xn sont remplacées par des valeurs issues uniquement de tirages aléatoires à partir de la distribution prédictive à posteriori de xn. Ce qui signifie que chaque variable sera imputée en utilisant sa propre distribution conditionnellement aux autres variables auxiliaires. Ceci apporte plus de flexibilité puisque pour une variable binaire on peut choisir une régression logistique comme modèle d'imputation puis une régression multinomiale pour une autre variable catégorielle et finalement une régression linéaire pour une variable continue.

Elle est disponible notamment dans SAS "Proc MI FCS" et R " smcfcs, mice" et Stata "mice". L'analyste peut spécifier le modèle d'imputation qui sera utilisé pour imputer chaque variable. Le tableau 4 résume les méthodes qui peuvent être utilisées.

Tableau 4 : Les méthodes disponibles S. van Buuren et Groothuis-Oudshoorn (2011)

Méthode	**Description**	**Type de variable**
pmm	moyenne prédite par « appariement »	Numérique
norm	régression linéaire bayésienne	Numérique
norm.nob	régression linéaire non bayésienne	Numérique
Mean	moyenne marginale	Numérique
2L.norm	Modèle linéaire à deux niveaux	Numérique
logreg	régression logistique.	Binaire
Polyreg	régression logistique multinomiale	Nominal
Polr	Régression logistique ordinale	Ordinal
lda	analyse discriminante linéaire	Nominal
Sample	échantillonnage aléatoire à partir des données observées	Tout type de variable

2.5) Modèle d'imputation

L'imputation multiple peut être paramétrique lorsqu'elle fait appel à un modèle d'imputation ou non paramétrique si aucun modèle n'est spécifié. Les méthodes paramétriques se basent sur une série d'hypothèses qui ne sont pas toujours respectées. La spécification du modèle d'imputation est cruciale dans ce cas de

figure. La construction d'un tel modèle est complexe et pas toujours facile à cause de la portion de données qui manquent et la mixité type des variables (binaires, continue, catégorielles). L'imputation non paramétrique évite ces contraintes. En effet, les méthodes non paramétriques sont moins contraignantes et peuvent tenir compte des interactions qui sont parfois ignorées par les méthodes paramétriques. Le choix des variables à inclure, y compris les variables auxiliaires qui ne sont pas utiles pour les analyses ultérieures, est déterminant pour une imputation réussie. Il est important de mentionner que le modèle d'imputation doit inclure toutes les variables du modèle d'analyse.

Les auteurs recommandent d'adopter une stratégie inclusive qui consiste à inclure un nombre raisonnable de variables auxiliaires pour rendre l'hypothèse MAR plausible Collins, Schafer et Kam (2001). En effet, le fait d'inclure dans le modèle d'imputation des variables auxiliaires qui ne sont pas importantes pour l'analyse, mais qui sont fortement corrélées avec la variable à imputer va augmenter la capacité prédictive du modèle d'imputation. Ces variables auxiliaires joueront un rôle uniquement à l'étape d'imputation. Au minimum, toutes les variables qui seront utilisées dans l'analyse principale devront être incluses lors de la phase d'imputation. Selon Collins, Schafer et Kam (2001), si des variables du modèle sont ignorées, cela risque d'atténuer les associations avec les autres variables. L'adoption d'une stratégie libérale durant l'étape d'imputation est toujours une bonne stratégie qui ne risque pas de générer des estimations biaisées. De plus, à cause de la difficulté de respecter l'hypothèse MAR de nombreux auteurs recommandent d'adopter une stratégie inclusive; qui consiste à inclure des variables auxiliaires dans le processus d'imputation (Collins, Schafer et Kam, 2001; Enders, 2010). Les mêmes auteurs suggèrent que le fait d'inclure des variables aide à transformer un processus de DM qui est MNAR en MAR.

C'est le terme « congeniality » qui est utilisé dans la littérature pour décrire cet aspect de la relation entre le modèle d'imputation et le modèle d'analyse.

L'imputation multiple est dite « congenial » si le modèle utilisé pour imputer les données manquantes englobe toutes les variables du modèle d'analyse.

Par conséquent, il est toujours bénéfique d'élargir le nombre de variables auxiliaires à inclure (Collins et al., 2001). La sélection de ces variables doit se baser sur la corrélation avec la variable à imputer. On peut inclure des variables faiblement corrélées à la variable à imputer (Collins et al., 2001); toutefois, les avantages de la stratégie inclusive sont plus évidents lorsque les variables sont fortement corrélées (Enders 2010). Pour Enders (2010), les variables auxiliaires sont utiles lorsqu'elles sont fortement corrélées (r > .40) avec la variable d'intérêt. En plus de nous éviter d'ignorer des variables importantes liées au processus qui génère les données manquantes, la stratégie inclusive permet aussi de réaliser à faible coût des gains en efficience et en réduction de biais.

2.6) Règles pour combiner les résultats

Le processus d'imputation multiple va produire D bases de données complètes ou les données manquantes ont été remplacées par d valeurs plausibles. Ces bases de données peuvent être analysées pour produire les paramètres d'intérêts (moyenne, paramètres de régression etc.). Rubin (1987) établi les règles suivantes pour combiner les résultats d'analyse des D bases de données. Dans ce qui suit, nous allons décrire ces règles. Soit $\hat{\theta}_d$, d=1, …,D les estimateurs des paramètres d'intérêts obtenus en réalisant les analyses sur les D bases de données et W_d d=1, …,D les variances qui sont associées à ces estimateurs.

L'estimateur global de θ est donné par la formule suivante :

$$\bar{\theta} = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d \quad .$$

La variance associée à ces estimateurs des paramètres est composée de deux parties : la matrice de variance covariance intra-imputation, notée $\overline{W}_D$, qui est en réalité la moyenne des variances :

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d$$

et la matrice de variance covariance inter-imputation notée B_D :

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D} (\hat{\theta}_d - \hat{\theta}_D)^2$$

La variance totale, notée T_D , associée au paramètre $\bar{\theta}_D$ est donnée par la formule suivante :

$T_D = \bar{W}_D + (\frac{D+1}{D})B_D$

Où (1 + 1/D) est un facteur de correction pour un nombre fini d'imputation D.

La distribution de référence pour construire l'intervalle des estimations est la distribution de Student :

$$(\theta - \bar{\theta}_D)T_D^{-1/2} \sim t_v$$

Le dégrée de liberté est :

$$v = (D-1)\left(1 + \frac{1}{(D+1)}\frac{\bar{W}_D}{B_D}\right)^2$$

Chapitre 3 : Choix méthodologiques

3.1) Objectif de la recherche

L'objectif principal de notre étude et de tester les méthodes d'imputations disponibles et d'en identifier celles qui sont le mieux adaptées à la base de données NCDB. Rappelons que nous disposons d'une base de données qui contient à la fois des données catégorielles et numériques. De nouvelles méthodes d'imputation simples ont été développées récemment. Nous aimerions vérifier si ces techniques équivalent l'imputation multiple comme l'affirment ses auteurs. L'objectif secondaire est d'évaluer à quel points ces méthodes génèrent des valeurs plausibles qui ne dénaturent pas les associations logiques qui existent entre les variables. En effet ces relations caractérisent les observations et sont une illustration de la validité et de la cohérence des données. Nous avons parcourus un nombre important d'articles et nous avons noté que très peu d'auteurs apportent une attention à cet aspect. La technique d'imputation adoptée tout en préservant la distribution des données risque de remplacer des données manquantes par des valeurs qui n'ont pas de sens. Par exemple proposer une valeur de remplacement pour l'âge du conducteur qui est en bas de l'âge légal. En dernier lieu nous allons étudier les conséquences concrètes des méthodes d'imputations choisies sur un modèle analytique nous avons choisi le modèle de prédiction du risque de décès ou de blessure grave développé par Fredette et al (2008). Nous allons dans un premiers temps utiliser un jeu de données qui ne contient pas de DM puis nous allons introduire des DM. Certains auteurs précisent que le taux de DM peut varier de 1% à 67% de la base à notre disposition. Peugh et Enders (2004) confirme qu'il varie de 26% à 72%. A l'instar de Yiran Dong(2013) nous avons choisi d'introduire des taux de DM, selon la typologie MAR, de façon artificiel et ce à des taux de 20%, 40% et 60%. Nous allons par la suite appliquer les méthodes d'imputation sélectionnées puis procéder à des diagnostics afin de

vérifier la validité des données et finalement appliquer le modèle de Fredette et al. (2008) dans le but d'étudier l'impact des jeux de données imputées sur ce modèle.

3.2) Description de l'échantillon

Les données proviennent de la base de données NCDB qui contient des informations sur les collisions de véhicules rapportées chaque année à Transport Canada par les provinces et territoires nationaux. Les VIN (Vehicule Indentification Number) des voitures sont décodés dans cette base pour en extraire des informations relatives aux caractéristiques des véhicules impliqués dans l'accident. Les données NCD couvrent une période qui s'étend de 2000 à 2006 et contient 9 332 819 observations. Pour les besoins de l'étude et dans le but d'avoir un échantillon facile à exploiter, nous nous sommes limités aux deux dernières années disponibles 2005 et 2006 soit 2 699 838 unités au total. De cette base nous avons extrait les observations relatives aux accidents impliquant des véhicules qui concernent uniquement les conducteurs. Dans le but d'éviter les duplications, deux provinces qui utilisent le même numéro d'accident par exemple, nous avons jumelé le numéro d'identification de l'accident à l'année et la province pour former un nouvel identifiant des observations. Nous avons par la suite comptabilisé le nombre d'unités relatives à chaque accident et gardé uniquement celles qui concernent les conducteurs dans les accidents impliquant deux véhicules. A l'instar de Fredette et al. (2008) nous avons exclue les accidents impliquant des motos, bicyclettes, motoneiges et les véhicules tout-terrain.

3.3) Détection de la configuration des DM (Voir VIM)

Il convient de distinguer deux caractéristiques qui permettent de décrire la nature des données manquantes: l'une est appelée configuration « pattern », et l'autre a trait au processus ou mécanisme à l'origine de l'apparition des données manquantes. En effet le pattern est un paramètre important pour sélectionner la

méthode d'imputation. Nous avons pu établir que nos données ont une configuration arbitraire (voir annexe 2). Étant donné cet aspect qui caractérise nos données nous allons utiliser Proc MI FCS. Les conditions d'utilisation de la procédure « Proc MI » selon la configuration ont été décrite dans le Tableau 1.

3.4) Préparation préliminaires des données et examen des DM

Dans un premier temps nous avons examiné les variables qui composent la base pour en écarter celle qui contiennent trop de valeurs manquantes. Les variables qui n'ont pas été utilisé par Fredette et al. (2008) et qui peuvent potentiellement servir comme variables auxiliaires ne doivent pas à leurs tour avoir un taux élevé de DM. C'est la raison principale qui nous a conduits à exclure les variables qui présentent plus de 90% de DM.

Nous avons aussi procédé aux nettoyages des données pour éliminer les observations qui peuvent conduire à prédire de fausses valeurs de remplacement pour les DM. Plus de 2998 observations indiquent un âge inférieur à l'âge légal du conducteur et aucune expérience de conduite. Pour le besoin de cette étude nous avons supprimé ces observations pour éviter que ces observations ne contribuent à produire des valeurs de remplacement non plausibles

La variables V_Litre (cylindré) a été transformé en classes de dix niveaux afin de réduire le nombre de catégories. Les variables catégorielles de plus de 10 catégories sont problématiques avec toutes les méthodes d'imputations que nous avons exploré. Nous avons recodé la variable P_SAFE qui contenait plus de 13 catégories en deux catégories; 'Pas de Ceinture de sécurité', 'Ceinture de sécurité portée'. Cette variable présentait un taux de DM qui avoisine le 50% dans la base originale. Nous avons déduit des valeurs de remplacement aux DM en utilisant l'information disponible dans la base. Nous avons utilisé la variable P_EJCT qui donne l'information si la personne a été éjectée hors de la voiture lors de l'impact.

Ainsi pour les personnes qui ont été éjectés de la voiture nous avons attribué la valeur « Pas de ceinture de sécurité » pour celle qui ne l'ont pas été nous avons attribué la valeurs « Ceinture de sécurité portée ».

La variable dépendante utilisé par Fredette et al. (2008) pour modéliser le risque de décès ou blessure grave présentait un taux élevé de DM (26%). Cette variable donne la sévérité de la blessure (Pas de blessure, Blessure minime, blessure majeure, fatale, mineure). Nous avons croisé cette variable avec d'autres variables disponibles afin de déduire facilement une valeur de remplacement pour les DM. En effet, la combinaison des variables C_SEV; qui renseigne sur la sévérité de l'accident et la variable C_INJ; qui donne le nombre de blessé a permis de remplacer de nombreuse DM de la variable P_ISEV. Si le nombre de blessé est égal à zéro (C_INJ=0) et il n'y a que de des dégâts matériels (C_SEV=3) on peut facilement affirmer qu'il n y a pas eu de blessé. Nous avons été en mesure de réduire ainsi le taux de DM à moins de 1%. Il est préférable d'être déterministe et de déduire des valeurs de remplacement quand les données disponibles le permettent.

Nous avons dressé une liste de toutes les modalités de la variable V_Bodyt puis nous avons regroupé ces catégories en 5 grandes classes. Ces groupes sont présentés en détails dans le tableau 5.

Tableau 5 : Classification des types de véhicule à l'aide du code Polk.

Type de véhicule	Code extrait de la variable V_BodyT
'Car'	2D,2H,2L,2P,2T,3D,3P,4D,4H,4L,4P,5D,CP,CV,HR,HT,IN,LB,LM,NB,RD,SD,SW,C4,CG,S1,S2,RS,SW
'Pickup truck'	3B,3C,4B,4C,CB,FB,PK,PC,PS,YY,CH,CL,CY
'Minivan'	SV,VN,EC,ES,EV,EW,IC,IE,MP,MH,MY,VC,VW,SN
'SUV'	CV,LL,SW,UT,2W,4W,8V,CW,FC
'Heavy Truck'	AC,CC,CM,DP,DS,FB,FT,GG,GL,TB,TL,TM,BU

La vérification de la performance des méthodes d'imputation, sur des données réelles, est une opération délicate voire impossible car les valeurs que prenaient les DM sont inexistantes et on ne peut faire des comparaisons avec les valeurs plausibles générées par le processus d'imputation. Afin de contourner ce problème nous avons éliminé les observations avec des DM et restreint l'échantillon de départ aux cas complets. Nous avons par la suite introduit artificiellement des taux de DM, selon la typologie MAR. Pour réaliser cette tâche nous avons utilisé la fonction «ampute » développé par Schouten, Lugtig et Vink (2018). Cette fonction est intégrée maintenant au module « mice » de R. Nous avons choisi de créer trois jeux de données qui contiennent respectivement, 20% de DM, 40% de DM et 60% de DM.

3.5) Sélection des variables auxiliaires et construction des modèles d'imputations

Le choix des variables auxiliaires à inclure dans l'étape d'estimation des valeurs de remplacement des données manquantes est important pour l'imputation multiple. Rubin (1996) conseille d'inclure autant de variables que possible lors de l'imputation multiple, même celles qui ne sont pas utilisées dans l'analyse

principale subséquente. Ces variables auxiliaires joueront un rôle uniquement à l'étape d'imputation. Au minimum, toutes les variables qui seront utilisées dans l'analyse principale devront être incluses lors de la phase d'imputation. Selon Collins et al. (2001), si des variables du modèle sont ignorées, cela risque d'atténuer les associations avec les autres variables. Le rôle des variables auxiliaire est double en plus de fournir de bons prédicteurs elles contribuent à renforcer l'hypothèse MAR.

CART (Classification and Regression Trees) est un concept statistique introduit par L. Breiman *et al.* (1984) qui permet de sélectionner des prédicteurs par arbre aussi bien en régression qu'en classification. Les solutions obtenues sont représentées sous formes graphiques simples à comprendre et à interpréter. CART séduit par sa simplicité et c'est pour cette raison qu'il sera utilisé comme outil de sélection des variables auxiliaires. En effet, au lieu de sélectionner toute les variables qui apparaissent dans l'arbre ce qui risque de gonfler le nombre de variable sélectionnées ou d'attribuer de l'importance à des variables qui ne sont pas critiques, une approche basée sur l'importance de la variable sera privilégiée comme critère de sélection. La sélection des variables se réduit donc à inclure les variables qui jouent un rôle dans la prédiction de la variable d'intérêt. Ce processus est appliqué de façon séparée sur chaque variable du modèle Fredette et al. (2008) . Le tableau 6 contient la liste des variables utilisées par Fredette et al. (2008) pour modéliser le risque de décès ou de blessure grave ainsi que les variables qui peuvent les prédire et qui leur sont fortement associées. Ce travail nous a servi à bâtir les modèles d'imputations qui vont servir à imputer les DM à l'aide SAS. Ces modèles sont résumés dans le Tableau 6.

Tableau 6: Modèles utilisés pour imputer les données à l'aide de "proc mi"

Variable dépendante	Description	Variable indépendante	Modèle utilisé
P_age	Age du conducteur	P_YLIC_N1,C_LITE,V_SEG C_Hour_N, V_WHLB1_N,V_NOCC, P_ID_N, V_BSWL_N, V_DISPL_N, P_sex, V_TYPE, V_CF2, C_HRUN, C_ALITE, P_USE, V_CF1,V_TRANST, V_DISPI_N, P_PSN,V_TRLR V_DSEV	RegPMM (Heitjan et Little, 1991)
Authorized Speed	Vitesse affichée	C_RCL1,C_RMTL,C_SCATT,C_PROV,C_RCFG,V_DIR	Régression logistique
V_BSWL_N	Poids du véhicule	V_SEG,V_DISPL, V_VTYPE V_CAB, V_MYEAR_N, V_GVWC, V_TRANS, V_WHLB2_N1,_DISPI_N,Driver_Vehicle_Type, V_RAXLE,V_WHEELS, V_DISPL_N V_WHLB1_N, V_BLOCK, V_WHEELD V_SECUR, V_RESTR, V_YEAR_N, V_FAXLE	RegPMM
Safety_Belt	Ceinture de sécurité	V_TYPE, P_EJCT, V_LICJ, P_PUSE, P_PSN, V_CAB, Major_or_Fatal, P_DLIC,V_EMER, Driver_Vehicle_Type V_CF1, C_SCATT, V_EVT1, V_DSEV, Authorized_Speed, V_VTYPE	Régression logistique

RegPMM est une approche d'imputation des variables continues à l'aide de régression à l'exception que cette méthode va attribuer une valeur de remplacement existante qui sera choisie parmi un ensemble d'observations qui ont des valeurs prédites proches pour la variable d'intérêt. La valeur proposée pour remplacer la DM sera sélectionnée de façon aléatoire à partir de cet ensemble d'observation.

En plus de ce rôle dans la construction des modèle d'imputation, les variable avec les plus grand score de prédiction seront utilisées par le module R « KNN » pour calculer les distance et avoir une meilleur identification des voisins proches.

Tableau 7 : Liste des variables avec leur pourcentage respectif de DM.

variable	Description	% de DM	Rôle de la variable
C_CONF	Collision Configuration	0%	Fredette et al.
C_POLC	Police Detachment/Region Code	0%	Auxiliaire
C_SEV	Severity of Collision	0%	Fredette et al.
P_ISEV	Medical Treatment Required	0%	Fredette et al.
C_Hour	Hour of Collision	0%	Auxiliaire
P_AGE	Person Age	0%	Fredette et al.
C_Conf	Collision Configuration	0%	Fredette et al.
P_AGE	Person Age	0%	Fredette et al.
C_RSUR	Road Surface	1%	Auxiliaire
V_MNVR	Vehicle Manoeuvre	1%	Auxiliaire
V_MYEAR	YEAR MODEL	1%	Auxiliaire
V_DIR	Direction of Travel	1%	Auxiliaire
V_DISPI	CUBIC INCH DISPLACEMENT	1%	Auxiliaire
C_INJ	Number of Persons Injured	1%	Auxiliaire
P_YLIC	Years Licensed in Jurisdiction	4%	Auxiliaire
P_SEX	Person Sex	4%	Fredette et al.
P_LICS	Licence Status	6%	Auxiliaire
P_SAFE	Safety Device Used	6%	Fredette et al.
V_SEG	SEGMENTATION CODE	16%	Auxiliaire
V_WHEELD	DRIVING WHEELS	16%	Auxiliaire
V_WHLB1	WHEEL BASE OF SERIES	16%	Auxiliaire
V_BSWL	BASE SHIPPING WEIGHT IN LBS	20%	Fredette et al.
V_BLOCK	ENGINE BLOCK TYPE	20%	Auxiliaire
V_BODYT	BODY TYPE	20%	Fredette et al.
V_LITRE	ENGINE LITRES	55%	Auxiliaire
V_HEAD	ENGINE HEAD CONFIGURATION	55%	Auxiliaire
C_RMTL	Road Material	75%	Auxiliaire
C_RCL1	Road Classification I	75%	Auxiliaire
C_TRAF	Traffic Control	75%	Auxiliaire
C_RCFG	Roadway Configuration	76%	Auxiliaire
C_SPED	Posted Speed Limit	77%	Fredette et al.
C_ALITE	Artificial Light Condition	78%	Auxiliaire
V_WHEELB	WHEEL BASES	88%	Auxiliaire

3.6) Méthodes d'imputations utilisées

Nous allons nous servir de l'imputation multiple et remédier au DM de l'échantillon à l'aide de la procédure SAS « proc MI » qui intègre la méthode « FCS ». En effet cet outil d'imputation multiple est adapté à nos données complexes qui contiennent à la fois des variables continues et des variables catégorielles. Il est possible de spécifier un modèle d'imputation pour chaque variable. Nous avons par la suite fourni des modèles d'imputations en se servant des résultats obtenus à l'aide de CART. Ces scores (Variable Importance) qui représentent la contribution de chaque variable indépendante à prédire la variable d'intérêt nous ont servi à la fois pour renforcer l'hypothèse MAR et pour bâtir les modèles de prédiction des variables du modèle (Fredette *et al.*, 2008).

De nombreux packages R sont disponibles et peuvent aider dans le processus d'imputation. Après en avoir essayé un bon nombre (NPBayesImpute, mice, mi, YaimputeR, mix, BabooN, HotDeck, DMwr, HotDeck, BBPMM) notre choix s'est arrêté sur quelques modules (package R) qui ont bien fonctionné avec nos jeux de données. Nous nous somme intéressé à en tester plusieurs qui sont porteurs de plein d'avantages, le fait de pas faire d'hypothèse sur la distribution par exemple, pour l'analyste qui cherche à imputer les données. Certain packages ne peuvent être utilisé qu'avec une matrice numérique ce qui va nous obliger à dénaturer les données. D'autre n'ont pas résisté à la nature multidimensionnelle complexe de notre jeu de données. D'autre étaient instable et produisait des erreurs si on refait l'essai avec un nouveau jeu de données.

La composante KNN du module VIM permet de déterminer une valeur d'imputation parmi les plus proches voisins de l'unité. Une distance entre les unités est calculée pour faire la sélection des observations candidates.

MissForest est le deuxième module R qui sera utilisé. Ce composant de R utilise les arbres de régression et les forêts aléatoires pour imputer les DM. Ce package,

développé récemment, permet de faire de l'imputation simple en produisant un seul jeu de donnée.

Le troisième module R qui sera utilisé est le MissMda. Ce module utilise les composantes principales pour sélectionner une valeur de remplacement. Le module MissMDA contient un module pour faire l'imputation multiple mais malheureusement il n'est pas adapté à notre jeu de donnée mixte. Nous avons utilisé la composante de ce module qui peut traiter des bases de données mixtes malgré qu'elle produise un seul jeu de données et qu'elle fait de l'imputation simple. Nous avons abandonné l'utilisation de ce module à cause de son instabilité.

Ces modules de R ont été rendu disponibles récemment. Ils présentent l'avantage d'être non paramétriques et permettent d'éviter de poser des hypothèses, difficile à vérifier, sur la distribution et la construction de modèle d'imputation complexe surtout en présence de plusieurs type de variables. Bien que ces trois module MissForest, MissMDA et KNN font de l'imputation simple, il sera intéressant de voir comment ils performent par rapport à l'imputation multiple. MissMDA a été abandonné car il n'offrait pas la stabilité recherché et il produisait des erreurs quand on essayait de changer de jeu de données.

3.7) Diagnostic des données imputées

Comme dans la plupart des articles scientifiques qui traitent de l'imputation multiple nous allons évaluer les techniques décrites en comparant les paramètres des modèles en utilisant les données complètes et les données imputées. Certains méthodistes suggèrent de procéder à des diagnostics graphiques et numériques (Raghunathan 2007; Abayomi 2011). Les auteurs réclament que l'imputation multiple doit générer des valeurs de remplacement plausibles. Les valeurs de remplacement doivent être des unités qui auraient pu exister si les données étaient

complètes. Les valeurs qui n'ont pas de sens (père enceinte, bébé fumeurs, cycliste avec ceinture de sécurité …) ne doivent pas exister dans une base après l'IM. Dans notre contexte on ne s'attend pas à ce que des valeurs imputées pour l'expérience du conducteur qui serait incompatible avec l'âge du conducteur ou attribuer « ceinture de sécurité » à des BUS.

Chapitre 4 : Résultats

Dans le présent chapitre, nous allons procéder à la vérification des jeux de données produits par les différentes méthodes d'imputation et ainsi présenter les résultats. Pour cela, nous allons utiliser à la fois la visualisation graphique et les contrôles numériques. Le but de ces opérations de diagnostic est de vérifier l'intégrité des jeux de données. De ce fait, nous pourrons voir si les relations logiques sont préservées et si les données produites ne contiennent pas des valeurs qui n'ont pas de sens et qui risquent d'induire à de fausses conclusions analytiques. En effet, l'objectif est de savoir si l'imputation sera en mesure de préserver la distribution des données tout en respectant les relations logiques qui caractérisent les données originales et qui sont la preuve de leur cohérence et leur validité. Puis, nous allons user du modèle Fredette et al. (2008) pour étudier l'impact de l'imputation sur la suite d'une analyse.

Le ratio de la masse est d'une importance cruciale pour le modèle de Fredette et al. (2008).

Ainsi, on pourra poser les questions suivantes :

Est-ce que la distribution du poids du véhicule sera préservée?

Est-ce que le poids selon le type de véhicule est cohérent avec la base originale?

Est-ce que les limites inférieures et supérieures du poids du véhicule seront maintenues ?

Pour ce qui est de la variable « âge du conducteur », nous avons fait un travail préliminaire pour éliminer les valeurs aberrantes. Le but est de savoir si la limite inférieure de l'âge du conducteur a été préservée et n'a pas été réduite. Dans ce sens, nous avons choisi de garder cette variable sous sa forme continue et ne pas la transformer en classe afin de l'utiliser pour tester le processus d'imputation.

Si la base d'origine ne contient pas de jeunes conducteurs de camions lourds, est ce que la base complétée par l'Imputation Multiple va en contenir ? Est-ce qu'il y a des chauffeurs d'autobus et de camion lourd qui ont moins de 20 ans?

En ce qui concerne la variable « blessure grave ou décès » nous cherchons à savoir si les proportions ont été préservées. L'étude Fredette et al. (2008) a conclu alors que les voitures offrent moins de protection aux conducteurs quand la deuxième voiture impliquée dans l'accident est de masse supérieure.

La question qui se pose est : est-ce que les proportions de ceux qui ont subi des blessures graves ou fatales seront maintenues dans les jeux de données générés par le processus d'imputation?

L'objet de ce travail de diagnostic est de s'assurer que les jeux de données résultant de l'Imputation Multiple ne risquent pas de contenir des valeurs aberrantes et d'engendrer des fausses conclusions.

Pour répondre à ces interrogations, nous allons utiliser des résumés graphiques et numériques qui seront présentés ultérieurement. Nous allons aussi utiliser la moyenne arithmétique et la variance pour nous assurer que les paramètres de l'échantillon sont bien maintenus. Et finalement nous allons nous servir des paramètres de position (Médiane, Centile…) en guise de second contrôle des représentations graphiques.

Pour ce qui est des variables nominales, nous nous servirons des proportions pour la vérification du maintien des distributions.

4.1) Diagnostic graphique

Dans cette section, nous allons aborder les distributions des variables qui ont été utilisées par Fredette et al. (2008) pour modéliser le risque de blessure grave ou décès du conducteur. Nous allons mettre le point sur les variables qui représentent un taux très élevé de Données Manquantes dans la base originale afin de nous assurer que les distributions sont préservées.

En principe, nous avons choisi de présenter les distributions des variables suivantes :

- Poids du véhicule

- Vitesse affichée
- Type de véhicule

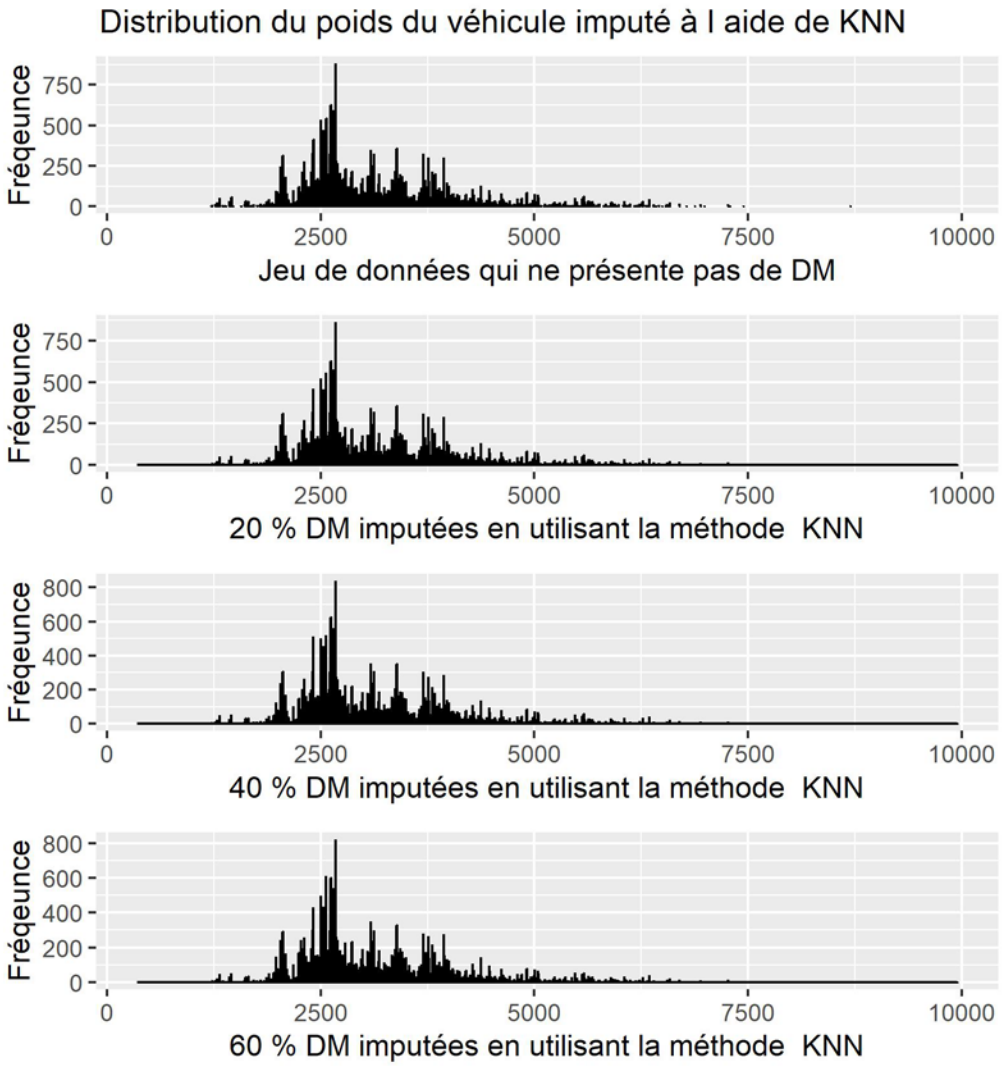
Ces variables avaient respectivement 37%, 49% et 25% de Données Manquantes dans la base initiale. Les représentations graphiques des autres variables peuvent être consultées dans les annexes.

Pour ce qui est des résumés graphiques, nous allons les utiliser pour vérifier si les distributions n'ont pas été dénaturées. Nous contrôlerons aussi qu'elles n'ont pas subi des changements causés par le taux élevé de Données Manquantes.

Il faut rappeler que nous avons choisi d'introduire des taux artificiels de Données Manquantes (20%, 40% et 60%) dans le but de vérifier la performance des méthodes d'imputation choisies.

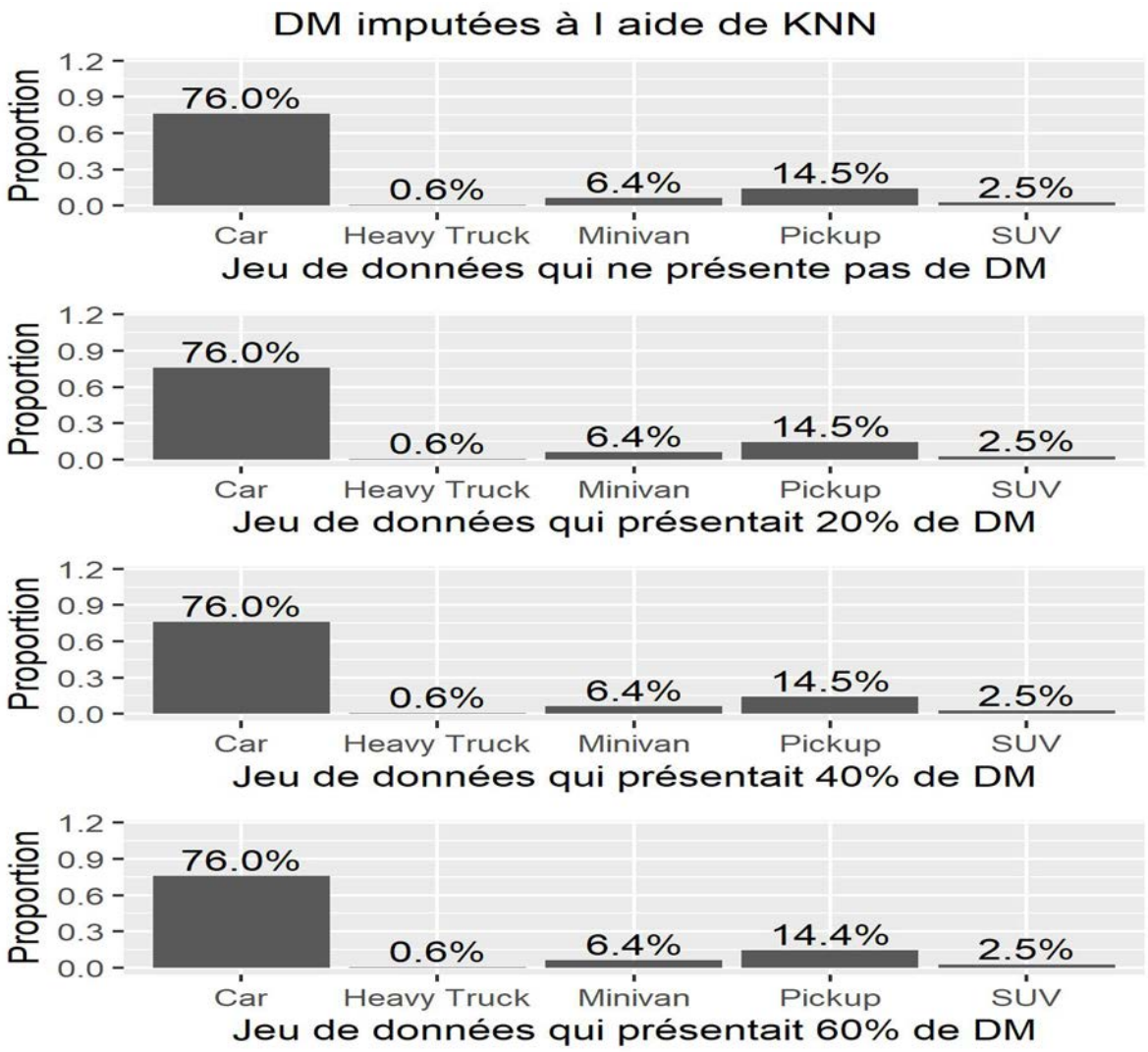
4.1.1) Représentation graphiques: DM imputées selon l'approche KNN

Figure 3 : Distribution du poids de véhicule (DM imputées à l'aide de KNN)



La Figure 1, qui résume graphiquement la distribution de la variable « poids du véhicule » dans les différents jeux de données, montre qu'il y a une grande similitude entre le jeu de données qui ne présente pas de données manquantes et celui avec 20%, 40% et 60% de données manquantes imputées. Nous allons procéder à la vérification de ce constat à travers d'autres contrôles.

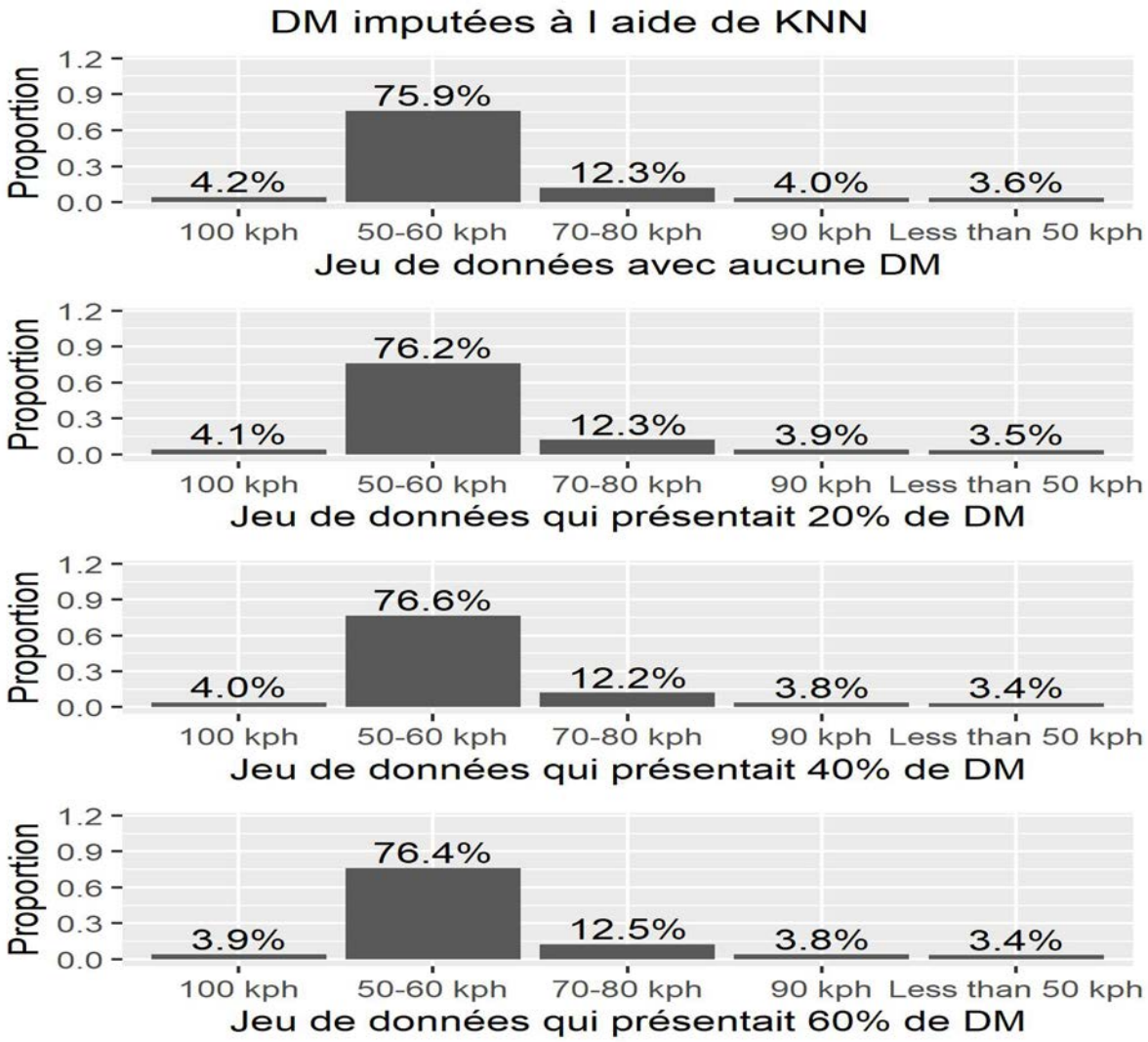
Figure 4: Distribution de la variable « type de véhicule » (DM imputées à l'aide de KNN)



En examinant la Figure 4 qui concerne la variable « type de véhicule », on peut constater que la distribution demeure sensiblement la même quel que soit le taux de données manquantes présent avant l'imputation.

On note aussi que les voitures sont les plus représentées dans l'échantillon de données. Ainsi, ces proportions sont maintenues dans les jeux de données imputés.

Figure 5: Distribution de la vitesse affichée (DM imputées à l'aide de KNN)

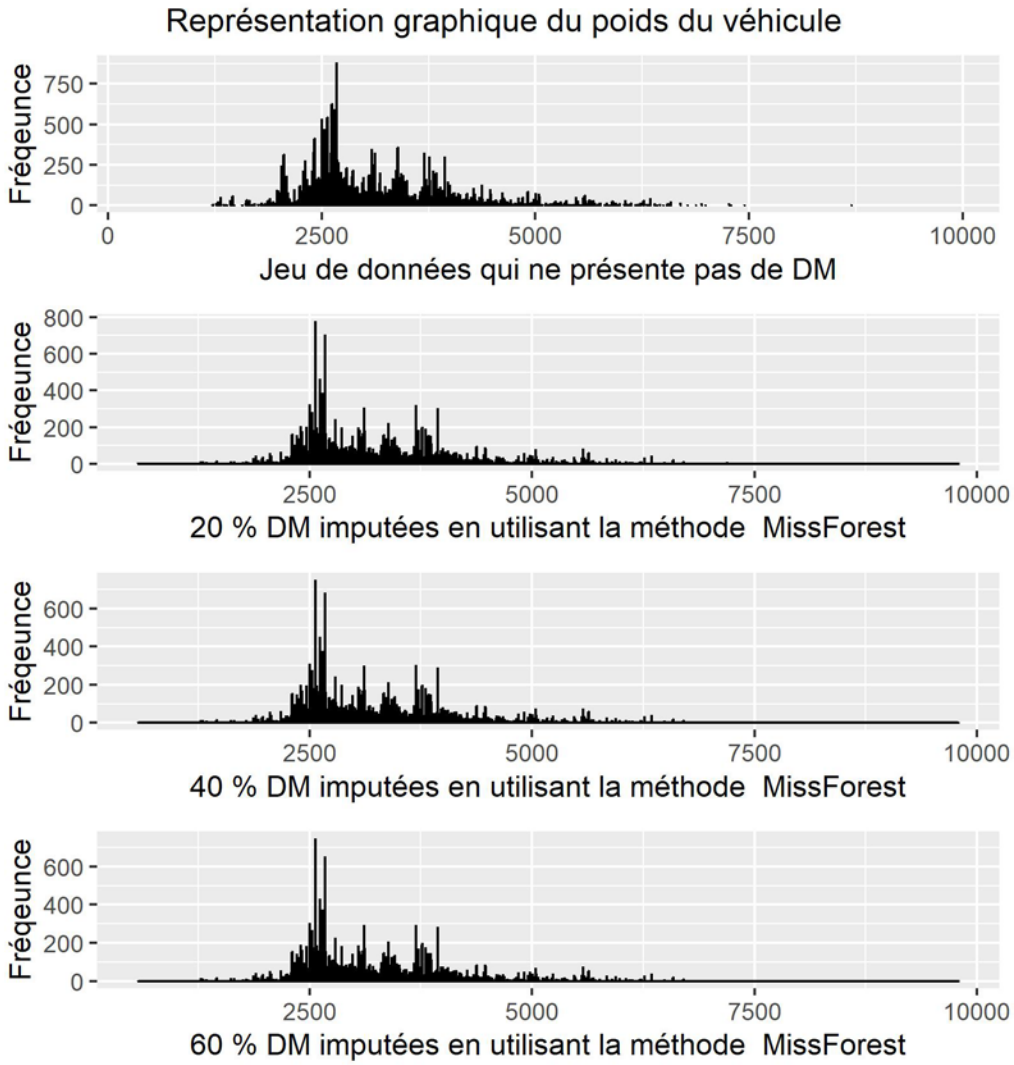


A la lumière de la Figure 5, on peut noter que la distribution de la variable « vitesse affichée » est sensiblement la même et ce quel que soit le jeu de données utilisé pour générer cette représentation graphique. Nous allons vérifier ce constat en nous servant des proportions avant et après imputation.

On constate aussi que la majorité des accidents, de l'échantillon de données, se sont passés dans des zones ou la vitesse affichée est entre 50 et 60 km/h.

4.1.2) Représentation graphiques: DM imputées selon l'approche MissForest :

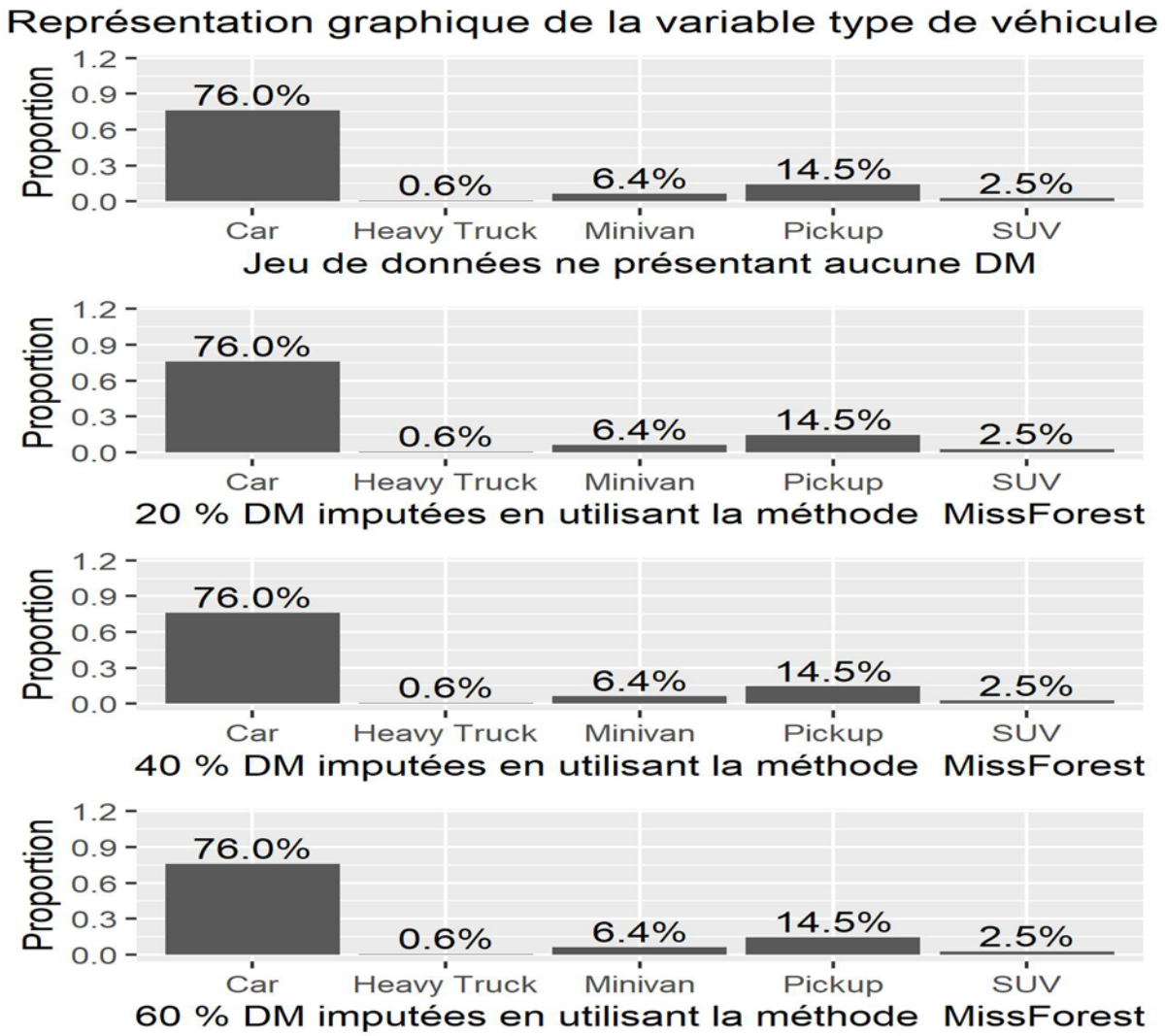
Figure 6 : Distribution du poids de véhicule (DM imputées à l'aide de MissForest)



La Figure 64 montre que l'imputation à l'aide du package MissForest donne de très bons résultats et que visuellement on peut affirmer que la distribution de la variable « poids du véhicule » a été préservée et reste identique à celle qui

prévalait avant d'introduire artificiellement des Données Manquantes dans le jeu de données initiale. Nous allons confirmer cette affirmation à l'aide des mesure de position notamment les centiles.

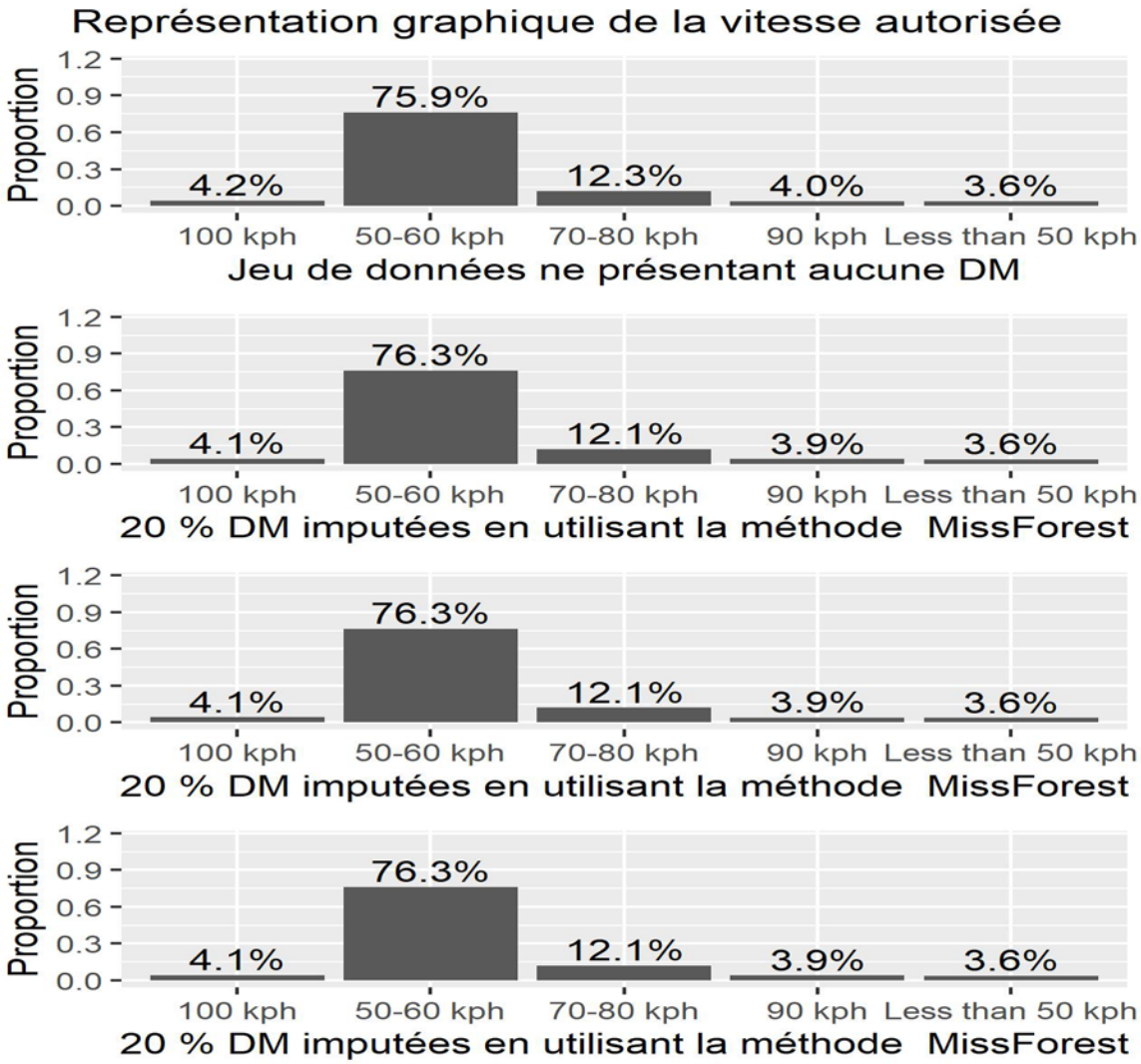
Figure 7 : Distribution de la variable « type de véhicule » (DM imputées à l'aide de MissForest)



La Figure 7, qui est une représentation graphique de la variable « type de véhicule », montre que, dans ce cas-ci aussi, dans ce cas également, les distributions sont préservées et que le taux de DM ne semble pas avoir d'impact puisque les proportions de chaque type de véhicule restent les mêmes.

On peut constater que les voitures sont prédominantes dans le jeu de données original et que cette proportion se retrouve dans les jeux de donnée imputée.

Figure 8 : Comparaison des distributions de la vitesse autorisée pré et post-imputation



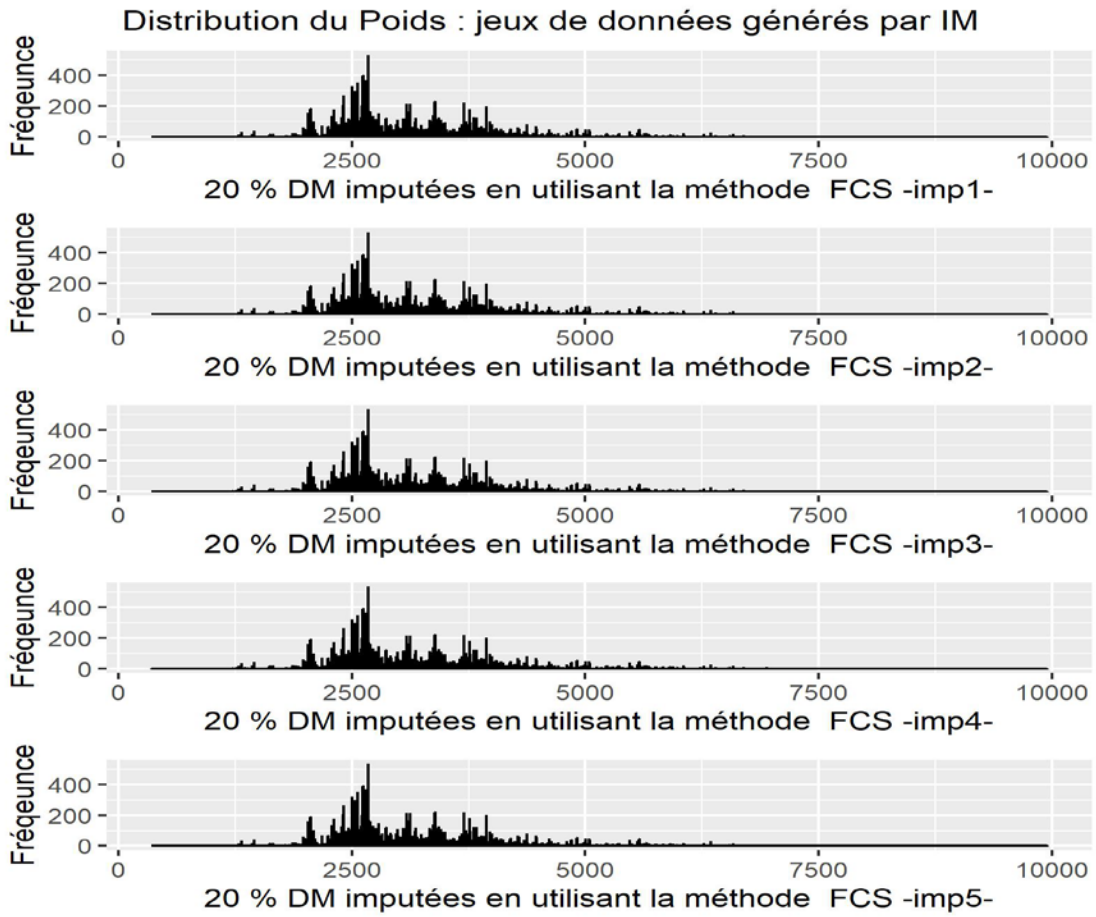
Si on s'intéresse à la Figure 8, on peut noter que le module, Missforest, qui a été utilisé pour imputer les DM préserve les distributions. La distribution est visuellement identique quel que soit le jeu de données qui a été utilisé pour la produire. Les proportions de

Le taux de DM (20%,40%,60%) ne semble pas avoir d'impact sur la distribution de cette variable.

4.1.3) Représentation graphiques: DM imputées selon l'approche FCS

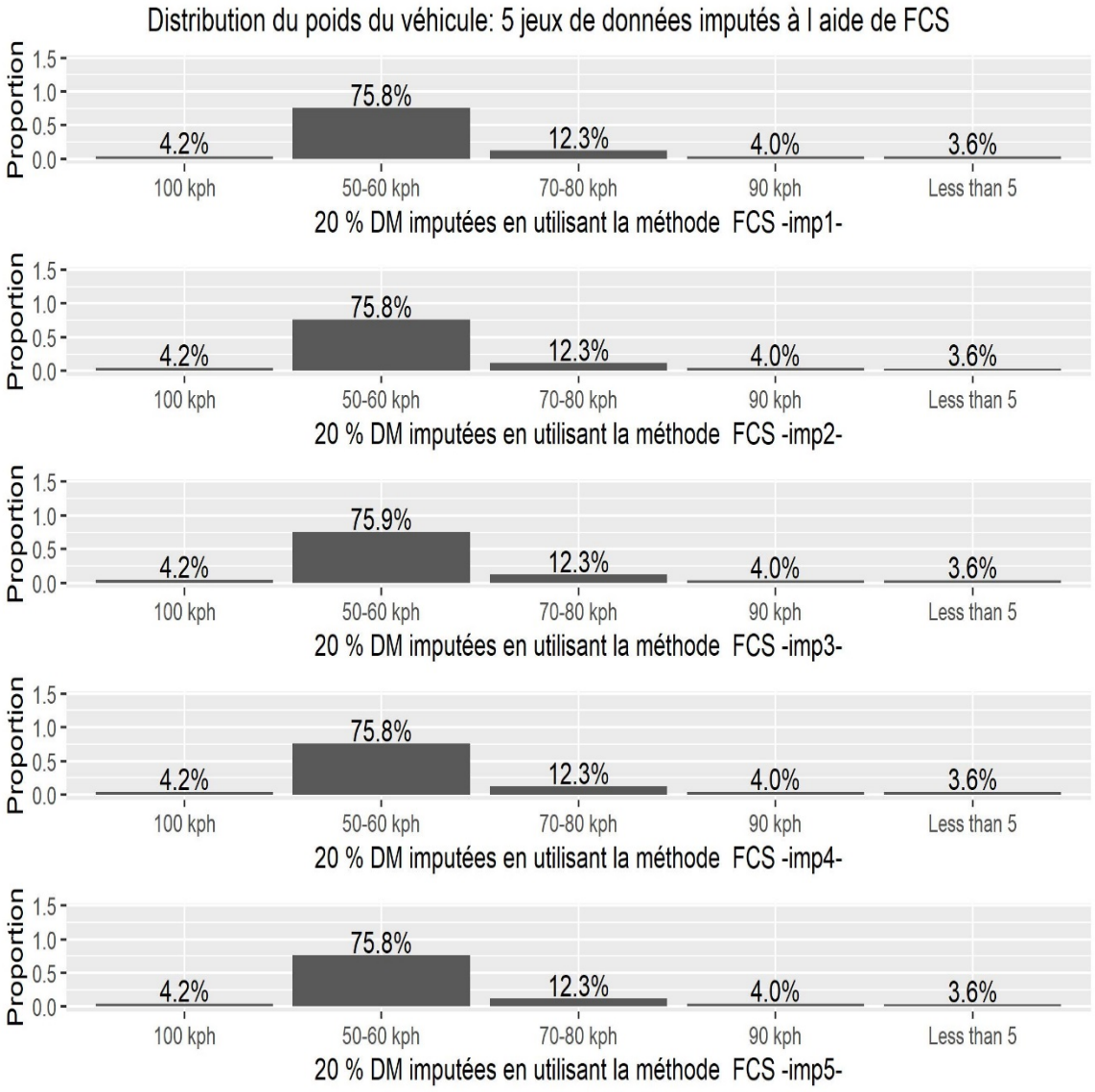
Les figures, figure 9 et la Figure 10, qui résument les distributions des jeux de données obtenus par l'imputation multiple, nous confirment que les différents jeux de données ont des distributions similaires. Pour ce qui est de la suite des représentations graphiques nous allons nous limiter à un seul jeu de données. Tous les jeux de données imputées seront utilisés pour les autres vérifications numériques. Les autres graphiques seront ajoutés aux annexes.

Figure 9 : Distribution du poids du véhicule (IM en utilisant FCS)



La distribution de la variable « poids du véhicule » est identique dans les différents jeux de données produits par l'imputation multiple. Cependant, le jeu de données présenté avait un taux de DM, introduit artificiellement, de 20%.

Figure 10: Distribution de la variable types de véhicule (DM imputés à l'aide de FCS)

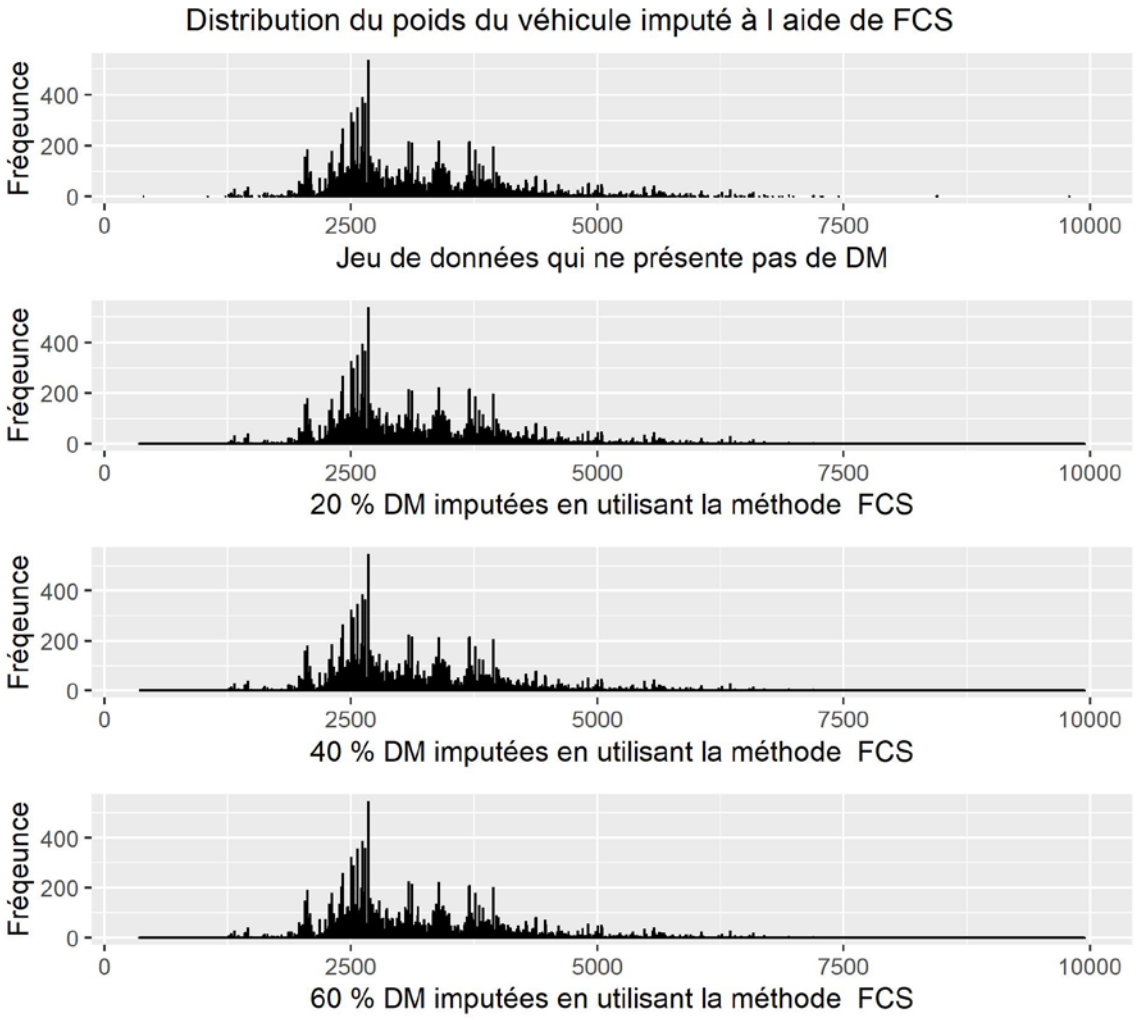


En analysant la Figure 10, on constate que les proportions que représente chaque type de véhicule sont identiques dans les différents jeux de données générés par l'Imputation Multiple.

Nous allons nous limiter à la présentation graphique d'un seul jeu de données en ce qui concerne l'imputation multiple.

Représentation graphiques à partir du jeu de données numéro 3 généré par l'imputation multiple.

Figure 11 : Distribution du poids du véhicule pré et post-imputation (DM imputées –FCS-)

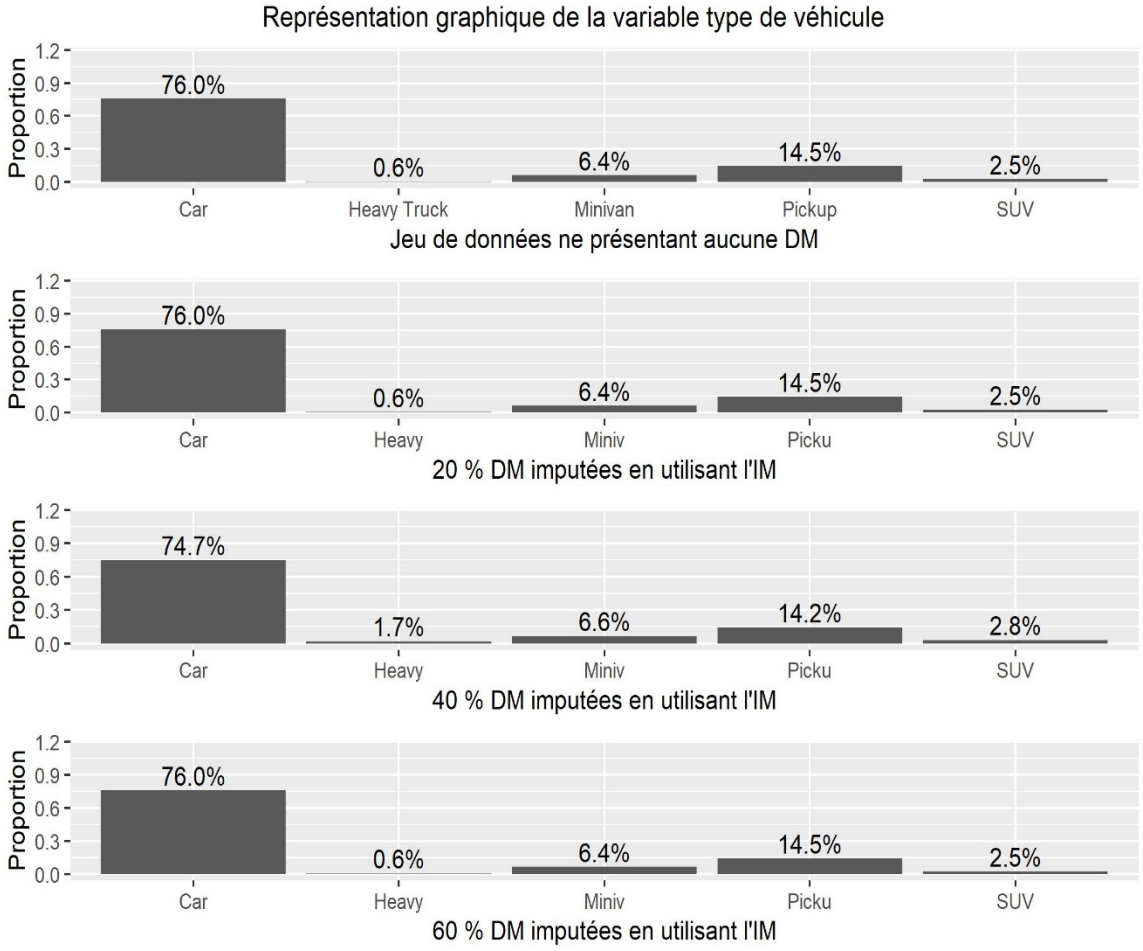


La Figure 11 résume les représentations graphiques des distributions de la variable « Poids du véhicule » dans les différents jeux de données pré et post-imputation.

On peut remarquer que la distribution des données, qui ont été imputées à l'aide du Fully Conditional Specification, a été préservée même en présence d'un taux élevé de Données Manquantes.

Visuellement, on peut dire que le processus d'imputation semble être robuste à la variation des taux de Données Manquantes. Quant aux jeux de données qui contenaient un taux important de Données Manquantes, ils semblent regagner les caractéristiques de la distribution d'origine.

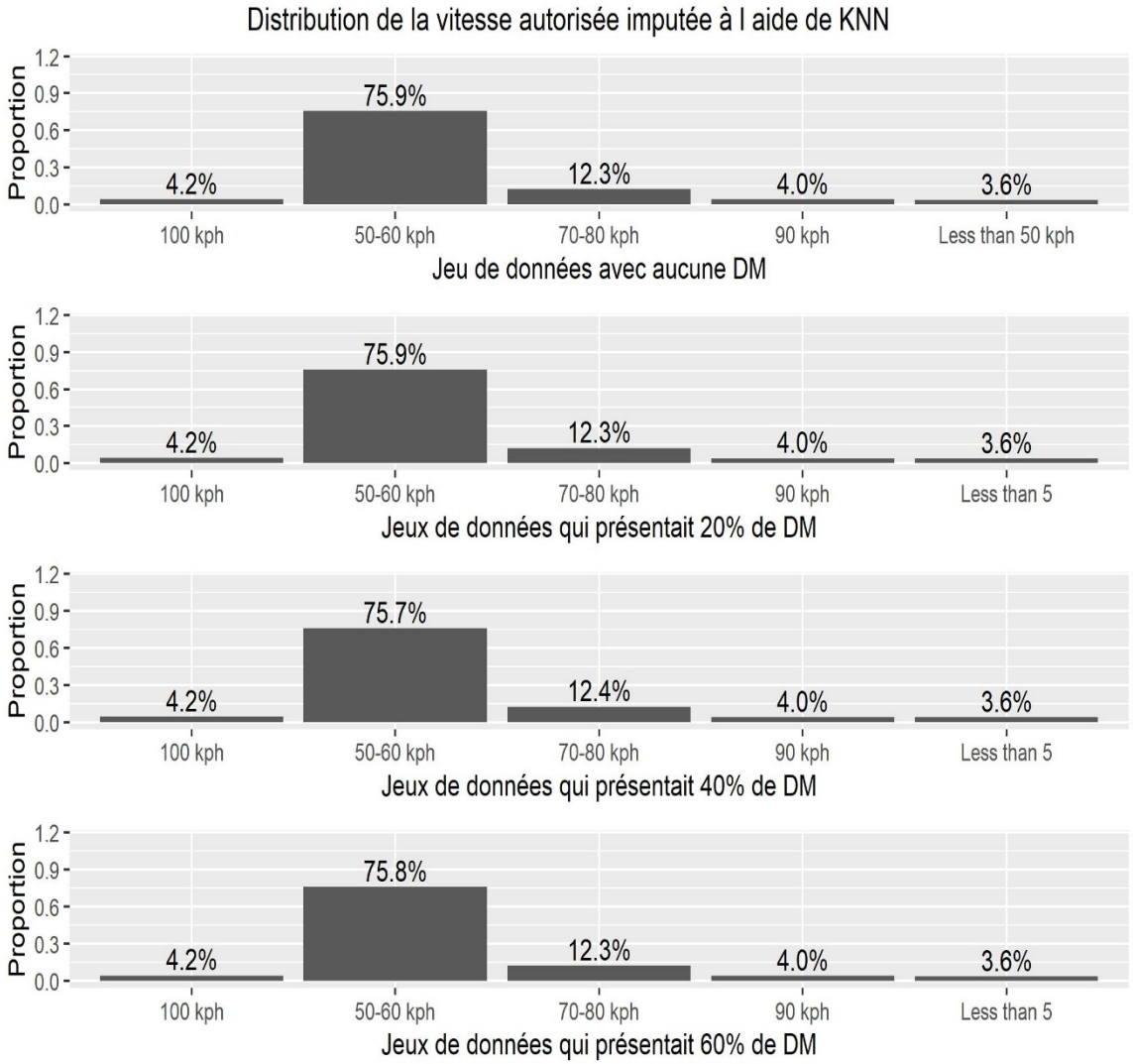
Figure 12 : Représentation graphique de la variable « type de véhicule » pré et post-imputation.



Focalisons nous maintenant sur les jeux de données imputés à l'aide de SAS suivant l'approche Fully conditional specification, on peut noter ainsi, une grande

similitude des différentes représentations graphiques (Figure 12 et Figure 13). On remarque que dans ce cas aussi les distributions ont été préservées et on retrouve les mêmes proportions d'un jeu de données à l'autre.

Figure 13 : Distributions de la vitesse autorisée pré et post-imputation (DM imputées à l'aide de FCS).



4.2) Diagnostic numérique

Dans cette section, nous allons tenter de confirmer les observations faites lors de l'analyse des représentations graphiques. Nous allons mener ce processus en deux étapes. Dans un premier temps nous allons nous intéresser à la totalité de la base qui a été générée par les différents processus d'imputation pour vérifier les distributions. Nous utiliserons les paramètres de distribution (moyenne, variance) et les mesures de position pour vérifier que les distributions des jeux de données sans DM et les jeux de données imputées à l'aide des approches sélectionnées sont identiques comme semble le montrer les représentations graphiques. Dans une deuxième étape nous allons identifier les observations qui contiennent des DM et récupérer les anciennes valeurs pour faire des comparaisons entre les valeurs initiales et les valeurs de remplacement générées par les processus d'imputation. Rappelons que l'échantillon sélectionné à partir de la base NCDB ne contenait aucune DM nous y avons introduit de façon artificielle des DM pour pouvoir mener ce genre de vérification.

4.2.1) Vérification de la distribution de tout l'échantillon

Le Tableau 7 montre que les paramètres de distribution des valeurs prises par la variable « Poids du véhicule » sont identiques d'un jeu de données à l'autre. Ce constat confirme ce qui a été observé en analysant les résumés graphiques. La distribution a été préservée par les différents modules d'imputation expérimentés. Les taux élevés de DM ne semblent avoir aucune influence sur la distribution.

Les paramètres des jeux de données imputées à l'aide de l'imputation multiple ont été agrégés selon les règles proposées par Rubin(1987) et décrite en détail dans le chapitre 2.

Tableau 8 : Statistiques descriptives de la variable poids du véhicule

Analysis Variable : V_BSWL_N						
Jeu de données	**% de DM**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
Données originales		87132	3202.54	808.97	192.00	9947.00
Données imputées en utilisant le module VIM(KNN) de R.	20%	87132	3202.00	807.74	192.00	9947.00
	40%	87132	3202.52	807.76	192.00	9947.00
	60%	87132	3202.32	806.35	192.00	9947.00
DM imputées à l'aide du module *missForest* de R	20%	87132	3202.39	807.91	192.00	9947.00
	40%	87132	3202.41	806.89	192.00	9947.00
	60%	87132	3202.49	805.60	192.00	9947.00
DM imputées à l'aide proc mi (FCS)	20%	87132	3202.11	808.06	192.00	9947.00
	40%	87132	3202.86	810.65	192.00	9947.00
	60%	87132	3202.88	809.20	192.00	9947.00

La variable « poids du véhicule » est au centre de l'étude de Fredette et al. (2008) qui a établi que la différence de masse joue en faveur ou en défaveur du

conducteur dépendamment du rapport entre les masses des deux véhicules. C'est pour cette raison que nous allons passer à un deuxième processus de vérification et utiliser les quintiles pour confirmer le constat observé dans les représentations graphiques des jeux de données pré et post imputation. Le jeu qui présentait 40% et le seul qui sera présenté dans le Tableau 9.

Tableau 9:Paramètres de la distribution de la variable "Poids du véhicule"

Niveau	Jeux de données sans aucune DM	DM(40%) imputées à l'aide de MissForest	DM(40%) imputées à l'aide de KNN	DM(40%) imputées à l'aide de proc Mi
100% Max	9947	9947.00	9947	9947
99%	5742	5698.29	5720	5720
95%	4700	4700.04	4701	4701
90%	4194	4194.00	4186	4194
75% Q3	3675	3675.00	3676	3675
50% Median	3085	3085.00	3085	3085
25% Q1	2606	2606.00	2605	2606
10%	2388	2388.00	2387	2388
5%	2242	2242.00	2242	2242
1%	1895	1895.00	1895	1895
0% Min	192	192	192	192

Le Tableau 9 confirme que les distributions sont identiques puisque les quintiles calculées sont similaires d'un jeu de données à l'autre. Rappelons que dans ce cas aussi les mesures qui concernent les jeux de données générés par l'imputation multiple ont été agrégées pour respecter les règles de Rubin (1987).

Comme on peut le constater la distribution a bel et bien été préservée par les différentes approches d'imputation utilisées et ce malgré des taux de Données Manquantes très élevés.

Si ensuite nous nous intéressons aux mesures intra-groupes, nous allons constater que parfois les limites inférieures ou supérieures du poids selon le type de véhicule ne sont pas respectées. Le Tableau 10 montre que le poids maximal a été presque multiplié par deux, pour les voitures, dans le jeu de données qui contenait 40% de Données Manquantes et qui a été imputé à l'aide de l'approche KNN.

Tableau 10 : distribution du risque de blessure selon le type de véhicule

	Blessure grave ou fatale=1	**Car**	**Heavy Vehicle**	**Minivan**	**Pickup**	**SUV**	**Total**
Jeux de données sans aucune DM	**0**	74.35	0.62	6.31	14.26	2.52	98.06
	1	1.61	0.01	0.08	0.21	0.02	1.94
DM(40%) imputées à l'aide de KNN	**0**	74.38	0.61	6.30	14.25	2.52	98.06
	1	1.61	0.01	0.08	0.21	0.02	1.94
DM(40%) imputées à l'aide de MissForest	**0**	74.35	0.61	6.31	14.26	2.52	98.06
	1	1.61	0.01	0.08	0.21	0.02	1.94
DM(40%) imputées à l'aide de proc Mi	**0**	74.32	0.64	6.32	14.25	2.53	98.06
	1	1.61	0.01	0.08	0.21	0.02	1.94

Concernant la variable dépendante binaire « blessure grave ou fatale », nous nous sommes intéressés à sa répartition selon le type de véhicule pour nous assurer que les proportions observées dans les jeux de données sans Données Manquantes sont reconduites dans les jeux de données imputés. On peut constater de légères différences. Dans le jeu de données sans Données Manquantes on observe que ce sont les conducteurs des voitures, de faible masse en générale, qui ont subi des blessures graves ou fatale. Cette proportion a été préservée dans les jeux de données imputés. Dans ce cas aussi les règles de Rubin (1987), décrite au chapitre

2, ont été appliquées pour agréger les proportions des 5 jeux de données générés par l'imputation multiples. Le Tableau 10 résume nos propos.

Tableau 11 : Distribution du poids selon le type de véhicule

Jeux de données	Type de véhicule	N	Mean	Std Dev	Minimum	Maximum
Données originaux avant de procéder à l'imputation	Car	66184	66184	2942.67	603.48	192.00
	Heavy vehicle	549	549	4602.21	1555.71	999.00
	Minivan	5574	5574	3796.30	450.94	542.00
	Pickup	12607	12607	4120.93	893.36	278.00
	SUV	2218	2218	3898.23	729.7615389	1426.00
Les DM(40%) ont été imputées en utilisant la méthode KNN	Car	66213	66213	2942.35	602.49	192.00
	Heavy vehicle	542	542	4605.16	1540.83	1045.00
	Minivan	5565	5565	3800.10	445.98	542.00
	Pickup	12598	12598	4123.76	889.77	278.00
	SUV	2214	2214	3895.76	724.6526087	1426.00
Les DM(40%) ont été imputées en utilisant l'approche MissForest	Car	66185	66185	2942.73	602.46	192.00
	Heavy vehicle	543	543	4596.70	1522.91	999.00
	Minivan	5574	5574	3796.33	449.30	542.00
	Pickup	12612	12612	4120.26	889.64	278.00
	SUV	2218	2218	3898.08	723.66	1426.00
Les DM(40%) de ce jeu de données ont été imputés à l'aide de proc MI FCS.	Car	66184	66184	2942.91	603.50	192.00
	Heavy vehicle	549	549	4615.61	1551.94	192.00
	Minivan	5574	5574	3796.95	448.70	1285.00
	Pickup	12607	12607	4120.86	890.65	278.00
	SUV	2218	2218	3903.40	729.08	1426.00

Les statistiques descriptives de chaque type de véhicule demeure sensiblement la même quelques soit la méthode utilisée.

Tableau 12 : Distribution de la variable port de ceinture de sécurité.

	Percent	24 or less	25-44	45-64	65 and +	Total
Jeux de données sans aucune DM	**No Safety belt**	0.28	0.50	0.36	0.07	1.21
	Safety belt	20.33	39.53	28.82	10.11	98.79
DM (40%) imputées à l'aide de MissForest	**No Safety bel**	0.26	0.49	0.34	0.07	1.16
	Safety belt	19.45	39.38	30.37	9.64	98.84
DM (40%) imputées à l'aide de KNN	**No Safety bel**	0.27	0.49	0.34	0.07	1.16
	Safety belt	19.56	40.90	28.75	9.62	98.84
DM (40%) imputées à l'aide de proc Mi	**No Safety bel**	0.27	0.50	0.26	0.08	1.22
	Safety belt	20.38	39.43	28.79	10.15	98.77

Si on se focalise sur la variable « Port de la ceinture de sécurité », on pourra constater que les proportions ont été maintenues sauf quelques faibles variations. Cette variable présente un taux de Données Manquantes qui avoisine les 78% dans la base originale. Quant à la valeur « Safety Belt », elle est en l'occurrence majoritaire dans les différents jeux de données imputés.

Tableau 13: Répartition des types de véhicule selon les groupes d'âge.

(Nombre d'observations=87132)	**Percent**	**24 or less**	**25-44**	**45-64**	**65 and +**	**Total**
Jeux de données sans aucune DM	*Car*	*15.62*	*24.46*	*17.63*	*7.53*	*65.24*
	Heavy vehicle	*0.03*	*0.13*	*0.12*	*0.01*	*0.29*
	Minivan	*1.39*	*6.23*	*4.66*	*1.34*	*13.61*
	Pickup	*1.86*	*4.12*	*3.32*	*0.69*	*9.99*
	SUV	*1.70*	*5.09*	*3.45*	*0.61*	*10.87*
DM (40%) imputées à l'aide de MissForest	*Car*	*15.05*	*24.44*	*18.45*	*7.31*	*65.24*
	Heavy vehicle	*0.03*	*0.14*	*0.12*	*0.01*	*0.29*
	Minivan	*1.30*	*6.15*	*4.96*	*1.20*	*13.61*
	Pickup	*1.77*	*4.04*	*3.55*	*0.63*	*9.99*
	SUV	*1.57*	*5.11*	*3.62*	*0.57*	*10.87*
DM (40%) imputées à l'aide de KNN	*Car*	*15.16*	*25.49*	*17.52*	*7.29*	*65.46*
	Heavy vehicle	*0.03*	*0.13*	*0.10*	*0.01*	*0.27*
	Minivan	*1.33*	*6.33*	*4.67*	*1.18*	*13.51*
	Pickup	*1.76*	*4.23*	*3.38*	*0.63*	*9.99*
	SUV	*1.55*	*5.22*	*3.43*	*0.57*	*10.77*
DM (40%) imputées à l'aide de proc Mi	*Car*	*15.60*	*24.54*	*17.71*	*7.51*	*65.37*
	Heavy vehicle	*0.03*	*0.13*	*0.10*	*0.010*	*0.28*
	Minivan	*1.42*	*6.09*	*4.64*	*1.37*	*13.53*
	Pickup	*1.86*	*4.15*	*3.28*	*0.70*	*10.01*
	SUV	*1.73*	*5.02*	*3.41*	*0.63*	*10.80*

Le Tableau 12 résume la répartition du type de véhicule selon le groupe d'âge. Nous avons utilisé cette relation entre le groupe d'âge et le type de véhicule pour vérifier si les proportions resteront inchangées après le processus d'imputation. Le but est de se servir de cette relation entre les deux variables pour vérifier si

l'imputation va attribuer des valeurs non plausibles au DM. Comme on peut le constater il y a très peu de jeunes conducteur de camion et très peu de conducteurs âgés dans les jeux de données présentés.

En observant le Tableau 12, on peut noter que ces proportions sont relativement identiques dans les différents jeux de données.

Tableau 14: Répartition de l'expérience de conduite selon les groupes d'âge.

	Percent	**24 or less**	**25-44**	**45-64**	**+65**	**Total**
Jeux de données sans aucune DM	**Conducteur peu expérimenté**	20.12	32.93	23.51	8.17	84.73
	Conducteur expérimenté	0.48	1.75	0.07	0.01	2.31
	Conducteur très expérimenté	0.00	5.35	5.60	2.00	12.96
DM (40%) imputées à l'aide de MissForest	**Conducteur peu expérimenté**	9.41	1.06	0.10	0.01	10.57
	Conducteur expérimenté	9.75	8.44	0.47	0.08	18.74
	Conducteur très expérimenté	0.56	30.37	30.14	9.62	70.69
DM (40%) imputées à l'aide de KNN	**Conducteur peu expérimenté**	5.13	2.37	0.74	0.22	8.47
	Conducteur expérimenté	14.39	37.32	26.79	9.02	87.52
	Conducteur très expérimenté	0.31	1.70	1.56	0.44	4.00
DM (40%) imputées à l'aide de proc Mi	**Conducteur peu expérimenté**	8.21	15.84	9.11	3.04	8.21
	Conducteur expérimenté	11.55	12.12	8.12	2.80	11.55
	Conducteur très expérimenté	0.29	12.47	12.43	4.02	0.29

Parmi les variables auxiliaires qui ont été intégrées à la base afin de renforcer l'hypothèse MAR et de servir dans les modèles d'imputation, nous allons nous intéresser à la variable « P_YLIC » qui donne l'expérience de conduite en années. Nous avons utilisé cette variable pour générer trois classe de niveau d'expérience (Peu expérimenté : P_Ylic<=5 ans, Expérimenté : 5 ans < P_Ylic <=10 ans, Très expérimenté : P_Ylic > 10 ans) et ainsi utiliser cette information pour vérifier si les valeurs attribuées à cette même variable par le processus d'imputation sont plausibles ou non. Nous nous attendons par exemple à ce que les jeunes

conducteurs ne soient pas classé comme conducteurs très expérimentés (expérience > 10ans). A partir du Tableau 13, on peut noter dans le jeu de données originale que la proportion des jeunes conducteurs (24 ans ou moins) qualifiés de très expérimentés est nulle mais que dans les jeux de données imputés cette proportion est faible mais reste non nulle. La variable P_Ylic a probablement reçu quelques valeurs non plausibles pour remplacer les données manquantes.

Tableau 15: Port de la ceinture de sécurité selon le véhicule

		Bus	Car	Heavy vehicle	Minivan	Other	Pickup	SUV	Total
Aucune DM	Pas de ceinture de sécurité	0.06	0.80	0.01	0.10	0.19	0.12	0.00	1.28
	Ceinture de sécurité portée	0.00	66.08	0.11	9.08	17.49	5.76	0.20	98.72
DM imputée par FCS	Pas de ceinture de sécurité	0.06	0.74	0.01	0.10	0.20	0.11	0.00	1.22
	Ceinture de sécurité portée	**0.04**	64.86	0.11	9.25	18.41	5.92	0.20	98.78
DM imputée par KNN	Pas de sécurité	0.06	0.72	0.01	0.09	0.18	0.11	0.00	1.16
	Ceint. Sécu. portée	0.00	64.92	0.11	9.29	18.43	5.90	0.20	98.84
DM imputée par MissForest	Pas de ceinture de sécurité	0.06	0.70	0.01	0.09	0.17	0.11	0.00	1.13
	Ceinture de sécurité portée	0.00	64.80	0.12	9.59	18.08	6.06	0.22	98.87

Il est connu que les bus ne contiennent pas de ceinture de sécurité, le Tableau 14 montre qu'à part l'imputation multiple qui en a attribué un faible taux les autres

imputations simple (KNN, MissForest) ont attribué les bonnes valeurs et aucune valeurs aberrante n'a été généré pour remplacer les DM des observations concernées.

4.2.1) Comparaison des valeurs prédites et des valeurs initiales

Nous allons comparer les valeurs de remplacement, proposées par les différentes approches utilisées, et les valeurs initiales que contenait la base avant d'introduire les DM. Nous allons sélectionner les observations qui ont fait l'objet d'introduction de DM selon les différents taux. Nous allons par la suite identifier ces observations dans la base originale qui ne contient que les données complètes. En renommant les variables et fusionnant les deux échantillons nous allons avoir un jeu de données qui va servir à des fins de comparaison. Ce jeux de données ne représente qu'une partie du jeu de données total et va varier selon le taux de DM. Pour les variables nominales nous allons produire des matrices de confusions. Des statistiques descriptives seront générées pour les variables continues afin de réaliser les comparaisons. Nous allons présenter le jeu de données qui contient 40% de DM. Les résultats des comparaisons des autres jeux de données seront placés en annexe.

Le Tableau 15 montre qu'en général les trois processus d'imputation ont générés des valeurs plausibles à un taux qui avoisine les 99%. Ce résultat confirme ce qui a été noté à l'aide des représentations graphiques pour la variable « type de véhicule ».

Tableau 16:Matrice de confusion de la variable type de véhicule.

		Valeurs prédites : *les taux des bons classements sont en diagonale.*					
		Car	**Heavy vehicle**	**Minivan**	**Pickup**	**SUV**	**Total**
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide de l'Imputation multiple.	**Car**	**19575** **99.83**	24 0.12	8 0.04	0 0.00	2 0.01	19609
	Heavy vehicle	4 1.14	**336** **95.45**	12 3.41	0 0.00	0 0.00	352
	Minivan	0 0.00	8 0.27	**2997** **99.37**	11 0.36	0 0.00	3016
	Pickup	0 0.00	0 0.00	8 0.15	**5460** **99.71**	8 0.15	5476
	SUV	0 0.00	0 0.00	0 0.00	1 0.08	**1247** **99.92**	1248
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide du module KNN.	**Car**	**19606** **99.98**	0 0.00	0 0.00	3 0.02	0 0.00	19609
	Heavy vehicle	2 0.57	**345** **98.01**	1 0.28	4 1.14	0 0.00	352
	Minivan	4 0.13	0 0.00	**3006** **99.67**	6 0.20	0 0.00	3016
	Pickup	24 0.44	0 0.00	0 0.00	**5451** **99.54**	1 0.02	5476
	SUV	2 0.16	0 0.00	0 0.00	3 0.24	**1243** **99.60**	1248
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide du module MissForest.	**Car**	**19609** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	19609
	Heavy vehicle	0 0.00	**349** **99.15**	1 0.28	2 0.57	0 0.00	352
	Minivan	1 0.03	0 0.00	**3015** **99.97**	0 0.00	0 0.00	3016
	Pickup	0 0.00	0 0.00	0 0.00	**5476** **100.00**	0 0.00	5476
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**1248** **100.00**	1248

Tableau 17: Matrice de confusion de la variable ceinture de sécurité

		Valeurs prédites (les taux des bons classements sont en diagonale)		
		No Safety belt	**Safety belt**	**Total**
Échantillon qui contenait 40% de DM imputées à l'aide de l'imputation multiple.	**No Safety belt**	**31** **47**	36 53	67
	Safety belt	487 1.64	**29147** **98.36**	29634
Échantillon qui contenait 40% de DM imputées à l'aide de KNN	**No Safety belt**	**25** **37.31**	42 62.69	67
	Safety belt	1 0.00	**29633** **100.00**	**29634**
Échantillon qui contenait 40% de DM imputées à l'aide de MissForest	**No Safety belt**	**62** **92.54**	5 7.46	67
	Safety belt	0 0.00	**29634** **100.00**	29634

Intéressons-nous maintenant au Tableau 16 qui représente la matrice de confusion pour la variable « ceinture de sécurité » nous constatons que les différents procédures d'imputations ont générés des valeurs plausibles à un taux qui avoisine parfois le 100%. On observe que, dans ce cas-ci, l'imputation multiple et KNN ont un faible taux de bon classement quand la valeur est très peu représentée dans le jeu de données original. Il s'agit de la catégorie « Ceinture non porté ». Pour ce qui est de KNN la valeur proposée sera basé sur les K proches voisins. Le fait de moins bien prédire les catégories peu représentées peut trouver son origine dans la nature du module KNN.

Le Tableau 17 confirme le constat fait à l'aide des représentations graphiques. On peut y observer qu'en général les valeurs proposées pour remplacer les DM sont identiques aux valeurs initiales et ce à un taux qui varie de 75% à 98%. Les procédures d'imputations semblent mieux prédire les catégories les plus représentées.

Tableau 18: Matrice de confusion de la variable vitesse autorisée

		Valeurs prédites : *les taux des bons classements sont en diagonale.*					
		100 kph	**50-60 kph**	**70-80 kph**	**90 kph**	**- 50**	**Total**
40% de DM imputées par IM	**100 kph**	**1561** **91.45**	112 6.35	11 0.69	4 0.28	20 1.23	1708
	50-60 kph	524 2.45	**19884** **93.10**	387 1.81	128 0.6	435 2.04	21358
	70-80 kph	84 1.98	362 4.48	**3732** **87.47**	23 0.51	66 1.56	4267
	90 kph	22 1.43	96 6.25	22 1.45	**1366** **89.35**	**23** **1.52**	1529
	- 50 kph	28 3.38	114 13.68	26 3.17	12 1.43	**659** **78.62**	839
40% de DM imputées par KNN	**100 kph**	**1645** **96.31**	55 3.22	6 0.35	2 0.12	0 0.00	1708
	50-60 kph	14 0.07	**21272** **99.60**	36 0.17	21 0.10	15 0.07	21358
	70-80 kph	5 0.12	234 5.48	**4013** **94.05**	12 0.28	3 0.07	4267
	90 kph	1 0.07	58 3.79	8 0.52	**1461** **95.55**	1 0.07	1529
	- 50 kph	0 0.00	43 5.13	2 0.24	1 0.12	**793** **94.52**	839
40% de DM imputées par MissForest	**100 kph**	**1539** **90.11**	149 8.72	20 1.17	0 0.00	0 0.00	1708
	50-60 kph	6 0.03	**20972** **98.19**	368 1.72	7 0.03	5 0.02	21358
	70-80 kph	0 0.00	541 12.68	**3723** **87.25**	3 0.07	0 0.00	4267
	90 kph	0 0.00	159 10.40	12 0.78	**1357** **88.75**	1 0.07	1529
	- 50 kph	0 0.00	184 21.93	19 2.26	0 0.00	**636** **75.80**	839

Tableau 19: Statistiques descriptives de la variable poids du véhicule pré et post-imputation.

	Variable	N	Mean	Std Dev	Minimum	Maximum
40% de DM imputées par IM.	Poids original	29701	3402	874.78	353	9947
	Poids imputé	29701	3403	877.40	281	9947
40% de DM imputées par KNN.	Poids original	29701	3402.15	871.52	353.00	9947.00
	Poids imputé	29701	3402.21	874.78	353.00	9947.00
40% de DM imputées par MissForest.	Poids original	29701	3401.68	871.10	353.00	9947.00
	Poids imputé	29701	3402.21	874.78	353.00	9947.00

A la lumière du Tableau 19 on peut noter que la moyenne et la variance du poids du véhicule est presque identique, dans l'échantillon composé des observations ayant fait l'objet d'introduction de DM, pour les valeurs originales et celle proposées pour remplacer les DM.

Tableau 20: Statistiques descriptives de la variable âge pré et post-imputation

	Variable	N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 40% du jeu de données initial.	Age original Age imputé	29701 29701	44.20 44.26	16.99 17.07	16.00 16.00	94 94.2
Échantillon représentant 40% du jeu de données initial. (KNN)	Age original Age imputé	29701 29701	44.20 44.03	16.99 16.14	16.00 16.00	94.00 94.00
Échantillon représentant 40% du jeu de données initial (MissForest)	Age original Age imputé	29701 29701	44.20 44.30	16.99 16.57	16.00 16.00	94.00 94.00

Pour les variables âges on peut voir que les moyennes et variances des valeurs initiales et celle des valeurs générées pour remplacer les DM sont sensiblement identiques.

4.3) Application du modèle Fredette et al. (2008) aux différents jeux de données

Le Tableau 14 résume les cotes (odd ratios) obtenues en appliquant le modèle Fredette et al. (2008) aux jeux de données cas complets et les jeux de données qui ont été transformés pour contenir des taux de DM de 20%,40% et 60% selon la typologie MAR et qui n'ont pas encore été imputés.

La première colonne contient les cotes (Odd ratios) et les intervalle de confiance obtenus en appliquant un modèle similaire au modèle Fredette et al. (2008) au jeu de données cas complets. Si on s'intéresse à la variable rapport des masses des deux véhicules on note que les conducteurs de véhicules dont le rapport de masse avec le deuxième véhicule est moins que .80 sont 69% plus exposé au risque de blessure grave ou fatale que les conducteurs dont ce même ratio est entre 0.80 et 1.20 (cette catégorie étant la référence). Les conducteurs dont les véhicules ont un rapport de masse se situant entre 1.2 et 2 sont moins exposés, d'environ 30%, que ceux dont le ratio des masses est entre 0.80 et 1.20. Pour les conducteurs dont les véhicules ont un rapport de masse avec l'autre véhicule qui est au moins le double, le risque est réduit de 54%. Ces résultats confirment l'impact de la masse du véhicule sur le risque de subir lors des accidents une blessure grave ou fatale.

Si on s'intéresse maintenant aux résultats de la colonne 2 qui concerne le jeu de données avec 40% de DM on constate que la deuxième conclusion ne s'applique plus car elle n'est plus significative statistiquement. L'intervalle de confiance contient la valeur 1. Il en est de même pour le cas où le rapport de masse entre les deux véhicules est au moins le double.

En analysant les résultats obtenus avec des jeux de données qui contiennent différents taux de DM, on peut constater que les DM peuvent induire de fausses conclusions analytiques. Les catégories, qui étaient significatives dans le jeu de données cas complets, ne le sont plus dans les jeux de données qui contiennent des DM. C'est la preuve de la nuisance qui peut être engendrée par les DM.

Nous avons aussi calculé un ratio (valeur entre deux crochets) entre l'étendu des intervalles de confiance des différent jeux de données [(Étendu intervalle imputé) / (Étendu intervalle cas complet)]. Notons que les jeux de données qui contiennent des taux de DM ont des intervalles de confiance plus étendu que le jeu de données des cas complets. Les intervalles avec une plus grande étendue sont moins précis. L'étendue des intervalles de confiance des jeux de données qui contiennent des DM est parfois le double de l'étendue du jeu de données cas complets. Il est établit que plus l'échantillon est grand plus l'intervalle de confiance est étroit et plus les estimations sont précises. La taille de l'échantillon est une des différentes causes derrières l'élargissement de l'intervalle de confiance. La précision croît donc avec l'augmentation de la taille de l'échantillon. Étant donné que les jeux de données qui contiennent des DM sont réduits de 20%,40% ou 60% l'étendu de l'intervalle de confiance est plus importante que pour le jeu de données cas complets.

Tableau 21: Modèle Fredette et al(2008). Appliqué aux jeux de données cas complet et ceux qui contiennent des DM et qui n'ont pas encore été imputés

	Aucune DM (n=43132)	20 % de DM non imputé (n=35842)	40 % de DM non imputé (n=28266)	60 % de DM non imputé (n=20678)
Véhicule 1 : Pickup	**0.65 (0.51 - 0.83)**	**0.66 (0.50 - 0.86) [1.1]**	**0.52 (0.38 - 0.71) [1.0]**	**0.68 (0.48 - 0.94) [1.4]**
Véhicule 1 : Minivan	0.74 (0.52 - 1.05)	0.57 (0.36 - 0.92) [1.0]	0.73 (0.45 - 1.17) [1.3]	0.59 (0.31 - 1.11) [1.5]
Véhicule 1 : SUV	0.75 (0.40 - 1.41)	0.67 (0.31 - 1.48) [1.2]	0.64 (0.25 - 1.62) [1.4]	0.54 (0.16 - 1.80) [1.6]
Véhicule 1 : Véhicule lourd	0.50 (0.20 - 1.26)	0.35 (0.09 - 1.31) [1.2]	0.48 (0.11 - 2.04) [1.8]	0.45 (0.12 - 1.76) [1.5]
Véhicule 2 : Pickup	1.32 (1.07 - 1.64)	1.38 (1.09 - 1.75) [1.2]	1.66 (1.29 - 2.13) [1.5]	1.71 (1.28 - 2.29) [1.8]
Véhicule 2 : Minivan	1.06 (0.78 - 1.45)	1.25 (0.88 - 1.76) [1.3]	1.30 (0.89 - 1.90) [1.5]	1.13 (0.70 - 1.82) [1.7]
Véhicule 2 : SUV	0.79 (0.45 - 1.42)	0.92 (0.50 - 1.69) [1.2]	0.86 (0.41 - 1.78) [1.4]	1.10 (0.50 - 2.43) [2.0]
Véhicule 2 : Véhicule lourd	**2.02 (1.00 - 4.12)**	**2.23 (1.01 - 4.90) [1.2]**	**3.55 (1.59 - 7.95) [2.0]**	**2.39 (0.82 - 6.95) [2.0]**
Poids1/Poids2 : Less Than 0.80	**1.69 (1.40 - 2.03)**	**1.43 (1.18 - 1.74) [0.9]**	**1.25 (1.02 - 1.54) [0.8]**	**1.14 (0.89 - 1.45) [0.9]**
Poids1/Poids2 : (1.20;2.00]	**0.70 (0.57 - 0.86)**	**0.73 (0.58 - 0.92) [1.2]**	**0.89 (0.68 - 1.16) [1.7]**	**0.70 (0.50 - 0.99) [1.7]**
Poids1/Poids2 : More Than 2.0	**0.46 (0.26 - 0.84)**	**0.36 (0.16 - 0.80) [1.1]**	**0.45 (0.18 - 1.16) [1.7]**	**0.43 (0.15 - 1.27) [1.9]**
Sex du conducteur	0.98 (0.84 - 1.15)	0.99 (0.83 - 1.18) [1.1]	1.02 (0.84 - 1.24) [1.3]	1.16 (0.92 - 1.47) [1.8]
Age : moins de 25 ans	0.79 (0.63 - 0.98)	0.77 (0.61 - 0.98) [1.1]	0.76 (0.59 - 0.98) [1.1]	0.66 (0.49 - 0.88) [1.1]
Age : (45;64]	1.22 (1.02 - 1.47)	1.24 (1.02 - 1.52) [1.1]	1.15 (0.92 - 1.43) [1.2]	1.07 (0.82 - 1.39) [1.3]
Age : 65 et +	1.52 (1.21 - 1.91)	1.37 (1.05 - 1.79) [1.0]	1.52 (1.14 - 2.02) [1.3]	1.33 (0.93 - 1.89) [1.4]
Ceinture de sécurité	**8.62 (6.47 - 11.5)**	**9.01 (6.71 - 12.10) [1.1]**	**8.05 (5.94 - 10.92) [1.0]**	**9.70 (7.03 - 13.39) [1.3]**
Vitesse : moins de 50 Kph	**0.34 (0.16 - 0.72)**	**0.38 (0.18 - 0.81) [1.1]**	**0.38 (0.17 - 0.85) [1.2]**	**0.45 (0.20 - 1.02) [1.5]**
Vitesse : (70-80 kph]	**3.89 (3.26 - 4.64)**	**3.74 (3.08 - 4.54) [1.1]**	**3.67 (2.96 - 4.54) [1.1]**	**3.47 (2.70 - 4.45) [1.3]**
Vitesse : 90 kph	**5.24 (4.13 - 6.65)**	**4.72 (3.60 - 6.18) [1.0]**	**5.03 (3.73 - 6.76) [1.2]**	**4.14 (2.86 - 6.00) [1.2]**
Vitesse : 100 kph	**4.65 (3.27 - 6.62)**	**4.20 (2.78 - 6.35) [1.1]**	**4.80 (3.03 - 7.60) [1.4]**	**2.60 (1.36 - 4.98) [1.1]**
Rear-end	0.15 (0.11 - 0.21)	0.16 (0.12 - 0.22) [1.2]	0.17 (0.12 - 0.24) [1.3]	0.18 (0.12 - 0.28) [1.7]
Side-swipe (samedirection)	0.80 (0.55 - 1.17)	0.83 (0.55 - 1.26) [1.1]	0.67 (0.42 - 1.09) [1.1]	0.79 (0.46 - 1.37) [1.5]
Left-turn (samedirection)	0.74 (0.38 - 1.45)	0.88 (0.45 - 1.73) [1.2]	0.94 (0.46 - 1.93) [1.4]	1.16 (0.51 - 2.60) [1.9]
Right-turn (samedirectio)	0.51 (0.25 - 1.05)	0.57 (0.26 - 1.23) [1.2]	0.47 (0.19 - 1.17) [1.2]	0.52 (0.19 - 1.45) [1.6]
Head-On	4.79 (3.93 - 5.84)	4.81 (3.86 - 5.99) [1.1]	4.61 (3.61 - 5.88) [1.2]	4.97 (3.73 - 6.62) [1.5]
Left-turn (diff.directio)	0.82 (0.62 - 1.07)	0.81 (0.60 - 1.09) [1.1]	0.84 (0.60 - 1.16) [1.2]	0.77 (0.52 - 1.14) [1.4]
Right-turn (diff.directi)	0.71 (0.31 - 1.62)	0.71 (0.29 - 1.75) [1.1]	1.12 (0.49 - 2.60) [1.6]	0.96 (0.35 - 2.68) [1.8]
Other	**0.60 (0.46 - 0.78)**	**0.58 (0.43 - 0.77) [1.1]**	**0.71 (0.52 - 0.97) [1.5]**	**0.69 (0.46 - 1.01) [1.8]**

Tableau 22: Modèle Fredette et al(2008) appliqué aux jeux de données cas complets et ceux qui ont été imputés en utilisant KNN.

	Jeux sans DM	**20% de DM avant imputation** (n=43566)	**40% de DM avant imputation** (n=43566)	**60% de DM avant imputation** (n=43566)
Véhicule 1 : Pickup	**0.65 (0.51 - 0.83)**	**0.67 (0.52 - 0.85) [1.0]**	**0.67 (0.53 - 0.86) [1.0]**	**0.68 (0.53 - 0.87) [1.0]**
Véhicule 1 : Minivan	0.74 (0.52 - 1.05)	0.70 (0.49 - 1.00) [0.9]	0.64 (0.45 - 0.92) [0.9]	0.67 (0.47 - 0.95) [0.9]
Véhicule 1 : SUV	0.75 (0.40 - 1.41)	0.76 (0.41 - 1.42) [1.0]	0.73 (0.39 - 1.37) [1.0]	0.74 (0.40 - 1.38) [1.0]
Véhicule 1 : Véhicule lourd	0.50 (0.20 - 1.26)	0.54 (0.22 - 1.34) [1.1]	0.57 (0.24 - 1.38) [1.1]	0.61 (0.25 - 1.50) [1.2]
Véhicule 2 : Pickup	**1.32 (1.07 - 1.64)**	**1.31 (1.06 - 1.62) [1.0]**	**1.35 (1.09 - 1.67) [1.0]**	**1.35 (1.09 - 1.66) [1.0]**
Véhicule 2 : Minivan	1.06 (0.78 - 1.45)	1.09 (0.80 - 1.48) [1.0]	1.07 (0.79 - 1.46) [1.0]	1.03 (0.76 - 1.41) [1.0]
Véhicule 2 : SUV	0.79 (0.45 - 1.42)	0.74 (0.42 - 1.31) [0.9]	0.75 (0.42 - 1.33) [0.9]	0.76 (0.43 - 1.35) [0.9]
Véhicule 2 : Véhicule lourd	**2.02 (1.00 - 4.12)**	**2.09 (1.04 - 4.20) [1.0]**	**2.03 (1.02 - 4.04) [1.0]**	**1.96 (0.96 - 3.99) [1.0]**
Poids1/Poids2 : Less Than 0.80	**1.69 (1.40 - 2.03)**	**1.66 (1.38 - 1.99) [1.0]**	**1.67 (1.39 - 2.00) [1.0]**	**1.58 (1.32 - 1.90) [0.9]**
Poids1/Poids2 : (1.20;2.00]	**0.70 (0.57 - 0.86)**	**0.69 (0.57 - 0.85) [1.0]**	**0.73 (0.59 - 0.89) [1.0]**	**0.68 (0.56 - 0.84) [1.0]**
Poids1/Poids2 : More Than 2.0	**0.46 (0.26 - 0.84)**	**0.48 (0.27 - 0.87) [1.0]**	**0.54 (0.30 - 0.95) [1.1]**	**0.49 (0.27 - 0.88) [1.1]**
Sex du conducteur	0.98 (0.84 - 1.15)	1.03 (0.88 - 1.20) [1.0]	1.11 (0.95 - 1.30) [1.1]	1.05 (0.90 - 1.22) [1.0]
Age : moins de 25 ans	**0.79 (0.63 - 0.98)**	**0.84 (0.68 - 1.04) [1.1]**	**0.78 (0.63 - 0.98) [1.0]**	**0.82 (0.66 - 1.02) [1.0]**
Age : (45;64]	**1.22 (1.02 - 1.47)**	**1.21 (1.01 - 1.45) [1.0]**	**1.21 (1.02 - 1.45) [1.0]**	**1.18 (0.99 - 1.41) [0.9]**
Age : 65 et +	**1.52 (1.21 - 1.91)**	**1.49 (1.19 - 1.87) [1.0]**	**1.47 (1.17 - 1.84) [1.0]**	**1.45 (1.15 - 1.82) [1.0]**
Ceinture de sécurité	**8.62 (6.47 - 11.49)**	**8.87 (6.66 - 11.81) [1.0]**	**8.60 (6.46 - 11.46) [1.0]**	**9.95 (7.43 - 13.35) [1.2]**
Vitesse : moins de 50 Kph	**0.34 (0.16 - 0.72)**	**0.36 (0.17 - 0.76) [1.1]**	**0.37 (0.17 - 0.78) [1.1]**	**0.39 (0.18 - 0.82) [1.1]**
Vitesse : (70-80 kph]	**3.89 (3.26 - 4.64)**	**3.93 (3.31 - 4.68) [1.0]**	**3.85 (3.24 - 4.57) [1.0]**	**4.16 (3.51 - 4.93) [1.0]**
Vitesse : 90 kph	**5.24 (4.13 - 6.65)**	**5.61 (4.44 - 7.09) [1.1]**	**5.27 (4.17 - 6.68) [1.0]**	**6.19 (4.92 - 7.78) [1.1]**
Vitesse : 100 kph	**4.65 (3.27 - 6.62)**	**3.79 (2.69 - 5.33) [0.8]**	**3.32 (2.37 - 4.65) [0.7]**	**3.20 (2.26 - 4.53) [0.7]**
Rear-end	**0.15 (0.11 - 0.21)**	**0.17 (0.13 - 0.23) [1.1]**	**0.20 (0.15 - 0.26) [1.2]**	**0.19 (0.14 - 0.25) [1.2]**
Side-swipe (samedirection)	0.80 (0.55 - 1.17)	0.85 (0.58 - 1.24) [1.1]	0.91 (0.62 - 1.33) [1.1]	0.92 (0.63 - 1.36) [1.2]
Left-turn (samedirection)	0.74 (0.38 - 1.45)	0.77 (0.39 - 1.51) [1.0]	0.77 (0.38 - 1.55) [1.1]	0.88 (0.45 - 1.71) [1.2]
Right-turn (samedirectio)	0.51 (0.25 - 1.05)	0.56 (0.27 - 1.15) [1.1]	0.60 (0.29 - 1.23) [1.2]	0.45 (0.20 - 1.03) [1.0]
Head-On	**4.79 (3.93 - 5.84)**	**3.75 (3.08 - 4.56) [0.8]**	**3.50 (2.88 - 4.25) [0.7]**	**2.79 (2.31 - 3.38) [0.6]**
Left-turn (diff.directio)	0.82 (0.62 - 1.07)	0.85 (0.66 - 1.11) [1.0]	0.87 (0.67 - 1.14) [1.0]	0.86 (0.66 - 1.13) [1.0]
Right-turn (diff.directi)	0.71 (0.31 - 1.62)	0.73 (0.32 - 1.68) [1.0]	0.76 (0.33 - 1.75) [1.1]	0.76 (0.33 - 1.74) [1.1]
Other	**0.60 (0.46 - 0.78)**	**0.60 (0.46 - 0.77) [1.0]**	**0.59 (0.45 - 0.77) [1.0]**	**0.61 (0.47 - 0.79) [1.0]**

Intéressons-nous maintenant à la comparaison du jeu de données cas complets et les jeux de données qui contenaient des taux de DM et qui ont été imputés à l'aide de l'approche KNN. Ces résultats sont résumés dans le Tableau 15. On peut constater que

le rapport [(Étendu intervalle imputé) / (Étendu intervalle cas complet)] est proche de 1 dans la plus part des cas. On peut en déduire que les données imputées à l'aide de KNN génèrent des intervalles de confiance d'étendu égale et aussi précis que les données du jeu cas complet. Notons que les catégories qui étaient statistiquement significatives le sont aussi dans les résultats obtenus à partir des jeux de données imputés. En imputant les données toute les observations sont utilisées, par conséquent les intervalles de confiance auront presque la même étendu que les intervalles obtenus avec le jeu de données cas complet. Nous avons fait la moyenne de tous les rapports des étendus de chaque jeu imputé et constaté que cette moyenne avoisine le 1. Les étendus des intervalles du jeu cas complet et des jeux imputés sont presque similaires.

Tableau 23: Modèle Fredette et al(2008). Appliqué aux jeux de données cas complet et ceux imputés en utilisant missForest.

	Jeux sans DM	**20% de DM avant imputation** (n=43566)	**40% de DM avant imputation** (n=43566)	**60% de DM avant imputation** (n=43566)
Véhicule 1 : Pickup	**0.65 (0.51 - 0.83)**	**0.65 (0.51 - 0.83) [1.0]**	**0.63 (0.50 - 0.81) [1.0]**	**0.63 (0.50 - 0.81) [1.0]**
Véhicule 1 : Minivan	0.74 (0.52 - 1.05)	0.74 (0.52 - 1.05) [1.0]	0.74 (0.52 - 1.06) [1.0]	0.73 (0.51 - 1.04) [1.0]
Véhicule 1 : SUV	0.75 (0.40 - 1.41)	0.77 (0.41 - 1.45) [1.0]	0.76 (0.40 - 1.42) [1.0]	0.74 (0.39 - 1.39) [1.0]
Véhicule 1 : Véhicule lourd	0.50 (0.20 - 1.26)	0.54 (0.22 - 1.37) [1.1]	0.56 (0.22 - 1.40) [1.1]	0.49 (0.19 - 1.24) [1.0]
Véhicule 2 : Pickup	**1.32 (1.07 - 1.64)**	**1.33 (1.07 - 1.65) [1.0]**	**1.33 (1.07 - 1.64) [1.0]**	**1.35 (1.09 - 1.67) [1.0]**
Véhicule 2 : Minivan	1.06 (0.78 - 1.45)	1.06 (0.77 - 1.44) [1.0]	1.05 (0.77 - 1.43) [1.0]	1.09 (0.80 - 1.48) [1.0]
Véhicule 2 : SUV	0.79 (0.45 - 1.42)	0.80 (0.45 - 1.43) [1.0]	0.80 (0.45 - 1.42) [1.0]	0.81 (0.46 - 1.45) [1.0]
Véhicule 2 : Véhicule lourd	**2.02 (1.00 - 4.12)**	**2.01 (0.99 - 4.08) [1.0]**	**2.07 (1.02 - 4.21) [1.0]**	**2.02 (0.99 - 4.10) [1.0]**
Poids1/Poids2 : Less Than 0.80	**1.69 (1.40 - 2.03)**	**1.68 (1.39 - 2.02) [1.0]**	**1.68 (1.40 - 2.03) [1.0]**	**1.66 (1.38 - 2.00) [1.0]**
Poids1/Poids2 : (1.20;2.00]	**0.70 (0.57 - 0.86)**	**0.70 (0.57 - 0.85) [1.0]**	**0.72 (0.59 - 0.89) [1.0]**	**0.72 (0.58 - 0.88) [1.0]**
Poids1/Poids2 : More Than 2.0	**0.46 (0.26 - 0.84)**	**0.48 (0.26 - 0.86) [1.0]**	**0.44 (0.24 - 0.80) [1.0]**	**0.48 (0.26 - 0.86) [1.0]**
Sex du conducteur	0.98 (0.84 - 1.15)	1.01 (0.86 - 1.18) [1.0]	1.02 (0.87 - 1.19) [1.0]	1.00 (0.85 - 1.17) [1.0]
Age : moins de 25 ans	**0.79 (0.63 - 0.98)**	**0.76 (0.61 - 0.95) [1.0]**	**0.76 (0.61 - 0.95) [1.0]**	**0.75 (0.60 - 0.94) [1.0]**
Age : (45;64]	**1.22 (1.02 - 1.47)**	**1.17 (0.97 - 1.40) [0.9]**	**1.23 (1.03 - 1.47) [1.0]**	**1.15 (0.97 - 1.38) [0.9]**
Age : 65 et +	**1.52 (1.21 - 1.91)**	**1.47 (1.16 - 1.85) [1.0]**	**1.57 (1.25 - 1.98) [1.0]**	**1.46 (1.16 - 1.85) [1.0]**
Ceinture de sécurité	**8.62 (6.47 - 11.49)**	**8.27 (6.17 - 11.09) [1.0]**	**8.26 (6.14 - 11.12) [1.0]**	**8.60 (6.39 - 11.59) [1.0]**
Vitesse : moins de 50 Kph	**0.34 (0.16 - 0.72)**	**0.35 (0.16 - 0.74) [1.0]**	**0.30 (0.13 - 0.68) [1.0]**	**0.35 (0.17 - 0.75) [1.0]**
Vitesse : (70-80 kph]	**3.89 (3.26 - 4.64)**	**4.09 (3.43 - 4.88) [1.1]**	**4.00 (3.35 - 4.78) [1.0]**	**3.89 (3.26 - 4.65) [1.0]**
Vitesse : 90 kph	**5.24 (4.13 - 6.65)**	**5.44 (4.29 - 6.90) [1.0]**	**5.50 (4.34 - 6.97) [1.0]**	**5.40 (4.26 - 6.85) [1.0]**
Vitesse : 100 kph	**4.65 (3.27 - 6.62)**	**4.78 (3.36 - 6.81) [1.0]**	**4.48 (3.12 - 6.44) [1.0]**	**4.85 (3.41 - 6.91) [1.0]**
Rear-end	**0.15 (0.11 - 0.21)**	**0.15 (0.11 - 0.20) [1.0]**	**0.15 (0.11 - 0.20) [1.0]**	**0.15 (0.11 - 0.20) [1.0]**
Side-swipe (samedirection)	0.80 (0.55 - 1.17)	0.78 (0.54 - 1.14) [1.0]	0.80 (0.55 - 1.17) [1.0]	0.77 (0.53 - 1.13) [1.0]
Left-turn (samedirection)	0.74 (0.38 - 1.45)	0.72 (0.37 - 1.41) [1.0]	0.68 (0.34 - 1.36) [1.0]	0.72 (0.37 - 1.41) [1.0]
Right-turn (samedirectio)	0.51 (0.25 - 1.05)	0.50 (0.24 - 1.04) [1.0]	0.51 (0.25 - 1.06) [1.0]	0.50 (0.24 - 1.04) [1.0]
Head-On	**4.79 (3.93 - 5.84)**	**4.64 (3.81 - 5.66) [1.0]**	**4.53 (3.71 - 5.53) [0.9]**	**4.61 (3.78 - 5.63) [1.0]**
Left-turn (diff.directio)	0.82 (0.62 - 1.07)	0.81 (0.61 - 1.06) [1.0]	0.81 (0.62 - 1.06) [1.0]	0.81 (0.61 - 1.06) [1.0]
Right-turn (diff.directi)	0.71 (0.31 - 1.62)	0.71 (0.31 - 1.62) [1.0]	0.72 (0.32 - 1.66) [1.0]	0.72 (0.32 - 1.66) [1.0]
Other	**0.60 (0.46 - 0.78)**	**0.60 (0.46 - 0.78) [1.0]**	**0.61 (0.47 - 0.79) [1.0]**	**0.60 (0.47 - 0.78) [1.0]**

Le Tableau 17 qui résume les résultats obtenus en appliquant un modèle proche du modèle Fredette *et al.* (2008) aux jeux de données cas complet et les jeux imputés à l'aide du module R missForest. On peut noter que le rapport [(Étendu intervalle imputé) /

(Étendu intervalle cas complet)] est proche de 1 dans la plus part des cas. Les intervalles ont la même étendu et donnent le même niveau de précision.

On observe aussi que les catégories qui sont significatives dans le jeu de données cas complets le sont aussi dans les autres jeux imputés à l'aide de missForest.

Tableau 24: Modèle Fredette et al(2008) appliqué aux jeux de données cas complets et ceux imputés en utilisant l'imputation multiple.

	Jeux sans DM (n=43566)	**20% de DM imputée** (n=43566)	**40% de DM imputée** (n=43566)	**60% de DM imputée** (n=43566)
Véhicule 1 : Pickup	**0.65 (0.51 - 0.83)**	**1.55 (1.21 - 1.99) [2.5]**	**1.63 (1.22 - 2.17) [3.0]**	**1.51 (1.17 - 1.95) [2.4]**
Véhicule 1 : Minivan	0.74 (0.52 - 1.05)	1.47 (1.00 - 2.15) [2.1]	1.22 (0.80 - 1.86) [2.0]	1.46 (0.98 - 2.16) [2.2]
Véhicule 1 : SUV	0.75 (0.40 - 1.41)	1.73 (0.83 - 3.59) [2.7]	1.40 (0.75 - 2.59) [1.8]	1.57 (0.80 - 3.09) [2.3]
Véhicule 1 : Véhicule lourd	0.50 (0.20 - 1.26)	2.38 (0.87 - 6.50) [5.3]	1.72 (0.61 - 4.83) [4.0]	1.73 (0.69 - 4.38) [3.5]
Véhicule 2 : Pickup	**1.32 (1.07 - 1.64)**	**0.80 (0.64 - 1.00) [0.6]**	**0.72 (0.58 - 0.90) [0.6]**	**0.73 (0.59 - 0.92) [0.6]**
Véhicule 2 : Minivan	1.06 (0.78 - 1.45)	0.98 (0.71 - 1.35) [1.0]	1.03 (0.73 - 1.46) [1.1]	0.88 (0.63 - 1.22) [0.9]
Véhicule 2 : SUV	0.79 (0.45 - 1.42)	1.22 (0.68 - 2.18) [1.5]	1.46 (0.83 - 2.56) [1.8]	1.16 (0.64 - 2.07) [1.5]
Véhicule 2 : Véhicule lourd	**2.02 (1.00 - 4.12)**	**0.52 (0.25 - 1.09) [0.3]**	**0.55 (0.22 - 1.41) [0.4]**	**0.50 (0.24 - 1.04) [0.3]**
Poids1/Poids2 : Less Than 0.80	**1.69 (1.40 - 2.03)**	**0.61 (0.50 - 0.74) [0.4]**	**0.59 (0.49 - 0.72) [0.4]**	**0.63 (0.52 - 0.77) [0.4]**
Poids1/Poids2 : (1.20;2.00]	**0.70 (0.57 - 0.86)**	**1.45 (1.18 - 1.79) [2.1]**	**1.41 (1.13 - 1.75) [2.2]**	**1.43 (1.15 - 1.78) [2.2]**
Poids1/Poids2 : More Than 2.0	**0.46 (0.26 - 0.84)**	**2.67 (1.37 - 5.24) [6.6]**	**2.01 (1.11 - 3.66) [4.4]**	**2.37 (1.23 - 4.58) [5.8]**
Sex du conducteur	0.98 (0.84 - 1.15)	1.01 (0.86 - 1.19) [1.1]	0.99 (0.83 - 1.17) [1.1]	0.93 (0.78 - 1.11) [1.1]
Age : moins de 25 ans	**0.79 (0.63 - 0.98)**	**1.29 (1.03 - 1.61) [1.7]**	**1.32 (1.04 - 1.66) [1.8]**	**1.39 (1.10 - 1.76) [1.9]**
Age : (45;64]	**1.22 (1.02 - 1.47)**	**0.81 (0.67 - 0.98) [0.7]**	**0.84 (0.69 - 1.02) [0.7]**	**0.90 (0.74 - 1.09) [0.8]**
Age : 65 et +	**1.52 (1.21 - 1.91)**	**0.70 (0.55 - 0.90) [0.5]**	**0.66 (0.52 - 0.84) [0.5]**	**0.74 (0.58 - 0.95) [0.5]**
Ceinture de sécurité	**8.62 (6.47 - 11.49)**	**0.12 (0.09 - 0.16) [0.0]**	**0.15 (0.04 - 0.48) [0.1]**	**0.13 (0.09 - 0.18) [0.0]**
Vitesse : moins de 50 Kph	**0.34 (0.16 - 0.72)**	**2.22 (1.02 - 4.83) [6.8]**	**2.03 (0.86 - 4.82) [7.1]**	**1.76 (0.68 - 4.57) [6.9]**
Vitesse : (70-80 kph]	**3.89 (3.26 - 4.64)**	**0.27 (0.22 - 0.32) [0.1]**	**0.26 (0.22 - 0.32) [0.1]**	**0.27 (0.23 - 0.33) [0.1]**
Vitesse : 90 kph	**5.24 (4.13 - 6.65)**	**0.20 (0.16 - 0.26) [0.0]**	**0.22 (0.17 - 0.29) [0.0]**	**0.22 (0.17 - 0.28) [0.0]**
Vitesse : 100 kph	**4.65 (3.27 - 6.62)**	**0.22 (0.15 - 0.31) [0.0]**	**0.24 (0.13 - 0.45) [0.1]**	**0.29 (0.19 - 0.42) [0.1]**
Rear-end	**0.15 (0.11 - 0.21)**	**6.42 (4.74 - 8.70) [43.0]**	**6.05 (4.46 - 8.22) [40.9]**	**5.74 (4.22 - 7.80) [38.9]**
Side-swipe (samedirection)	0.80 (0.55 - 1.17)	1.25 (0.85 - 1.83) [1.6]	1.17 (0.78 - 1.76) [1.6]	1.11 (0.75 - 1.64) [1.4]
Left-turn (samedirection)	0.74 (0.38 - 1.45)	1.25 (0.64 - 2.45) [1.7]	1.25 (0.62 - 2.53) [1.8]	1.01 (0.52 - 1.96) [1.3]
Right-turn (samedirectio)	0.51 (0.25 - 1.05)	1.81 (0.88 - 3.75) [3.6]	1.71 (0.83 - 3.53) [3.4]	2.22 (0.97 - 5.09) [5.1]
Head-On	**4.79 (3.93 - 5.84)**	**0.21 (0.17 - 0.26) [0.0]**	**0.19 (0.16 - 0.24) [0.0]**	**0.20 (0.16 - 0.25) [0.0]**
Left-turn (diff.directio)	0.82 (0.62 - 1.07)	1.24 (0.94 - 1.64) [1.5]	1.18 (0.89 - 1.57) [1.5]	1.23 (0.92 - 1.64) [1.6]
Right-turn (diff.directi)	0.71 (0.31 - 1.62)	1.34 (0.58 - 3.05) [1.9]	1.29 (0.56 - 2.97) [1.8]	1.29 (0.56 - 2.96) [1.8]
Other	**0.60 (0.46 - 0.78)**	**1.70 (1.31 - 2.21) [2.9]**	**1.74 (1.33 - 2.29) [3.1]**	**1.70 (1.29 - 2.24) [3.0]**

Le Tableau 24 résume les résultats obtenus en appliquant un modèle similaire au modèle Fredette *et al.* (2008) au jeu de données qui ne contient pas de DM et aux jeux données qui contiennent des taux respectifs de 20%,40% et 60% de DM. Ces derniers ont été créés en introduisant des DM, selon la typologie MAR, au jeu de données qui n'en contenait pas. Les jeux de données contenant des DM ont été imputés en utilisant l'imputation multiple.

A la différence des autres méthodes d'imputation utilisées, KNN et MissForest, on observe que les intervalles de confiance obtenus à l'aide des jeux de données imputés sont plus étendus. Les catégories qui été significatives dans le résultat obtenus à l'aide du jeu de données complètes le sont aussi dans les résultats obtenus à partir des jeux de données imputés.

Les étendus plus larges obtenus à l'aide des jeux imputés à l'aide de l'imputation multiple peuvent être expliqués par l'augmentation de la variance. En effet l'imputation multiple propose plusieurs valeurs plausibles pour imputer une DM pour en tenant compte de l'incertitude qui concerne les DM. Ce qui peut contribuer à faire augmenter la variance. L'augmentation de la variance est une des causes de l'élargissement de l'étendu de l'intervalle de confiance.

Les résultats qui proviennent de l'imputation multiple et qui sont résumés dans le Tableau 24 ont été regroupés en suivant les règles de Rubin décrites dans le chapitre 2.

4.4) Récapitulatif du temps d'exécution des méthodes d'imputations utilisés

Tableau 25: Récapitulatif des méthodes utilisées.

Méthode	Stabilité	Niveau de difficulté	Temps d'exécution avec un échantillon de 43596 observations		
			20% de DM	40% de DM	60% de DM
Proc mi FCS	+ou- stable	élevé	15h58	20h07	44h58
MissForest	Très Stable	Moyen	18h50	18h30	18h15
KNN	Très Stable	Bas	6h43	7h05	7h30

Le temps d'exécution de l'approche KNN prenait environ 20h00 avant d'observer les variables qui ont un très grand score de prédiction (Variable Importance) et de les introduire comme paramètres dans KNN pour qu'elles servent au calcul des distances. Cette action nous a permis de réduire le temps d'exécution de 60% et de le ramener au temps raisonnable d'environ 8h30. Rappelons que nous nous sommes limités à travailler avec deux ans de données après avoir fait la majeure partie des travaux préliminaires sur toute la base à savoir 6 ans et 9.4 millions d'observations. La même démarche de réduction du temps d'exécution a été conduite avec la méthode MissForest. Nous avons utilisé un aspect, qui augmente l'intérêt pour cette méthode, à savoir la possibilité de paralléliser le processus d'imputation. De nos jours presque tous les processeurs des ordinateurs ont plus d'un cœur et à moins de paralléliser les opérations ou d'utiliser des logiciels qui le font déjà, une bonne partie des traitements va se dérouler sur un seul cœur. En utilisant le module « parallèle » nous avons réduit le temps d'exécution d'environ 50%. Le temps actuel de déroulement de l'imputation est d'environ 20h00 pour un échantillon d'environ 88000 observations qui représentent le nombre d'unités qui n'ont pas de DM.

Discussion et conclusion

Dans ce mémoire, nous avons comme objectif principal de comparer et d'appliquer les méthodes d'imputation sur la base nationale de données sur les collisions (NCDB). Nous avons pu sélectionner de nombreux module R ou SAS pour ne garder à la fin que ceux qui seraient adaptés à notre base de données. Nous en avons choisi trois : la méthode FCS (SAS), MissForest (R) et KNN(R).

Nous devons prévoir le moyen de vérifier les résultats des méthodes sélectionnées en procédant à des diagnostics post-imputations. Cette tâche représentait un défi majeur, car les valeurs qui manquent ne sont pas disponibles pour pouvoir les comparer par la suite aux valeurs plausibles qui sont générées par les outils sélectionnés. Nous avons choisi de bâtir nous même une base qui ne contient que les cas complets et d'y introduire des DM à des taux prédéterminés selon la typologie MAR pour évaluer, par la suite, la qualité des méthodes d'imputation.

Nous avons été en mesure d'observer que quand seules les observations avec des données complètes sont analysées et que ces unités ne constituent pas un échantillon représentatif de l'ensemble des données, les conclusions engendrées peuvent être biaisés et ne sont pas, parfois, valide pour en déduire de l'inférence statistique. L'utilisation de certaines techniques d'imputation requièrent de poser des hypothèses, sur la normalité de la distribution, par exemple, qu'il n'est pas facile de respecter surtout en présence de données multidimensionnelles mixtes qui contiennent à la fois des données binaires, des données continue et des données catégorielles. Parfois les méthodes d'imputation nécessitent la définition de modèles basés sur la normalité des données. Au travail que doit consacrer l'analyste, à l'étude d'intérêt, s'ajoute alors un effort colossal pour bâtir des modèles d'imputations. Nous avons utilisé des méthodes de sélection de variables pour pouvoir construire des modèles d'imputation.

C'est pour cela que nous avons orienté notre étude vers une comparaison à la fois des méthodes dites paramétriques qui nécessitent la définition de modèles d'imputations et des méthodes dites non paramétrique qui libèrent l'analyste des contraintes liées au non-respect des hypothèses. Certaines méthodes d'imputation nécessitent de redéfinir les variables ce qui les dénaturent et leur enlèvent leur sens. Par exemple les variables catégorielles sont traitées comme des variables numériques, imputées en suivant le modèle multivarié normale puis arrondies pour leur redonner leur nature discrète. Nous avons testé des méthodes développées récemment qui produisent un seul jeu de données, donc qui font de l'imputation dite simple et une méthode qui génère des imputations multiples.

La présente étude nous a permis de démonter l'impact des DM sur les études analytiques. En effet, en prenant comme exemple le modèle de Fredette *et al.* (2008) nous avons pu observer que la présence des DM peut induire de fausses conclusions analytiques. Nous avons noté que certains intervalles de confiance engendrées en présence des DM sont trop large ce qui cause des imprécisions et des biais.

Nous avons pu démontrer aussi que les méthodes non paramétriques développées récemment sont aussi performantes et donnent d'aussi excellents résultats que les méthodes d'imputations multiples. Nous avons pu établir en se servant du modèle Fredette *et al.* (2008) que les données imputées génèrent des paramètres précis et proches de la réalité. Nous avons aussi expérimenté les méthodes d'imputations sur une base de données multidimensionnelle et volumineuse de quelques millions d'observations et réalisé que cette tâche est fastidieuse mais réalisable.

Au cours de cette recherche nous avons rencontré des variables qui nous ont poussés à se questionner sur l'imputabilité des variables en d'autre termes quelle variables doivent être imputées et celles qui doivent plutôt être laissées à l'analyste pour les déduire à partir des données présentes. Nous pensons en particulier à la variable « sévérité de la blessure ». Ce genre de variable ne peut être confié à un modèle pour imputer une valeur qui statue sur le décès de la personne.

En guise d'avenue de recherche nous croyons qu'il est utile d'avoir des outils qui peuvent détecter et attribuer un score d'imputabilité à chaque observation pour savoir si elle peut être imputé ou non et ce en fonction du nombre de valeurs présente et de la nature de l'information. Ce score serait un préambule au processus d'imputation. Nous avons passé en revue de nombreuses recherches qui font des études sur les pourcentages des données manquantes par colonne mais nous n'en avons pas rencontré qui s'intéressent au pourcentage de données manquantes par ligne. Nous entendons par ça quel ratio de nombre de variable avec DM sur le nombre total des variables serait acceptable pour faire l'imputation.

Bibliographie

Allan, FE et J Wishart (1930). « A method of estimating the yield of a missing plot in field experimental work », *The Journal of Agricultural Science*, vol. 20, no 3, p. 399-406.

Allison, Paul D. (2001). *Missing data*, Thousand Oaks, Calif.; London, Sage.

Baraldi, A. N. et C. K. Enders (2010). « An introduction to modern missing data analyses », *Journal of School Psychology*, vol. 48, no 1, p. 5-37.

Beretta, Lorenzo et Alessandro Santaniello (2016). « Nearest neighbor imputation algorithms: A critical evaluation », *BMC medical informatics and decision making*, vol. 16, no 3, p. 74.

Breiman, L., J. Friedman, C.J. Stone et R.A. Olshen (1984). *Classification and regression trees*, Taylor & Francis.

Breiman, Leo (2001). « Random forests », *Machine learning*, vol. 45, no 1, p. 5-32.

Buck, Samuel F (1960). « A method of estimation of missing values in multivariate data suitable for use with an electronic computer », *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 302-306.

Buuren, Stef van (2012). *Flexible imputation of missing data*, Boca Raton, FL, CRC Press.

Casella, G. et E. I. George (1992). « Explaining the gibbs sampler », *American Statistician*, vol. 46, no 3, p. 167-174.

Collins, L. M., J. L. Schafer et C. M. Kam (2001). « A comparison of inclusive and restrictive strategies in modern missing data procedures », *Psychological Methods*, vol. 6, no 4, p. 330-351.

Crookston, NL et AO Finley (2008). « Main content area yaimpute: An r package for knn imputation », *Journal of Statistical Software*, vol. 23, no 10, p. 1-16.

Dear, Robert Ernest (1959). *A principal-component missing-data method for multiple regression models*, System Development Corporation.

Dempster, Arthur P, Nan M Laird et Donald B Rubin (1977). « Maximum likelihood from incomplete data via the em algorithm », *Journal of the royal statistical society. Series B (methodological)*, p. 1-38.

Donders, A. R. T., G. J. M. G. van der Heijden, T. Stijnen et K. G. M. Moons (2006). « Review: A gentle introduction to imputation of missing values », *Journal of Clinical Epidemiology*, vol. 59, no 10, p. 1087-1091.

Donzé, Laurent (2001). « L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données ».

Enders, Craig K. (2010). *Applied missing data analysis*, New York, Guilford Press.

Faisal, Shahla et Gerhard Tutz (2017). « Nearest neighbor imputation for categorical data by weighting of attributes », *arXiv preprint arXiv:1710.01011*.

Frane, J. W. (1976). « Some simple procedures for handling missing data in multivariate-analysis », *Psychometrika*, vol. 41, no 3, p. 409-415.

Fredette, Marc, Lema Sikoti Mambu, Aline Chouinard et François Bellavance (2008). « Safety impacts due to the incompatibility of suvs, minivans, and pickup trucks in two-vehicle collisions », *Accident Analysis & Prevention*, vol. 40, no 6, p. 1987-1995.

Graham, John W (2009). « Missing data analysis: Making it work in the real world », *Annual review of psychology*, vol. 60, p. 549-576.

Heitjan, Daniel F. et Roderick J. A. Little (1991). « Multiple imputation for the fatal accident reporting system », *Applied Statistics*, p. 13-29 %@ 0035-9254.

Jones, M. P. (1996). « Indicator and stratification methods for missing explanatory variables in multiple linear regression », *Journal of the American Statistical Association*, vol. 91, no 433, p. 222-230.
Josse, J. et F. Husson (2016). « Missmda: A package for handling missing values in multivariate data analysis », *Journal of Statistical Software*, vol. 70, no 1.
Kenward, Michael G et Geert Molenberghs (1998). « Likelihood based frequentist inference when data are missing at random », *Statistical Science*, vol. 13, no 3, p. 236-247.
Kiers, Henk AL (1997). « Weighted least squares fitting using ordinary least squares algorithms », *Psychometrika*, vol. 62, no 2, p. 251-266.
Knol, M. J., K. J. M. Janssen, A. R. T. Donders, A. C. G. Egberts, E. R. Heerdink, D. E. Grobbee*, et al.* (2010). « Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example », *Journal of Clinical Epidemiology*, vol. 63, no 7, p. 728-736.
Kramer, Clyde Young et Suzanne Glass (1960). « Analysis of variance of a latin square design with missing observations », *Applied Statistics*, p. 43-50.
Little, Roderick J. A. (1987). « Multiple imputation for nonresponse in surveys:Rubin, d.B. (1987). New york: John wiley & sons. 258 pages, $32.95 », *CEUS</cja:jid> Computers, Environment and Urban Systems*, vol. 14, no 1, p. 75-75.
Little, Roderick J. A. et Donald B Rubin (2016). *Statistical analysis with missing data*, Chicester, John Wiley and Sons Ltd.
Little, Roderick J. A. et Donald B. Rubin (1987). *Statistical analysis with missing data*, New York, Wiley, coll. Wiley series in probability and mathematical statistics applied probability and statistics, xiv, 278 p. p.
Little, Roderick J. A. et Donald B. Rubin (2002). *Statistical analysis with missing data*, 2nd^{e} éd., Hoboken, N.J., Wiley, coll. Wiley series in probability and statistics, xv, 381 p. p.
Liu, Huawen et Shichao Zhang (2012). « Noisy data elimination using mutual k-nearest neighbor for classification mining », *Journal of Systems and Software*, vol. 85, no 5, p. 1067-1074.
Matthai, Abraham (1951). « Estimation of parameters from incomplete data with application to design of sample surveys », *Sankhyā: The Indian Journal of Statistics*, p. 145-152.
McKnight, Patrick E. (2007). *Missing data : A gentle introduction*, New York, Guilford Press.
Meinfelder, Florian, Thorsten Schnapp et Maintainer Florian Meinfelder (2015). « Package 'baboon' ».
Nordholt, E. S. et J. H. VanHuijsduijnen (1997). « The treatment of item nonresponse during the editing of survey results », *New Techniques and Technologies for Statistics Ii*, p. 55-61.
Pedersen, Alma B., Ellen M. Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R. Kristensen, Tra My Pham, Lars Pedersen*, et al.* (2017). « Missing data and multiple imputation in clinical epidemiological research », *Clinical epidemiology*, vol. 9, p. 157-166.
Peugh, J. L. et C. K. Enders (2004). « Missing data in educational research: A review of reporting practices and suggestions for improvement », *Review of Educational Research*, vol. 74, no 4, p. 525-556.
Roth, P. L. (1994). « Missing data - a conceptual review for applied psychologists », *Personnel Psychology*, vol. 47, no 3, p. 537-560.
Rubin, Donald B (1976). « Inference and missing data », *Biometrika*, vol. 63, no 3, p. 581-592.
Rubin, Donald B (1996). « Multiple imputation after 18+ years », *Journal of the American statistical Association*, vol. 91, no 434, p. 473-489.
Sande, Innis G (1983). « Hot-deck imputation procedures », *Incomplete data in sample surveys*, vol. 3, p. 339-349.
Schafer, J. L. (1997). « Analysis of incomplete multivariate data ».

Schafer, J. L. et J. W. Graham (2002). « Missing data: Our view of the state of the art », *Psychological Methods*, vol. 7, no 2, p. 147-177.
Schouten, Rianne Margaretha, Peter Lugtig et Gerko Vink (2018). « Generating missing values for simulation purposes: A multivariate amputation procedure », *Journal of Statistical Computation and Simulation*, vol. 88, no 15, p. 2909-2930.
Schwender, Holger (2012). « Imputing missing genotypes with weighted k nearest neighbors », *Journal of Toxicology and Environmental Health, Part A*, vol. 75, no 8-10, p. 438-446.
Stacklies, Wolfram, Henning Redestig, Matthias Scholz, Dirk Walther et Joachim Selbig (2007). « Pcamethods—a bioconductor package providing pca methods for incomplete data », *Bioinformatics*, vol. 23, no 9, p. 1164-1167.
Stekhoven, Daniel J et Peter Bühlmann (2011). « Missforest—non-parametric missing value imputation for mixed-type data », *Bioinformatics*, vol. 28, no 1, p. 112-118.
Su, Yu-Sung, Andrew Gelman, Jennifer Hill et Masanao Yajima (2011). « Multiple imputation with diagnostics (mi) in r: Opening windows into the black box », *Journal of Statistical Software*, vol. 45, no 2, p. 1-31.
Tanner, M. A. et H. W. Wing (1987). « The calculation of posterior distributions by data augmentation », *Journal of the American Statistical Association*, vol. 82, no 398, p. 528-540.
Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani*, et al.* (2001). « Missing value estimation methods for DNA microarrays », *Bioinformatics*, vol. 17, no 6, p. 520-525.
van Buuren, S. (2007). « Multiple imputation of discrete and continuous data by fully conditional specification », *Statistical Methods in Medical Research*, vol. 16, no 3, p. 219-242.
van Buuren, S. et K. Groothuis-Oudshoorn (2011). « Mice: Multivariate imputation by chained equations in r », *Journal of Statistical Software*, vol. 45, no 3, p. 1-67.
Van Buuren, Stef et Karin Oudshoorn (1999). « Flexible multivariate imputation by mice », *Leiden, The Netherlands: TNO Prevention Center*.
Wilks, Samuel S (1932). « Certain generalizations in the analysis of variance », *Biometrika*, p. 471-494.
Yates, Frank (1933). « The analysis of replicated experiments when the field results are incomplete », *Empire Journal of Experimental Agriculture*, vol. 1, no 2, p. 129-142.
Yuan, Y. (2011). « Multiple imputation using sas software », *Journal of Statistical Software*, vol. 45, no 6, p. 1-25.

Annexe1 : Configuration des DM

Missing Data Patterns

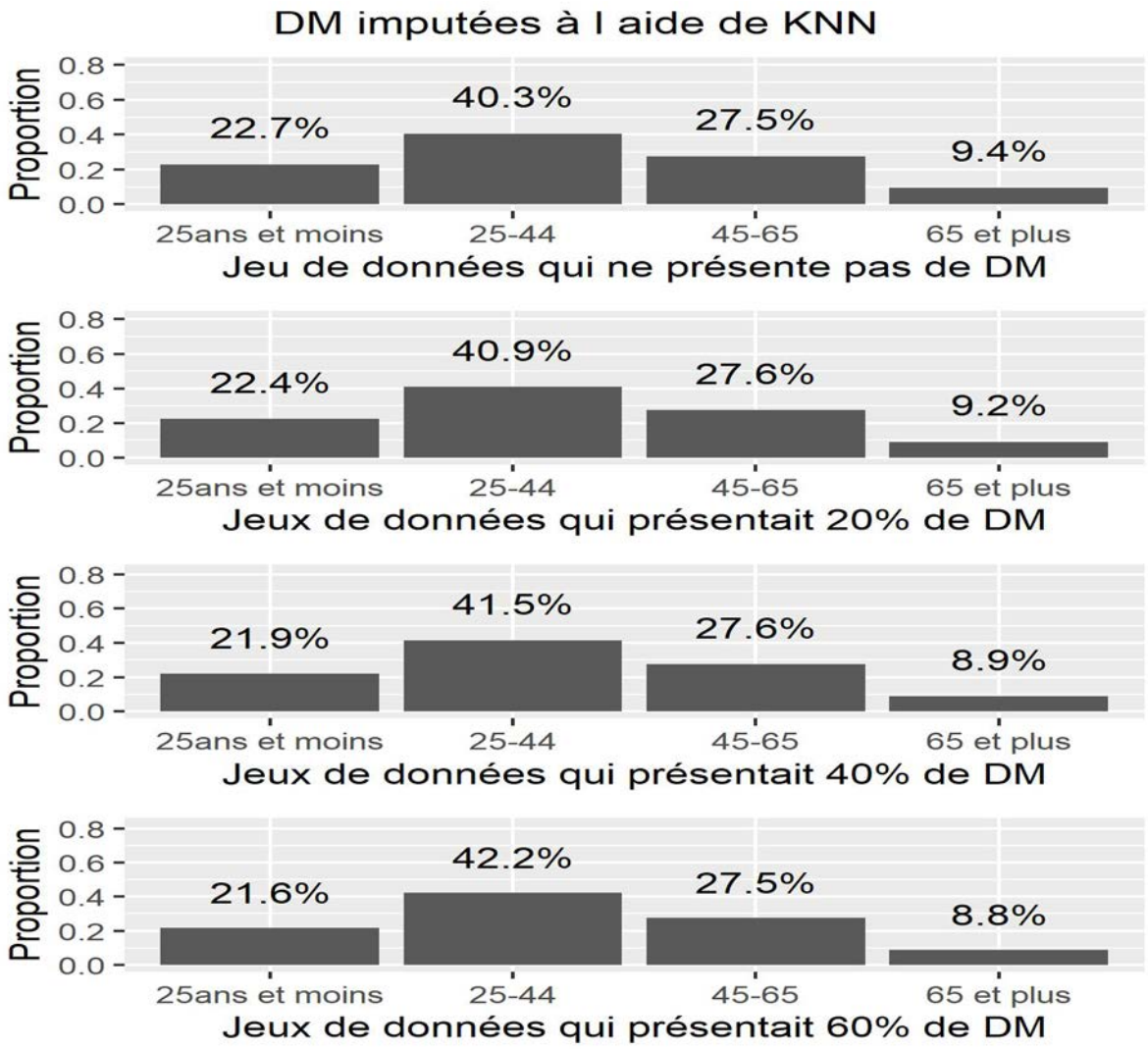
P_AGE_N	P_ID_N	P_YLIC_N1	V_BSWL_N	V_DISPI_N	V_DISPL_N	V_MYEAR_N	V_WHLB1_N	V_WHLB2_N1	V_YEAR_N1	Safety_Belt	P_EJCT	V_LICJ	P_PUSE
X	X	X	X	X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X	X	.
X	X	X	X	X	X	X	X	.	X	X	X	X	X
X	X	X	X	X	X	X	X	.	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
.	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	.	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	.	X	X	.
X	X	.	.	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	X	X	X	X	X	X	X	X	X	X	X	.
X	X	X	X	X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	.	X	.	X	X	X
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	.	X	X	.
.	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	.	X	X	X	X	.
X	X	.	.	X	X	X	X	X	X	X	X	X	.
X	X	.	.	X	X	X	X	X	X	X	X	X	.
X	X	.	X	X	X	X	X	X	X	X	X	X	.

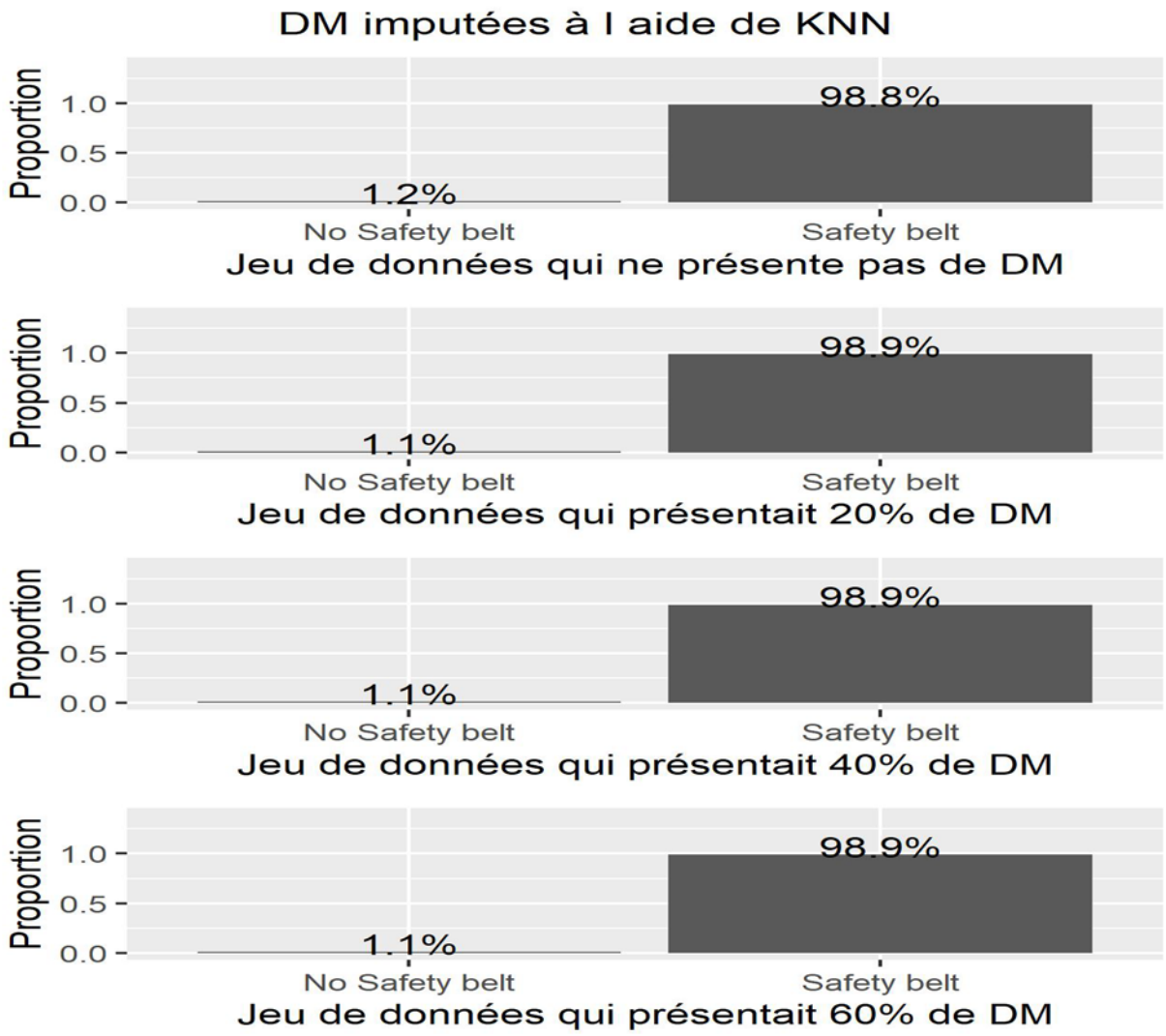
Annexe 2 : Variables auxiliaires avec leur importance dans les prédictions des variables du modèle Fredette et al. (2008)

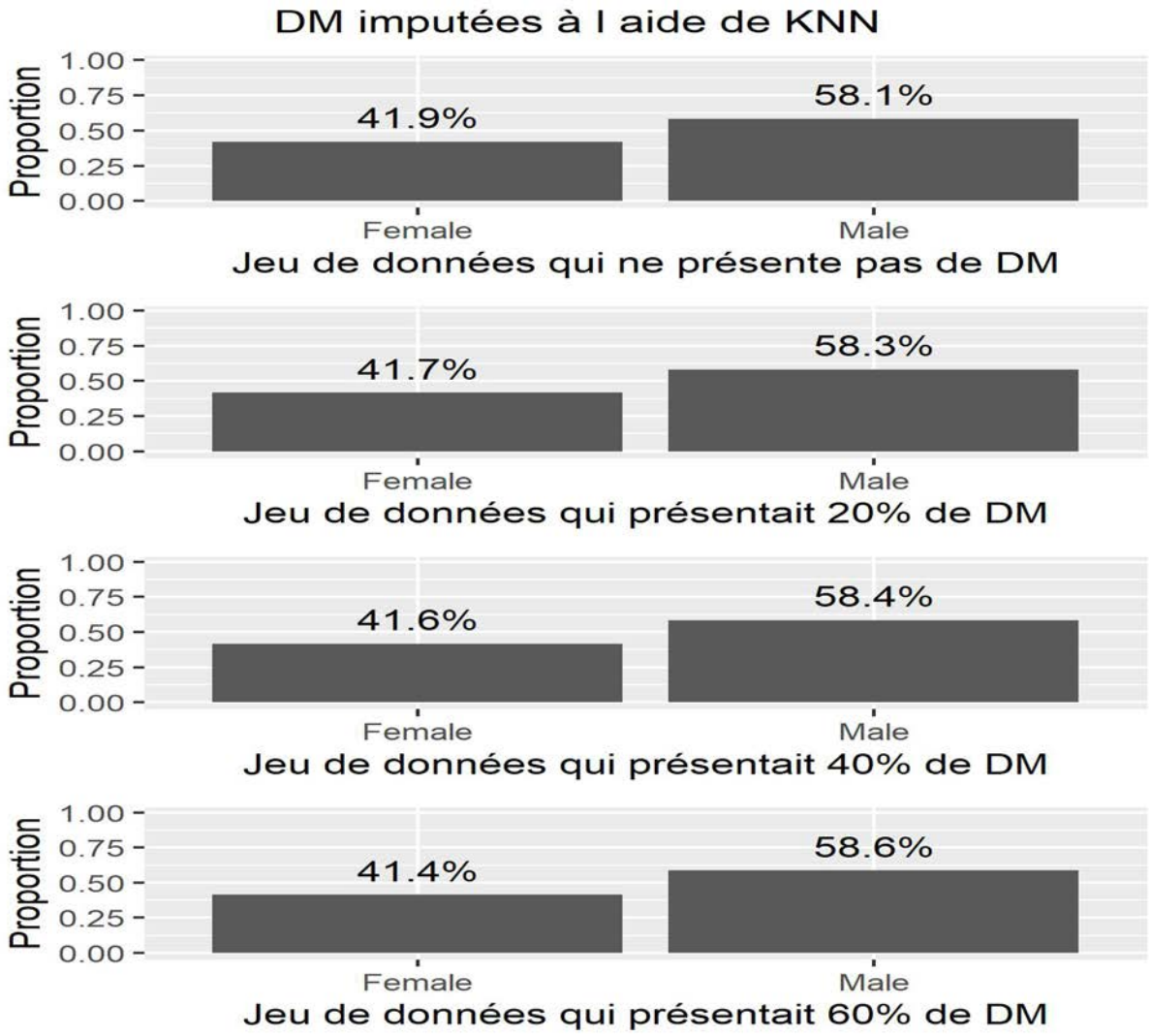
Variable à prédire	Variable prédictrice	Description	score d'Importance
Driver Vehicle Type	V_GVWC	** GVW/CYCLES **	1
Driver Vehicle Type	V_SEG	** SEGMENTATION CODE **	0.8165
Driver Vehicle Type	V_CAB	** CAB CONFIGURATION **	0.4361
Driver Vehicle Type	V_TRANS	** TRANSMISSION **	0.1251
Vehicle Mass	V_SEG	** SEGMENTATION CODE **	1
Vehicle Mass	V_SERIES	** SERIES NAME **	0.6479
Vehicle Mass	V_WHEELB	** WHEEL BASES **	0.5867
Vehicle Mass	V_WHLB1_N		0.2288
Vehicle Mass	V_DISPI_N		0.1773
Driver Sex	V_TYPE	** VEHICLE TYPE **	1
Driver Sex	P_ISEV	** INJURY SEVERITY OF PERSON **	0.3409
Driver Sex	V_WHEELD	** DRIVING WHEELS **	0.3185
Driver Sex	V_BODYT	** BODY TYPE **	0.3047
Driver Sex	P_DLIC	** PROVINCE OF DRIVER LICENSE **	0.2688
Driver Sex	V_MYEAR_N		0.2472
Driver Sex	V_SEG	** SEGMENTATION CODE **	0.2185
Driver Sex	V_SERIES	** SERIES NAME **	0.1431
Driver Sex	C_INJ_N		0.1351
Driver Sex	C_PROV	** PROVINCE OF COLLISION **	0.1289
Driver Sex	P_AGE_N		0.0929
Driver Age	P_YLIC_N		1
Driver Age	V_SEG	** SEGMENTATION CODE **	0.2957
Driver Age	V_DISPL_N		0.2786
Driver Age	V_SERIES	** SERIES NAME **	0.2272
Driver Age	C_Hour_N		0.2203
Driver Age	C_LITE	** LIGHT CONDITIONS **	0.1841
Driver Age	V_WHEELB	** WHEEL BASES **	0.1794
Driver Age	P_YLIC	** YRS LICENSED IN JURISDICTION **	0.1385
Authorized Speed	C_RCL1	** ROAD CLASSIFICATION 1 **	1
Authorized Speed	C_RMTL	** ROAD MATERIAL TYPE **	0.9968
Authorized Speed	C_SCATT	** SCENE ATTENDED **	0.5503
Authorized Speed	C_PROV	** PROVINCE OF COLLISION **	0.4794
Authorized Speed	C_RCFG	** ROADWAY CONFIGURATION **	0.2396
Authorized Speed	V_DIR	** DIRECTION OF TRAVEL **	0.1979

Variable à prédire	Variable prédictrice	Description	score d'Importance
Authorized Speed	P_AGE_N		0.1519
Authorized Speed	C_RCL2	** ROAD CLASSIFICATION 2 **	0.1372
Collision	V_MNVR	** VEHICLE MANOEUVRE **	1
Collision	C_PROV	** PROVINCE OF COLLISION **	0.4361
Collision	C_RCFG	** ROADWAY CONFIGURATION **	0.4301
Collision	C_TRAF	** TRAFFIC CONTROL TYPE **	0.3283
Collision	C_SCATT	** SCENE ATTENDED **	0.2105
Safety	V_SERIES	** SERIES NAME **	1
Safety	Driver_Vehicle_Type		0.1646
Safety	C_PROV	** PROVINCE OF COLLISION **	0.1426
Safety	P_EJCT	** OCCUPANT EJECTION FROM VEHICLE **	0.1323
Safety	Major_Fat		0.1251
Safety	V_USE	** VEHICLE USE **	0.1102
Safety	P_ABAG	** AIR BAG DEPLOYMENT **	0.1021

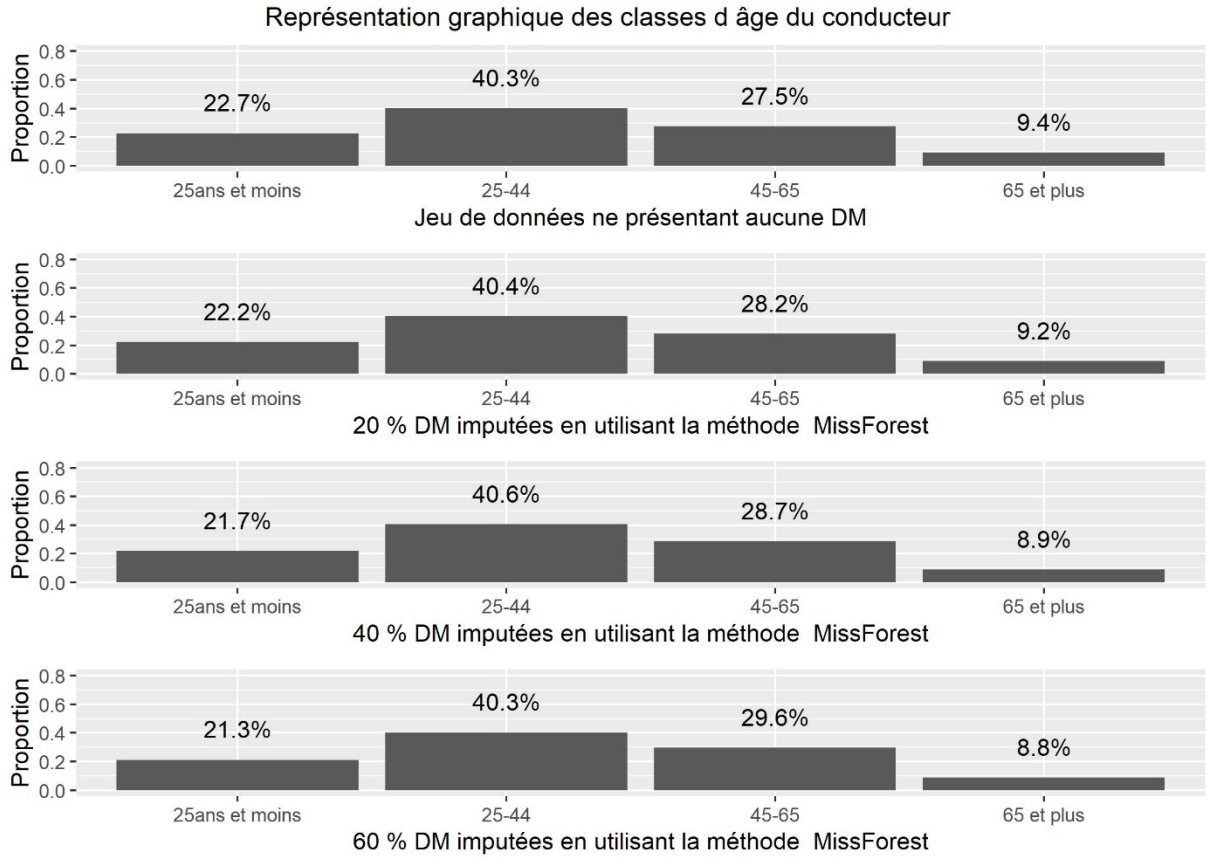
Annexe 3 : Représentation graphiques des variables : Jeu de données imputé à l'aide de KNN

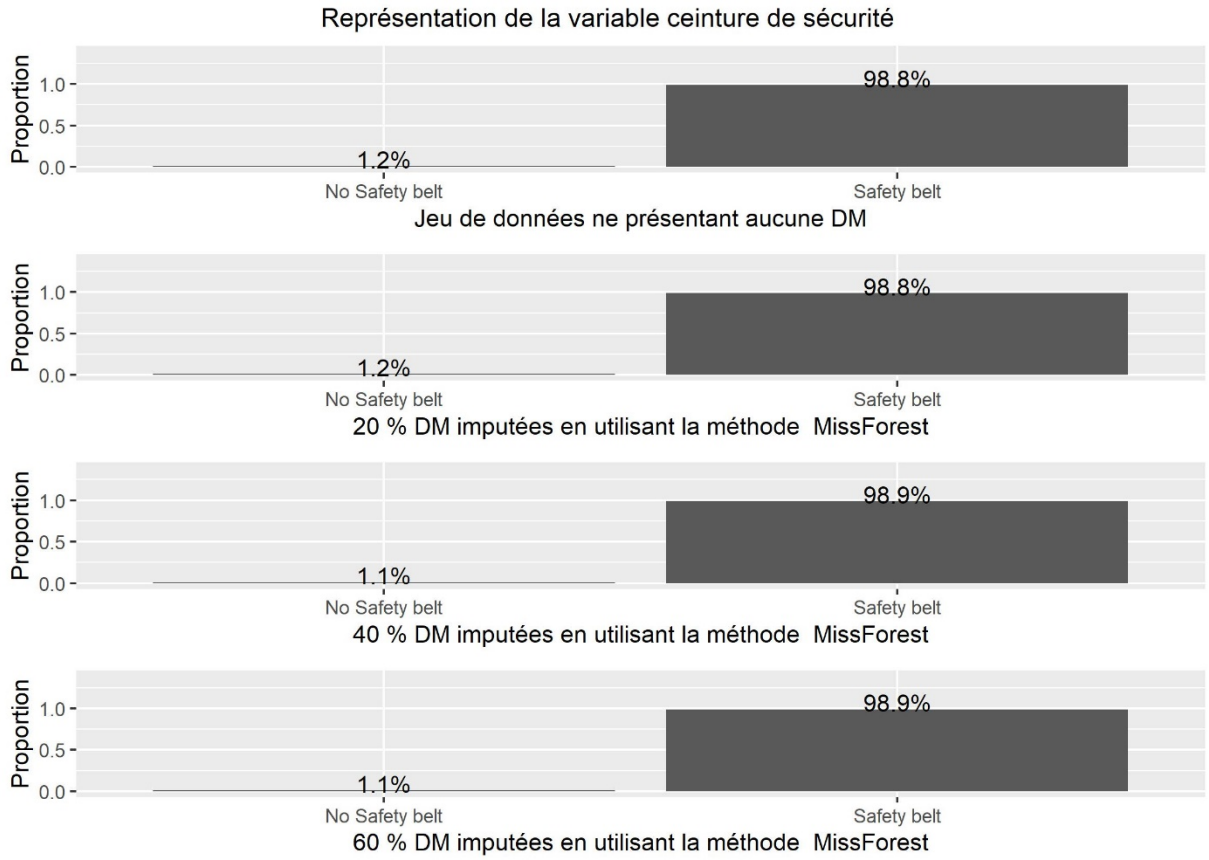


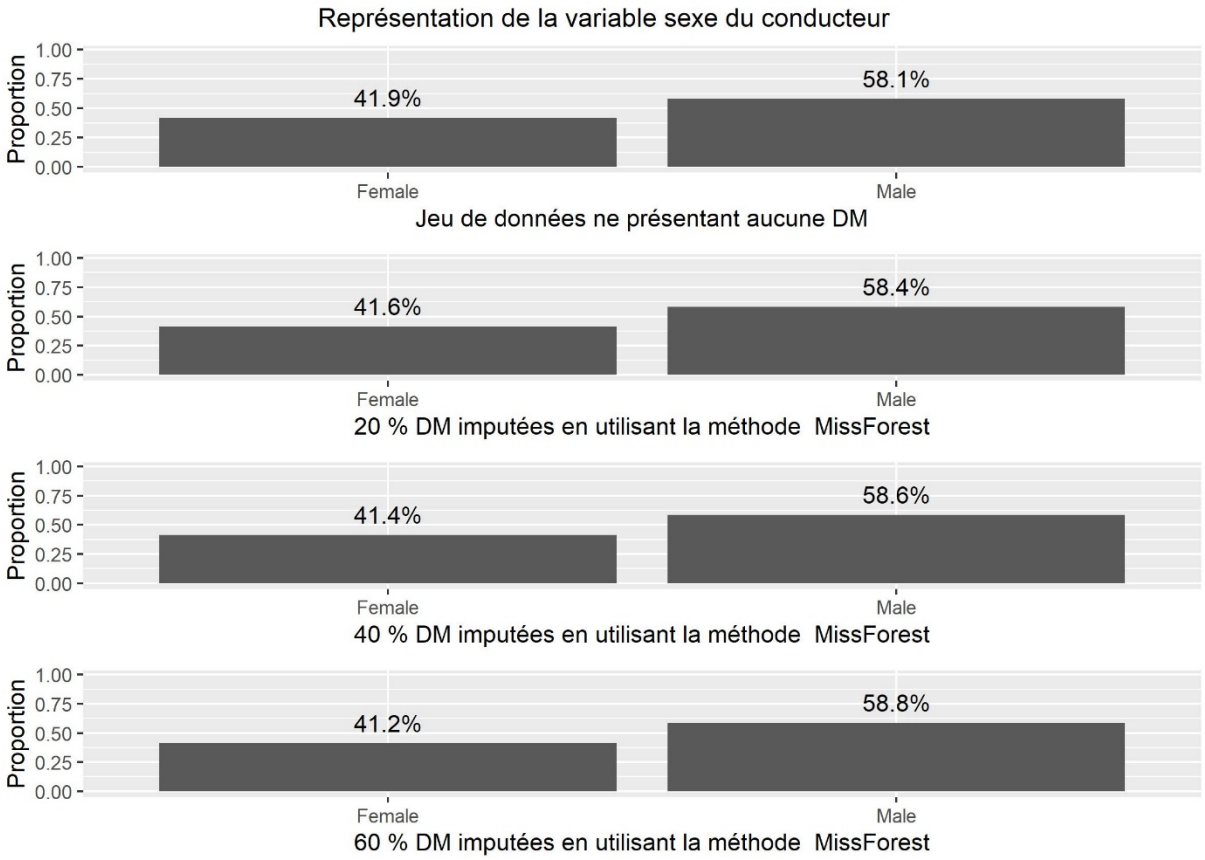

DM imputées à l aide de KNN
Proportion
1.0
0.5
0.0
98.8%
1.2%
No Safety belt
Safety belt
Jeu de données qui ne présente pas de DM
Proportion
1.0
0.5
0.0
98.9%
1.1%
No Safety belt
Safety belt
Jeu de données qui présentait 20% de DM
Proportion
1.0
0.5
0.0
98.9%
1.1%
No Safety belt
Safety belt
Jeu de données qui présentait 40% de DM
Proportion
1.0
0.5
0.0
98.9%
1.1%
No Safety belt
Safety belt
Jeu de données qui présentait 60% de DM


DM imputées à l aide de KNN
Proportion
1.00
0.75
0.50
0.25
0.00
41.9%
58.1%
Female
Male
Jeu de données qui ne présente pas de DM
Proportion
1.00
0.75
0.50
0.25
0.00
41.7%
58.3%
Female
Male
Jeu de données qui présentait 20% de DM
Proportion
1.00
0.75
0.50
0.25
0.00
41.6%
58.4%
Female
Male
Jeu de données qui présentait 40% de DM
Proportion
1.00
0.75
0.50
0.25
0.00
41.4%
58.6%
Female
Male
Jeu de données qui présentait 60% de DM

Annexe 4 : Représentations graphiques des variables : Jeu de données imputé à l'aide de Missforest.




Représentation de la variable ceinture de sécurité
Proportion
1.0
0.5
0.0
98.8%
1.2%
No Safety belt
Safety belt
Jeu de données ne présentant aucune DM
98.8%
1.2%
20 % DM imputées en utilisant la méthode MissForest
98.9%
1.1%
40 % DM imputées en utilisant la méthode MissForest
98.9%
1.1%
60 % DM imputées en utilisant la méthode MissForest


Représentation de la variable sexe du conducteur
Proportion
1.00
0.75
0.50
0.25
0.00
41.9%
58.1%
Female
Male
Jeu de données ne présentant aucune DM
41.6%
58.4%
20 % DM imputées en utilisant la méthode MissForest
41.4%
58.6%
40 % DM imputées en utilisant la méthode MissForest
41.2%
58.8%
60 % DM imputées en utilisant la méthode MissForest

Annexe 5 : Comparaison des valeurs générées par l'IM et des valeurs initiales

		(les pourcentages des bonnes catégories prédites sont en diagonale)					
		Car	**Heavy**	**Minivan**	**Pickup**	**SUV**	**Total**
Échantillon, qui représente 20% du jeu de données, dont les DM ont été imputées à l'aide de l'imputation multiple	**Car**	**8895** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	8895
	Heavy Truck	0 0.00	**291** **100.00**	0 0.00	0 0.00	0 0.00	291
	Minivan	0 0.00	0 0.00	**1805** **100.00**	0 0.00	0 0.00	1805
	Pickup	0 0.00	0 0.00	0 0.00	**2868** **100.00**	0 0.00	2868
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**816** **100.00**	816
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide de l'imputation multiple	**Car**	**19575** **99.83**	24 0.12	8 0.04	0 0.00	2 0.01	19609
	Heavy Truck	4 1.14	**336** **95.45**	12 3.41	0 0.00	0 0.00	352
	Minivan	0 0.00	8 0.27	**2997** **99.37**	11 0.36	0 0.00	3016
	Pickup	0 0.00	0 0.00	8 0.15	**5460** **99.71**	8 0.15	5476
	SUV	0 0.00	0 0.00	0 0.00	1 0.08	**1247** **99.92**	1248
Échantillon, qui représente 60% du jeu de données, dont les DM ont été imputées à l'aide de l'imputation multiple	**Car**	**26121** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	26121
	Heavy Truck	0 0.00	**331** **100.00**	0 0.00	0 0.00	0 0.00	331
	Minivan	0 0.00	0 0.00	**3160** **100.00**	0 0.00	0 0.00	3160
	Pickup	0 0.00	0 0.00	0 0.00	**6361** **100.00**	0 0.00	6361
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**1308** **100.00**	1308

		Catégories prédites *(les pourcentages des bonnes catégories prédites sont en diagonale)*		
		No Safety belt	**Safety belt**	**Total**
Échantillon représentant 20% de DM imputées à l'aide de l'imputation multiple.	**No Safety belt**	**12** **30.25**	27 69.75	39
	Safety belt	85 0.6	**14551** **99.4**	14636
	Total	97	14637	14675
Échantillon représentant 40% de DM imputées à l'aide de l'imputation multiple.	**No Safety belt**	**31** **47**	36 53	67
	Safety belt	487 1.64	**29147** **98.36**	29634
	Total	518	29183	29701
Échantillon représentant 60% de DM imputées à l'aide de l'imputation multiple.	**No Safety belt**	**47** **37.30**	79 62.70	126
	Safety belt	290 0.78	**36865** **99.22**	37155
	Total	337	36944	37281

		(les taux des bons classements sont en diagonale)					
		100 kph	**50-60 kph**	**70-80 kph**	**90 kph**	**> 50**	**Total**
Échantillon représentant 20% de DM imputées à l'aide de l'IM	**100 kph**	**956** **93.43**	53 5.17	5 0.50	3 0.27	4 0.63	1021
	50-60 kph	79 0.77	**9807** **95.58**	233 2.27	75 0.71	67 0.65	10261
	70-80 kph	13 .58	244 10.94	**1942** **87.28**	13 0.59	13 0.59	2225
	90 kph	3 0.33	48 6.09	10 1.27	**718** **91.6**	5 0.68	784
	Less than 50 kph	3 0.83	69 18.02	15 3.9	8 2	**289** **75.15**	384
Échantillon représentant 40% de DM imputées à l'aide de l'IM	**100 kph**	**1561** **91.45**	112 6.35	11 0.69	4 0.28	20 1.23	1708
	50-60 kph	524 2.45	**19884** **93.10**	387 1.81	128 0.6	435 2.04	21358
	70-80 kph	84 1.98	362 4.48	**3732** **87.47**	23 0.51	66 1.56	4267
	90 kph	22 1.43	96 6.25	22 1.45	**1366** **89.35**	**23** **1.52**	1529
	Less than 50 kph	28 3.38	114 13.68	26 3.17	12 1.43	**659** **78.62**	839
Échantillon représentant 60% de DM imputées à l'aide de l'IM	**100 kph**	**1673** **87.21**	206 10.77	26 1.35	7 0.34	6 0.33	1918
	50-60 kph	247 0.9	**26148** **95.05**	702 2.55	217 0.79	196 0.71	27510
	70-80 kph	35 0.7	703 13.99	**4193** **83.48**	48 0.95	44 0.88	5022
	90 kph	9 0.51	193 11.10	36 2.06	**1488** **85.55**	14 0.78	1740
	Less than 50 kph	11 1.03	198 18.11	43 3.9	17 1.52	**822** **75.44**	1091

	Variable		N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Age	original	14675	45.55	17.03	16.00	95.00
	Age imputé		14675	45.52	17.20	16.00	95.00
Échantillon représentant 40% du jeu de données initial.	Age	original	29701	44.20	16.99	16.00	94
	Age imputé		29701	44.26	17.07	16.00	94.2
Échantillon représentant 60% du jeu de données initial.	Age	original	37281	43.18	16.94	16	95
	Age imputé		37281	43.18	16.98	16	95

	Variable		N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Poids	original	14675	3494.97	917.94	353.00	9947.00
	Poids imputé		14675	3492.86	915.70	353.00	9947.00
Échantillon représentant 40% du jeu de données initial.	Poids	original	29701	3402	874.78	353	9947
	Poids imputé		29701	3403	877.40	281	9947
Échantillon représentant 60% du jeu de données initial.	Poids	original	37281	3334.75	849.54	572	9947
	Poids imputé		37281	3334.75	849.54	572	9947

Annexe 6 : Comparaison des valeurs générées par le module KNN et des valeurs initiales (*matrice de confusion et comparaison des distributions des variables continues)*

	Valeurs initiales avant l'introduction des DM	**Catégories prédites** *(les pourcentages des bonnes catégories prédites sont en diagonale)*		
		No Safety belt	**Safety belt**	**Total**
Échantillon représentant 20% de DM imputées à l'aide de KNN	**No Safety belt**	**11** **28.21**	28 71.79	39
	Safety belt	0 0.00	**14636** **100.00**	14636
	Total	11	14664	14675
Échantillon représentant 40% de DM imputées à l'aide de KNN	**No Safety belt**	**25** **37.31**	42 62.69	67
	Safety belt	1 0.00	**29633** **100.00**	**29634**
	Total	26	29675	29701
Échantillon représentant 60% de DM imputées à l'aide de KNN	**No Safety belt**	**40** **31.75**	86 68.25	126
	Safety belt	0 0.00	**37155** **100.00**	37155
	Total	40	37241	37281

	Valeurs initiales avant introduction des DM	Catégories prédites *(les pourcentages des bonnes catégories prédites sont en diagonale)*					
		Car	**Heavy**	**Minivan**	**Pickup**	**SUV**	**Total**
Échantillon, qui représente 20% du jeu de données, dont les DM ont été imputées à l'aide de KNN.	**Car**	**8892** **99.97**	0 0.00	1 0.01	2 0.02	0 0.00	8895
	Heavy Truck	1 0.34	**287** **98.63**	0 0.00	3 1.03	0 0.00	291
	Minivan	4 0.22	0 0.00	**1801** **99.78**	0 0.00	0 0.00	1805
	Pickup	10 0.35	0 0.00	0 0.00	**2858** **99.65**	0 0.00	2868
	SUV	2 0.25	0 0.00	0 0.00	1 0.12	**813** **99.63**	816
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide de KNN.	**Car**	**19606** **99.98**	0 0.00	0 0.00	3 0.02	0 0.00	19609
	Heavy Truck	2 0.57	**345** **98.01**	1 0.28	4 1.14	0 0.00	352
	Minivan	4 0.13	0 0.00	**3006** **99.67**	6 0.20	0 0.00	3016
	Pickup	24 0.44	0 0.00	0 0.00	**5451** **99.54**	1 0.02	5476
	SUV	2 0.16	0 0.00	0 0.00	3 0.24	**1243** **99.60**	1248
Échantillon, qui représente 60% du jeu de données, dont les DM ont été imputées à l'aide de KNN.)	**Car**	**26110** **99.96**	0 0.00	0 0.00	11 0.04	0 0.00	26121
	Heavy Truck	3 0.91	**325** **98.19**	1 0.30	2 0.60	0 0.00	331
	Minivan	11 0.35	0 0.00	**3143** **99.46**	5 0.16	1 0.03	3160
	Pickup	35 0.55	1 0.02	2 0.03	**6320** **99.36**	3 0.05	6361
	SUV	10 0.76	0 0.00	0 0.00	2 0.15	**1296** **99.08**	1308

		Catégories prédites *(les pourcentages des bonnes catégories prédites sont en diagonale)*					
		100 kph	**50-60 kph**	**70-80 kph**	**90 kph**	**Less than 50**	**Total**
Échantillon représentant 40% de DM imputées à l'aide de KNN	**100 kph**	**953** **93.34**	60 5.88	6 0.59	2 0.20	0 0.00	1021
	50-60 kph	0 0.00	**10009** **97.54**	217 2.11	29 0.28	6 0.06	10261
	70-80 kph	0 0.00	284 12.76	**1938** **87.10**	2 0.09	1 0.04	2225
	90 kph	0 0.00	64 8.16	5 0.64	**715** **91.20**	0 0.00	784
	Less than 50 kph	0 0.00	93 24.22	8 2.08	0 0.00	**283** **73.70**	384
Échantillon représentant 40% de DM imputées à l'aide de KNN	**100 kph**	**1539** **90.11**	149 8.72	20 1.17	0 0.00	0 0.00	1708
	50-60 kph	6 0.03	**20972** **98.19**	368 1.72	7 0.03	5 0.02	21358
	70-80 kph	0 0.00	541 12.68	**3723** **87.25**	3 0.07	0 0.00	4267
	90 kph	0 0.00	159 10.40	12 0.78	**1357** **88.75**	1 0.07	1529
	Less than 50 kph	0 0.00	184 21.93	19 2.26	0 0.00	**636** **75.80**	839
Échantillon représentant 60% de DM imputées à l'aide de KNN	**100 kph**	**1657** **86.39**	204 10.64	54 2.82	3 0.16	0 0.00	1918
	50-60 kph	1 0.00	**26509** **96.36**	854 3.10	117 0.43	29 0.11	27510
	70-80 kph	0 0.00	764 15.21	**4219** **83.99**	30 0.60	10 0.20	5023
	90 kph	0 0.00	217 12.48	39 2.24	**1482** **85.22**	1 0.06	1739
	Less than 50 kph	0 0.00	233 21.36	35 3.21	3 0.27	**820** **75.16**	1091

	Variable		N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Age	original	14675	45.55	17.03	16.00	95.00
	Age imputé		14675	45.425	16.21	16.00	95.00
Échantillon représentant 40% du jeu de données initial.	Age	original	29701	44.20	16.99	16.00	94.00
	Age imputé		29701	44.03	16.14	16.00	94.00
Échantillon représentant 60% du jeu de données initial.	Age	original	37281	43.19	16.94	16.00	95.00
	Age imputé		37281	42.97	16.01	16.00	95.00

	Variable		N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Poids	original	14675	3491.77	911.68	353.00	9947.00
	Poids impute		14675	3494.97	917.14	353.00	9947.00
Échantillon représentant 40% du jeu de données initial.	Poids	original	29701	3402.15	871.52	353.00	9947.00
	Poids imputé		29701	3402.21	874.78	353.00	9947.00
Échantillon représentant 60% du jeu de données initial.	Poids	original	37281	3334.76	849.55	572.00	9947.00
	Poids imputé		37281	3334.76	849.55	572.00	9947.00

Annexe 7 : Comparaison des valeurs générées par le module MissForest et des valeurs initiales (*matrice de confusion et comparaison des distributions des variables continues)*

		Catégories prédites *(les taux des bons classements sont en diagonale)*		
		No Safety belt	**Safety belt**	**Total**
Échantillon représentant 20% de DM imputées à l'aide de MissForest	**No Safety belt**	**38** **97.44**	1 2.56	39
	Safety belt	0 0.00	**14636** **100.00**	14636
Échantillon représentant 40% de DM imputées à l'aide de KNN	**No Safety belt**	**62** **92.54**	5 7.46	67
	Safety belt	0 0.00	**29634** **100.00**	29634
Échantillon représentant 60% de DM imputées à l'aide de KNN	**No Safety belt**	**115** **91.27**	11 8.73	126
	Safety belt	1 0.00	**37154** **100.00**	37155

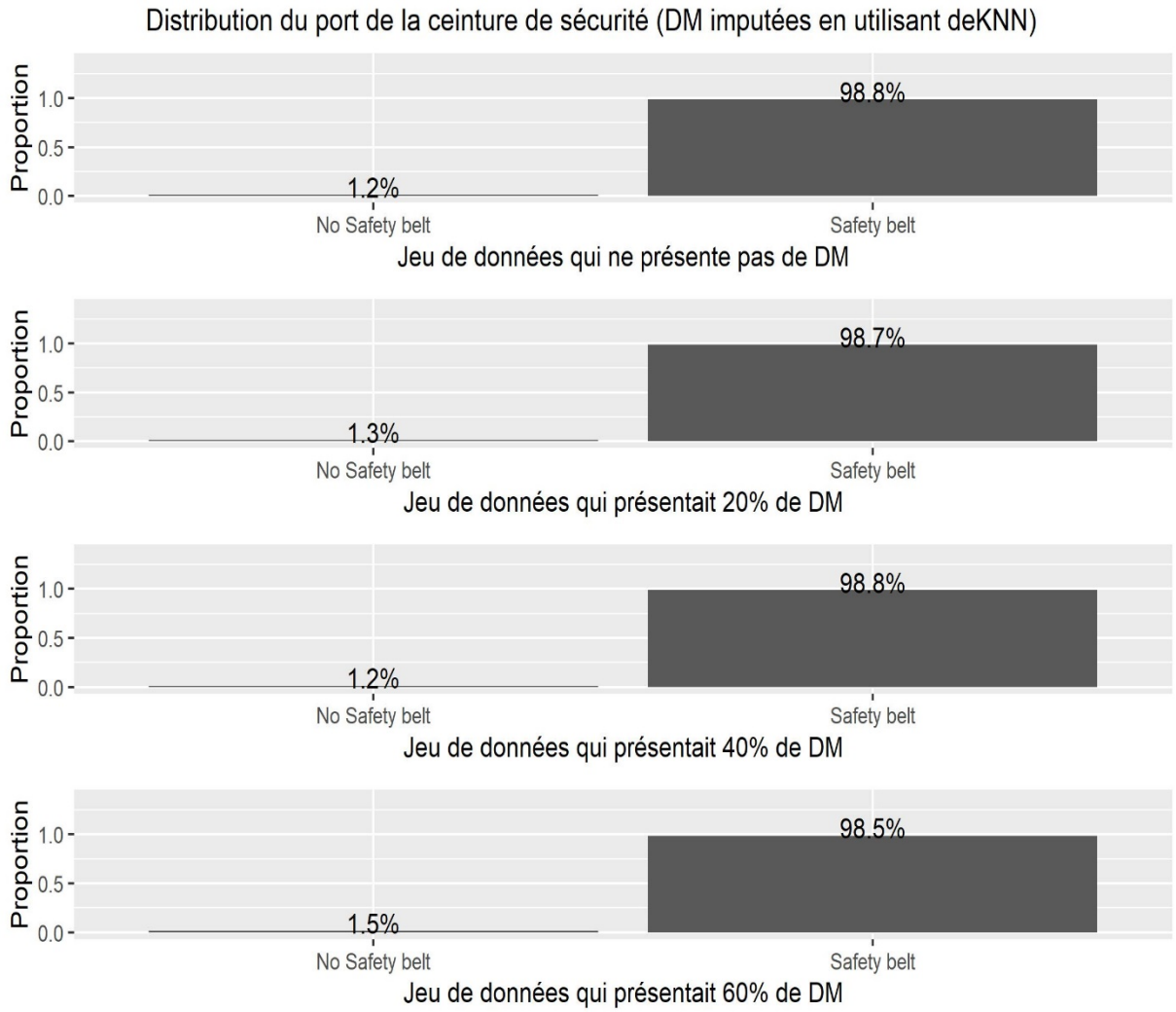
		Catégories prédites *(les taux des bons classements sont en diagonale)*					
		Car	**Heavy**	**Minivan**	**Pickup**	**SUV**	**Total**
Échantillon, qui représente 20% du jeu de données, dont les DM ont été imputées à l'aide du module MissForest.	**Car**	**8895** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	8895
	Heavy Truck	0 0.00	**290** **99.66**	1 0.34	0 0.00	0 0.00	291
	Minivan	0 0.00	0 0.00	**1805** **100.00**	0 0.00	0 0.00	1805
	Pickup	0 0.00	0 0.00	0 0.00	**2868** **100.00**	0 0.00	2868
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**816** **100.00**	816
Échantillon, qui représente 40% du jeu de données, dont les DM ont été imputées à l'aide du module MissForest.	**Car**	**19609** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	19609
	Heavy Truck	0 0.00	**349** **99.15**	1 0.28	2 0.57	0 0.00	352
	Minivan	1 0.03	0 0.00	**3015** **99.97**	0 0.00	0 0.00	3016
	Pickup	0 0.00	0 0.00	0 0.00	**5476** **100.00**	0 0.00	5476
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**1248** **100.00**	1248
Échantillon, qui représente 60% du jeu de données, dont les DM ont été imputées à l'aide du module MissForest.	**Car**	**26121** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	26121
	Heavy Truck	1 0.30	**325** **98.19**	4 1.21	1 0.30	0 0.00	331
	Minivan	0 0.00	0 0.00	**3160** **100.00**	0 0.00	0 0.00	3160
	Pickup	0 0.00	0 0.00	0 0.00	**6361** **100.00**	0 0.00	6361
	SUV	0 0.00	0 0.00	0 0.00	0 0.00	**1308** **100.00**	1308

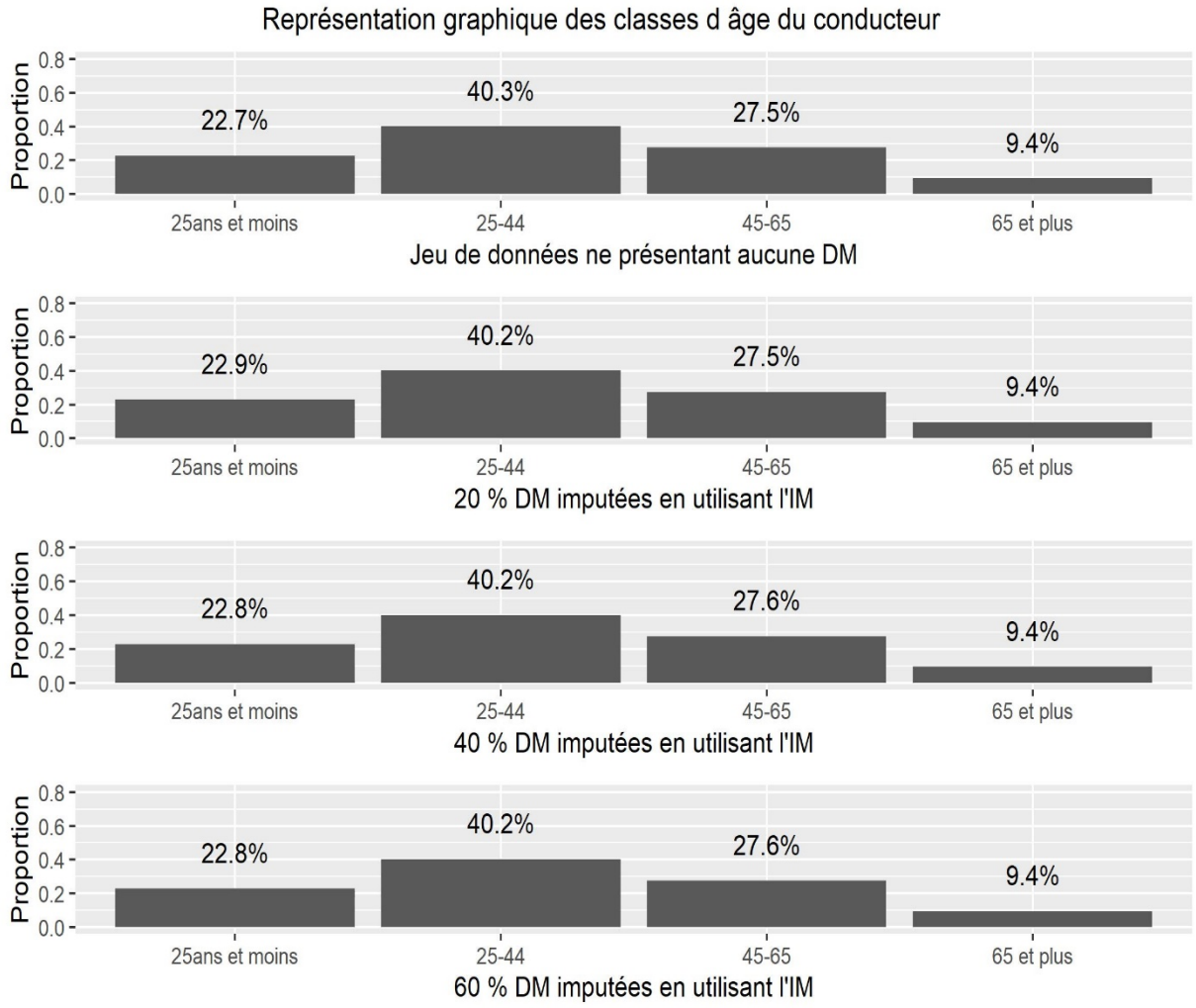
	Variable	N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Poids original	14675	3494.01	914.97	353.00	9947.00
	Poids imputé	14675	3494.97	917.14	353.00	9947.00
Échantillon représentant 40% du jeu de données initial.	Poids original	29701	3401.68	871.10	353.00	9947.00
	Poids imputé	29701	3402.21	874.78	353.00	9947.00
Échantillon représentant 60% du jeu de données initial.	Poids original	37281	3333.96	845.41	572.00	9947.00
	Poids imputé	37281	3334.76	849.55	572.00	9947.00

	Variable	N	Mean	Std Dev	Minimum	Maximum
Échantillon représentant 20% du jeu de données initial.	Age original	14675	45.55	17.03	16.00	95.00
	Age imputé	14675	45.60	16.77	16.00	95.00
Échantillon représentant 40% du jeu de données initial.	Age original	29701	44.20	16.99	16.00	94.00
	Age imputé	29701	44.30	16.57	16.00	94.00
Échantillon représentant 60% du jeu de données initial.	Age original	37281	43.18	16.94	16.00	95.00
	Age imputé	37281	43.49	16.45	16.00	95.00

		(les taux des bons classements sont en diagonale)					
		100 kph	**50-60 kph**	**70-80 kph**	**90 kph**	**Less than 50**	**Total**
Échantillon représentant 20% de DM imputées par MissForest	**100 kph**	**992** **97.16**	26 2.55	2 0.20	1 0.10	0 0.00	1021
	50-60 kph	4 0.04	**10233** **99.73**	13 0.13	5 0.05	6 0.06	10261
	70-80 kph	0 0.00	79 3.55	**2140** **96.18**	5 0.22	1 0.04	2225
	90 kph	1 0.13	23 2.93	4 0.51	**756** **96.43**	0 0.00	784
	Less than 50 kph	0 0.00	9 2.34	0 0.00	0 0.00	**375** **97.66**	384
Échantillon représentant 40% de DM imputées par MissForest	**100 kph**	**1645** **96.31**	55 3.22	6 0.35	2 0.12	0 0.00	1708
	50-60 kph	14 0.07	**21272** **99.60**	36 0.17	21 0.10	15 0.07	21358
	70-80 kph	5 0.12	234 5.48	**4013** **94.05**	12 0.28	3 0.07	4267
	90 kph	1 0.07	58 3.79	8 0.52	**1461** **95.55**	1 0.07	1529
	Less than 50 kph	0 0.00	43 5.13	2 0.24	1 0.12	**793** **94.52**	839
Échantillon représentant 60% de DM imputées par MissForest	**100 kph**	**1794** **93.53**	116 6.05	4 0.21	4 0.21	0 0.00	1918
	50-60 kph	35 0.13	**27363** **99.47**	58 0.21	32 0.12	22 0.08	27510
	70-80 kph	11 0.22	383 7.62	**4601** **91.60**	25 0.50	3 0.06	5023
	90 kph	2 0.12	114 6.56	9 0.52	**1614** **92.81**	0 0.00	1739
	Less than 50 kph	0 0.00	89 8.16	1 0.09	2 0.18	**999** **91.57**	1091

Annexe 8 : Représentations graphiques des variables (*jeu de données imputées à l'aide FCS*).




Représentation graphique des classes d âge du conducteur
Proportion
0.8
0.6
0.4
0.2
0.0
22.7%
40.3%
27.5%
9.4%
25ans et moins
25-44
45-65
65 et plus
Jeu de données ne présentant aucune DM
22.9%
40.2%
27.5%
9.4%
20 % DM imputées en utilisant l'IM
22.8%
40.2%
27.6%
9.4%
40 % DM imputées en utilisant l'IM
22.8%
40.2%
27.6%
9.4%
60 % DM imputées en utilisant l'IM

Annexe 9 : Résultat du modèle Fredette et al.(2008) Appliqué au jeu de données 20% de DM imputé à l'aide de KNN

Model Information	
Data Set	WORK.FREDETTE_AL
Response Variable	Major_or_Fatal
Number of Response Levels	2
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	43570
Number of Observations Used	43570

Response Profile		
Ordered Value	**Major_or_Fatal**	**Total Frequency**
1	1	792
2	0	42778

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	7919.495	6419.503
SC	7928.177	6706.013
-2 Log L	7917.495	6353.503

R-Square	0.0353	Max-rescaled R-Square	0.2122

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1563.9917	32	<.0001
Score	2563.0553	32	<.0001
Wald	1449.7398	32	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Mass_Ratio	3	46.0934	<.0001
Drivers_Age	3	24.5619	<.0001
Drivers_sex	1	0.0742	0.7854
Authorized_Speed	4	369.1543	<.0001
Collision	8	563.2396	<.0001
Safety_Belt	1	221.9148	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Driver_Vehicle_Type	6	11.7346	0.0682
Driver2_Vehicle_Type	6	14.2450	0.0270

Analysis of Maximum Likelihood Estimates						
Parameter		**Major_or_Fatal**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**
Intercept		**1**	1	-4.4308	0.1149	1488.3131
Mass_Ratio	**(0.50;0.80]**	**1**	1	0.5167	0.0987	27.4251
Mass_Ratio	**(1.20;2.00]**	**1**	1	-0.3307	0.1080	9.3848
Mass_Ratio	**(2.00;+)**	**1**	1	-0.8139	0.3181	6.5452
Drivers_Age	**24 or less**	**1**	1	-0.1944	0.1107	3.0833
Drivers_Age	**45-64**	**1**	1	0.1775	0.0921	3.7096
Drivers_Age	**65 and +**	**1**	1	0.4156	0.1167	12.6757
Drivers_sex	**Male**	**1**	1	0.0215	0.0788	0.0742
Authorized_Speed	**100 kph**	**1**	1	1.2849	0.1752	53.7571
Authorized_Speed	**70-80 kph**	**1**	1	1.3957	0.0885	248.6588
Authorized_Speed	**90 kph**	**1**	1	1.7134	0.1195	205.5779
Authorized_Speed	**Less than 5**	**1**	1	-1.0400	0.3852	7.2879
Collision	**Head-On**	**1**	1	1.3236	0.1003	174.1126
Collision	**Leftturnconflict(diff.directio**	**1**	1	-0.1465	0.1341	1.1939
Collision	**Leftturnconflict(samedirection**	**1**	1	-0.1426	0.3293	0.1874
Collision	**Other**	**1**	1	-0.4607	0.1304	12.4747
Collision	**Rear-end**	**1**	1	-1.8524	0.1545	143.7027
Collision	**Rightturnconflict(diff.directi**	**1**	1	-0.2709	0.4221	0.4118

Analysis of Maximum Likelihood Estimates						
Parameter		**Major_or_Fatal**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**
Collision	**Rightturnconflict(samedirectio**	**1**	1	-0.5874	0.3680	2.5471
Collision	**Side-swipe(samedirection)**	**1**	1	-0.1254	0.1925	0.4246
Safety_Belt	**No Safety belt**	**1**	1	2.2138	0.1486	221.9148
Driver_Vehicle_Type	**Bus**	**1**	1	-0.6722	0.7890	0.7259
Driver_Vehicle_Type	**Heavy**	**1**	1	-1.3711	1.1667	1.3810
Driver_Vehicle_Type	**Miniv**	**1**	1	-0.3110	0.1495	4.3268
Driver_Vehicle_Type	**Other**	**1**	1	-0.2839	0.1201	5.5867
Driver_Vehicle_Type	**Picku**	**1**	1	-0.3644	0.1708	4.5536
Driver_Vehicle_Type	**SUV**	**1**	1	0.1553	1.0215	0.0231
Driver2_Vehicle_Type	**Bus**	**1**	1	1.4873	1.1426	1.6945
Driver2_Vehicle_Type	**Heavy**	**1**	1	0.7353	0.8022	0.8402
Driver2_Vehicle_Type	**Miniv**	**1**	1	0.0323	0.1380	0.0549
Driver2_Vehicle_Type	**Other**	**1**	1	0.1286	0.1072	1.4407
Driver2_Vehicle_Type	**Picku**	**1**	1	0.4144	0.1476	7.8824
Driver2_Vehicle_Type	**SUV**	**1**	1	1.1480	0.5299	4.6947

Analysis of Maximum Likelihood Estimates				
Parameter		**Major_or_Fatal**	**Pr > ChiSq**	**Exp(Est)**
Intercept		**1**	<.0001	0.012
Mass_Ratio	**(0.50;0.80]**	**1**	<.0001	1.677
Mass_Ratio	**(1.20;2.00]**	**1**	0.0022	0.718
Mass_Ratio	**(2.00;+)**	**1**	0.0105	0.443
Drivers_Age	**24 or less**	**1**	0.0791	0.823
Drivers_Age	**45-64**	**1**	0.0541	1.194

Analysis of Maximum Likelihood Estimates				
Parameter		**Major_or_Fatal**	**Pr > ChiSq**	**Exp(Est)**
Drivers_Age	**65 and +**	**1**	0.0004	1.515
Drivers_sex	**Male**	**1**	0.7854	1.022
Authorized_Speed	**100 kph**	**1**	<.0001	3.614
Authorized_Speed	**70-80 kph**	**1**	<.0001	4.038
Authorized_Speed	**90 kph**	**1**	<.0001	5.548
Authorized_Speed	**Less than 5**	**1**	0.0069	0.353
Collision	**Head-On**	**1**	<.0001	3.757
Collision	**Leftturnconflict(diff.directio**	**1**	0.2745	0.864
Collision	**Leftturnconflict(samedirection**	**1**	0.6651	0.867
Collision	**Other**	**1**	0.0004	0.631
Collision	**Rear-end**	**1**	<.0001	0.157
Collision	**Rightturnconflict(diff.directi**	**1**	0.5211	0.763
Collision	**Rightturnconflict(samedirectio**	**1**	0.1105	0.556
Collision	**Side-swipe(samedirection)**	**1**	0.5146	0.882
Safety_Belt	**No Safety belt**	**1**	<.0001	9.151
Driver_Vehicle_Type	**Bus**	**1**	0.3942	0.511
Driver_Vehicle_Type	**Heavy**	**1**	0.2399	0.254
Driver_Vehicle_Type	**Miniv**	**1**	0.0375	0.733
Driver_Vehicle_Type	**Other**	**1**	0.0181	0.753
Driver_Vehicle_Type	**Picku**	**1**	0.0328	0.695
Driver_Vehicle_Type	**SUV**	**1**	0.8792	1.168
Driver2_Vehicle_Type	**Bus**	**1**	0.1930	4.425
Driver2_Vehicle_Type	**Heavy**	**1**	0.3593	2.086
Driver2_Vehicle_Type	**Miniv**	**1**	0.8148	1.033
Driver2_Vehicle_Type	**Other**	**1**	0.2300	1.137

Analysis of Maximum Likelihood Estimates				
Parameter		**Major_or_Fatal**	**Pr > ChiSq**	**Exp(Est)**
Driver2_Vehicle_Type	**Picku**	**1**	0.0050	1.513
Driver2_Vehicle_Type	**SUV**	**1**	0.0303	3.152

Odds Ratio Estimates				
Effect	**Major_or_Fatal**	**Point Estimate**	**95% Wald Confidence Limits**	
Mass_Ratio (0.50;0.80] vs (0.80;1.20]	**1**	1.677	1.382	2.034
Mass_Ratio (1.20;2.00] vs (0.80;1.20]	**1**	0.718	0.581	0.888
Mass_Ratio (2.00;+) vs (0.80;1.20]	**1**	0.443	0.238	0.827
Drivers_Age 24 or less vs 25-44	**1**	0.823	0.663	1.023
Drivers_Age 45-64 vs 25-44	**1**	1.194	0.997	1.431
Drivers_Age 65 and + vs 25-44	**1**	1.515	1.205	1.905
Drivers_sex Male vs Female	**1**	1.022	0.875	1.192
Authorized_Speed 100 kph vs 50-60 kph	**1**	3.614	2.564	5.095
Authorized_Speed 70-80 kph vs 50-60 kph	**1**	4.038	3.395	4.803
Authorized_Speed 90 kph vs 50-60 kph	**1**	5.548	4.389	7.012
Authorized_Speed Less than 5 vs 50-60 kph	**1**	0.353	0.166	0.752
Collision Head-On vs Right-angle(sideundetermined)	**1**	3.757	3.086	4.573
Collision Leftturnconflict(diff.directio vs Right-angle(sideundetermined)	**1**	0.864	0.664	1.123
Collision Leftturnconflict(samedirection vs Right-angle(sideundetermined)	**1**	0.867	0.455	1.653
Collision Other vs Right-angle(sideundetermined)	**1**	0.631	0.489	0.815
Collision Rear-end vs Right-angle(sideundetermined)	**1**	0.157	0.116	0.212
Collision Rightturnconflict(diff.directi vs Right-angle(sideundetermined)	**1**	0.763	0.333	1.744

Odds Ratio Estimates				
Effect	**Major_or_Fatal**	**Point Estimate**	**95% Wald Confidence Limits**	
Collision Rightturnconflict(samedirectio vs Right-angle(sideundetermined)	**1**	0.556	0.270	1.143
Collision Side-swipe(samedirection) vs Right-angle(sideundetermined)	**1**	0.882	0.605	1.286
Safety_Belt No Safety belt vs Safety belt	**1**	9.151	6.838	12.245
Driver_Vehicle_Type Bus vs Car	**1**	0.511	0.109	2.397
Driver_Vehicle_Type Heavy vs Car	**1**	0.254	0.026	2.498
Driver_Vehicle_Type Miniv vs Car	**1**	0.733	0.547	0.982
Driver_Vehicle_Type Other vs Car	**1**	0.753	0.595	0.953
Driver_Vehicle_Type Picku vs Car	**1**	0.695	0.497	0.971
Driver_Vehicle_Type SUV vs Car	**1**	1.168	0.158	8.648
Driver2_Vehicle_Type Bus vs Car	**1**	4.425	0.471	41.541
Driver2_Vehicle_Type Heavy vs Car	**1**	2.086	0.433	10.049
Driver2_Vehicle_Type Miniv vs Car	**1**	1.033	0.788	1.354
Driver2_Vehicle_Type Other vs Car	**1**	1.137	0.922	1.403
Driver2_Vehicle_Type Picku vs Car	**1**	1.513	1.133	2.021
Driver2_Vehicle_Type SUV vs Car	**1**	3.152	1.116	8.904

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.1	**Somers' D**	0.663
Percent Discordant	16.8	**Gamma**	0.664
Percent Tied	0.1	**Tau-a**	0.024
Pairs	33880176	**c**	0.832

Partition for the Hosmer and Lemeshow Test					
		Major_or_Fatal = 1		Major_or_Fatal = 0	
Group	Total	Observed	Expected	Observed	Expected
1	4244	3	4.60	4241	4239.40
2	4174	14	7.64	4160	4166.36
3	4356	19	13.80	4337	4342.20
4	4341	23	22.25	4318	4318.75
5	4380	30	30.17	4350	4349.83
6	4452	41	39.15	4411	4412.85
7	4371	45	49.26	4326	4321.74
8	4359	56	67.68	4303	4291.32
9	4375	97	113.17	4278	4261.83
10	4518	464	444.31	4054	4073.69

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
13.6999	8	0.0899

Annexe 10 : Résultat du modèle Fredette et al. (2008) Appliqué au jeu de données 40% de DM avant qu'il soit imputé

Model Information	
Data Set	WORK.FREDETTE_AL
Response Variable	Major_or_Fatal
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	43570
Number of Observations Used	22765

Response Profile		
Ordered Value	**Major_or_Fatal**	**Total Frequency**
1	1	418
2	0	22347

Probability modeled is Major_or_Fatal='1'.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	4172.186	3307.084
SC	4180.219	3572.173
-2 Log L	4170.186	3241.084

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	929.1017	32	<.0001
Score	1745.4695	32	<.0001
Wald	870.6618	32	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Mass_Ratio	3	23.2464	<.0001
Drivers_Age	3	30.8301	<.0001
Drivers_sex	1	0.1247	0.7240
Authorized_Speed	4	145.7123	<.0001
Collision	8	375.8433	<.0001
Safety_Belt	1	154.1455	<.0001
Driver_Vehicle_Type	6	4.2071	0.6487
Driver2_Vehicle_Type	6	8.1515	0.2272

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-4.4802	0.1632	754.0242	<.0001
Mass_Ratio	**(1.20;2.00]**	1	-0.3335	0.1537	4.7091	0.0300
Mass_Ratio	**(2.00;+)**	1	-0.5931	0.4701	1.5922	0.2070
Mass_Ratio	**(Less than 0.80]**	1	0.5166	0.1360	14.4313	0.0001
Drivers_Age	**24 or less**	1	-0.3538	0.1492	5.6210	0.0177
Drivers_Age	**45-64**	1	0.3035	0.1289	5.5459	0.0185
Drivers_Age	**65 and +**	1	0.5872	0.1698	11.9588	0.0005
Drivers_sex	**Male**	1	0.0404	0.1143	0.1247	0.7240
Authorized_Speed	**100 kph**	1	0.8872	0.3295	7.2493	0.0071
Authorized_Speed	**70-80 kph**	1	1.2235	0.1252	95.5572	<.0001
Authorized_Speed	**90 kph**	1	1.5458	0.1786	74.8671	<.0001
Authorized_Speed	**Less than 5**	1	-0.9391	0.4226	4.9392	0.0263
Collision	**Head-On**	1	1.6976	0.1428	141.2489	<.0001
Collision	**Leftturnconflict(diff.directio**	1	-0.0869	0.1861	0.2179	0.6407
Collision	**Leftturnconflict(samedirection**	1	-0.0682	0.4504	0.0230	0.8796
Collision	**Other**	1	-0.2927	0.1953	2.2473	0.1338
Collision	**Rear-end**	1	-1.7049	0.2088	66.6424	<.0001
Collision	**Rightturnconflict(diff.directi**	1	-0.3097	0.5960	0.2701	0.6033
Collision	**Rightturnconflict(samedirectio**	1	-0.3957	0.4706	0.7071	0.4004
Collision	**Side-swipe(samedirection)**	1	-0.0970	0.2555	0.1441	0.7043
Safety_Belt	**No Safety belt**	1	2.1320	0.1717	154.1455	<.0001
Driver_Vehicle_Type	**Bus**	1	-1.2440	1.1218	1.2299	0.2674
Driver_Vehicle_Type	**Heavy**	1	-0.9191	1.2881	0.5092	0.4755
Driver_Vehicle_Type	**Miniv**	1	-0.3482	0.2328	2.2359	0.1348

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Driver_Vehicle_Type	**Other**	1	-0.1476	0.1840	0.6436	0.4224
Driver_Vehicle_Type	**Picku**	1	-0.1807	0.2305	0.6150	0.4329
Driver_Vehicle_Type	**SUV**	1	-10.5016	283.0	0.0014	0.9704
Driver2_Vehicle_Type	**Bus**	1	1.9324	1.2162	2.5247	0.1121
Driver2_Vehicle_Type	**Heavy**	1	1.2015	1.1542	1.0836	0.2979
Driver2_Vehicle_Type	**Miniv**	1	-0.0884	0.2016	0.1925	0.6609
Driver2_Vehicle_Type	**Other**	1	0.1069	0.1512	0.5003	0.4794
Driver2_Vehicle_Type	**Picku**	1	0.4038	0.2056	3.8555	0.0496
Driver2_Vehicle_Type	**SUV**	1	0.5332	1.0274	0.2694	0.6038

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Mass_Ratio (1.20;2.00] vs (0.80;1.20]	0.716	0.530	0.968
Mass_Ratio (2.00;+) vs (0.80;1.20]	0.553	0.220	1.388
Mass_Ratio (Less than 0.80] vs (0.80;1.20]	1.676	1.284	2.188
Drivers_Age 24 or less vs 25-44	0.702	0.524	0.941
Drivers_Age 45-64 vs 25-44	1.355	1.052	1.744
Drivers_Age 65 and + vs 25-44	1.799	1.290	2.509
Drivers_sex Male vs Female	1.041	0.832	1.303
Authorized_Speed 100 kph vs 50-60 kph	2.428	1.273	4.633
Authorized_Speed 70-80 kph vs 50-60 kph	3.399	2.660	4.344
Authorized_Speed 90 kph vs 50-60 kph	4.692	3.306	6.659
Authorized_Speed Less than 5 vs 50-60 kph	0.391	0.171	0.895

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Collision Head-On vs Right-angle(sideundetermined)	5.461	4.128	7.225
Collision Leftturnconflict(diff.directio vs Right-angle(sideundetermined)	0.917	0.637	1.320
Collision Leftturnconflict(samedirection vs Right-angle(sideundetermined)	0.934	0.386	2.258
Collision Other vs Right-angle(sideundetermined)	0.746	0.509	1.094
Collision Rear-end vs Right-angle(sideundetermined)	0.182	0.121	0.274
Collision Rightturnconflict(diff.directi vs Right-angle(sideundetermined)	0.734	0.228	2.359
Collision Rightturnconflict(samedirectio vs Right-angle(sideundetermined)	0.673	0.268	1.693
Collision Side-swipe(samedirection) vs Right-angle(sideundetermined)	0.908	0.550	1.497
Safety_Belt No Safety belt vs Safety belt	8.432	6.022	11.806
Driver_Vehicle_Type Bus vs Car	0.288	0.032	2.598
Driver_Vehicle_Type Heavy vs Car	0.399	0.032	4.980
Driver_Vehicle_Type Miniv vs Car	0.706	0.447	1.114
Driver_Vehicle_Type Other vs Car	0.863	0.602	1.237
Driver_Vehicle_Type Picku vs Car	0.835	0.531	1.311
Driver_Vehicle_Type SUV vs Car	<0.001	<0.001	>999.999
Driver2_Vehicle_Type Bus vs Car	6.906	0.637	74.893
Driver2_Vehicle_Type Heavy vs Car	3.325	0.346	31.931
Driver2_Vehicle_Type Miniv vs Car	0.915	0.617	1.359
Driver2_Vehicle_Type Other vs Car	1.113	0.827	1.497
Driver2_Vehicle_Type Picku vs Car	1.498	1.001	2.241
Driver2_Vehicle_Type SUV vs Car	1.704	0.228	12.767

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.8	**Somers' D**	0.678
Percent Discordant	16.0	**Gamma**	0.679
Percent Tied	0.1	**Tau-a**	0.024
Pairs	9341046	**c**	0.839