Université de Montréal

Comparaison de quatre méthodes pour le traitement des données manquantes au sein d'un modèle multiniveau paramétrique visant l'estimation de l'effet d'une intervention

Par
Stéphane Paquin

Département de sociologie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et sciences
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Sociologie

Mars, 2010

Université de Montréal
Faculté des arts et sciences

Ce mémoire intitulé :

Comparaison de quatre méthodes pour le traitement des données manquantes au sein d'un modèle multiniveau paramétrique visant l'estimation de l'effet d'un programme de prévention

présenté par :
Stéphane Paquin

a été évalué par un jury composé des personnes suivantes :

Claire Durand
Président-rapporteur
Éric Lacourse
Directeur de recherche
Frank Vitaro
Membre du jury

Résumé

Les données manquantes sont fréquentes dans les enquêtes et peuvent entraîner d'importantes erreurs d'estimation de paramètres. Ce mémoire méthodologique en sociologie porte sur l'influence des données manquantes sur l'estimation de l'effet d'un programme de prévention. Les deux premières sections exposent les possibilités de biais engendrées par les données manquantes et présentent les approches théoriques permettant de les décrire. La troisième section porte sur les méthodes de traitement des données manquantes. Les méthodes classiques sont décrites ainsi que trois méthodes récentes. La quatrième section contient une présentation de l'Enquête longitudinale et expérimentale de Montréal (ELEM) et une description des données utilisées. La cinquième expose les analyses effectuées, elle contient : la méthode d'analyse de l'effet d'une intervention à partir de données longitudinales, une description approfondie des données manquantes de l'ELEM ainsi qu'un diagnostic des schémas et du mécanisme. La sixième section contient les résultats de l'estimation de l'effet du programme selon différents postulats concernant le mécanisme des données manquantes et selon quatre méthodes : l'analyse des cas complets, le maximum de vraisemblance, la pondération et l'imputation multiple. Ils indiquent (I) que le postulat sur le type de mécanisme *MAR* des données manquantes semble influencer l'estimation de l'effet du programme et que (II) les estimations obtenues par différentes méthodes d'estimation mènent à des conclusions similaires sur l'effet de l'intervention.

Mots-clés : Données manquantes, imputation multiple, maximum de vraisemblance, pondération, mécanisme de données manquantes, schéma de données manquantes, multiniveau, effet d'une intervention, données longitudinales.

Abstract

Missing data are common in empirical research and can lead to significant errors in parameters' estimation. This dissertation in the field of methodological sociology addresses the influence of missing data on the estimation of the impact of a prevention program. The first two sections outline the potential bias caused by missing data and present the theoretical background to describe them. The third section focuses on methods for handling missing data, conventional methods are exposed as well as three recent ones. The fourth section contains a description of the Montreal Longitudinal Experimental Study (MLES) and of the data used. The fifth section presents the analysis performed, it contains: the method for analysing the effect of an intervention from longitudinal data, a detailed description of the missing data of MLES and a diagnosis of patterns and mechanisms. The sixth section contains the results of estimating the effect of the program under different assumptions about the mechanism of missing data and by four methods: complete case analysis, maximum likelihood, weighting and multiple imputation. They indicate (I) that the assumption on the type of *MAR* mechanism seems to affect the estimate of the program's impact and, (II) that the estimates obtained using different estimation methods leads to similar conclusions about the intervention's effect.

Keywords : Missing data, multiple imputation, maximum likelihood, weighting, missing data mechanism, pattern of missing data, multilevel, effect of an intervention, longitudinal data.

Table des matières

Liste des tableaux

Liste des figures

À tous ceux qui m'entourent,
parce que sans eux, je ne serais
qu'un parmi d'autres

Remerciements

Je tiens à remercier toutes les personnes qui m'ont appuyé au cours de ce parcours. Mon père, pour qui la science est une religion. Ma mère, pour qui la famille est une institution. Mon frère, ainsi que ta belle-famille, pour qui la fratrie est aussi importante que pour moi. Je tiens à remercier tout spécialement mon directeur Éric Lacourse, pour son appui constant, ses idées et son temps. Je n'aurais pas été aussi fier de ce mémoire, n'eusse été de ces innombrables corrections et réorganisations. Produire des résultats de recherche est une chose, mais communiquer et, dans une certaine mesure, vulgariser les résultats en est une autre. Je tiens également à remercier tous mes amis pour leur présence constante dans ma vie, votre appui m'est indispensable.

À tous, sans votre exemple de respect, compréhension et persévérance, la dernière année aurait été beaucoup moins agréable.

Avant-propos

Toute collecte d'information à grande échelle est, nécessairement, confrontée au problème des données manquantes. Un exemple peut être tiré de la loi des grands nombres, un des théorèmes fondateurs de la statistique. Elle stipule qu'un échantillon tiré aléatoirement et d'assez grande taille permet d'obtenir une estimation sans biais de la moyenne d'une valeur au sein de la population d'où provient l'échantillon. Par exemple, pour obtenir une marge d'erreur d'environ 3 %, 19 fois sur 20, un échantillon de 1000 individus est généralement utilisé pour représenter la population canadienne. Ces 1000 individus représentent à peine 0,003 % de l'ensemble de la population et permettent pourtant d'obtenir une estimation de moyenne sans biais. Est-ce que les réponses des 99,997 % de la population sont des données manquantes? Non, puisque leurs réponses n'étaient pas visées par la collecte. En pratique, les sondeurs tirent généralement un échantillon beaucoup plus grand que celui qu'ils désirent obtenir à la fin de l'enquête. En effet, il est très difficile d'obtenir une réponse de la part de l'ensemble des 1000 individus sélectionnés. Le taux de réponse est un indice indiquant la proportion des individus sélectionnés pour constituer l'échantillon et qui ont effectivement répondu à quelques questions. Dans le même exemple mentionné, posant un taux de réponse de 50 %, le sondeur aurait contacté 2000 individus pour réussir à récolter l'information pour 1000 d'entre eux. Les 1000 individus n'ayant pas répondu représentent des données manquantes. Le principe de la loi des grands nombres suppose que l'ensemble des éléments de l'échantillon aléatoirement tiré contient l'information recherchée. C'est dans ce cas qu'un échantillon est dit représentatif. Si les 1000 individus dont on a obtenu les réponses constituent un sous-échantillon aléatoire des 2000 initiaux, ce postulat de la loi des grands nombres est toujours respecté.

Chapitre 1 : Introduction

D'un point de vue institutionnel, l'évaluation de l'effet d'une intervention est utile pour guider le choix d'une intervention plutôt qu'une autre. D'un point de vue technique, une telle évaluation est sujette à de nombreux obstacles : l'identification d'indicateurs, la mesure de ces indicateurs, l'analyse des données récoltées et la comparaison des effets de différentes interventions. Il s'agit d'appliquer la méthode scientifique à l'étude des interventions sociales. Les techniques d'attribution de la causalité ont une longue histoire qui embrasse sensiblement toutes les sciences. L'attribution de la causalité constitue un des plus grands défis de la recherche scientifique. Ce défi est d'autant plus grand dans le champ des sciences humaines où il n'existe pas de théorie unique englobant l'ensemble des phénomènes humains. Un outil méthodologique comporte toutefois de nombreuses qualités désirables au regard des critères de reproductibilité et de systématisation, la statistique.

Bien qu'elle comporte de nombreux avantages, la statistique a également ses défauts, notamment la complexité de son application. En effet, pour obtenir l'estimation de paramètres sans biais, de nombreux postulats doivent être respectés. Ces postulats sont propres à chacune des méthodes d'estimation. L'attribution de la causalité par des méthodes statistiques est dépendante de plusieurs facteurs, notamment des caractéristiques du plan d'échantillonnage et du type d'analyse menée à partir des données récoltées.

Le plan d'échantillonnage d'une enquête a des implications quant aux hypothèses pouvant être testées, chaque type d'échantillon permet de poser des hypothèses différentes et nécessitent des méthodes d'analyse particulières. Deux critères sont utilisés pour qualifier les propriétés d'une enquête, la validité interne et la validité externe.

La validité interne renvoie à une enquête qui, par son plan et sa structure, permet de conclure qu'une variable indépendante (VI) est bien la cause d'une variable dépendante (VD). Si la relation observée entre la VI et la VD laisse place à de multiples autres interprétations que celle présentée dans l'interprétation, une enquête a une faible validité interne. Par exemple, une enquête contenant de possibles variables médiatrices possède généralement une plus grande validité interne qu'une enquête qui n'a pas testé ces possibles relations puisque la première permet de réduire le nombre d'explications alternatives

possibles (Baron & Kenny, 1986). La validité interne comprend la validité et la fiabilité de la relation entre A et B. La validité de la relation renvoie à la vraie relation entre A et B, c'est-à-dire qu'A cause vraiment B. La fiabilité de la relation désigne le fait que la relation peut être reproduite (McKnight, 2007).

La validité externe, quant à elle, renvoie à la généralisation des résultats. Une enquête avec une validité externe élevée permet de généraliser les résultats à une population. La validité externe est donc liée à la représentativité d'un échantillon, mais également à la validité interne. En effet, un échantillon peut être très représentatif d'une population, mais advenant une faible validité interne, il est possible que la généralisation soit impossible puisqu'aucune relation ne peut être clairement démontrée, donc généralisée.

Dans un échantillon expérimental, l'objectif n'est pas l'inférence à une population, mais l'évaluation de l'effet d'un traitement sur une variable dépendante. Le postulat principal permettant d'évaluer cet effet est l'équivalence entre les groupes contrôle et expérimental. Cette équivalence est assurée en assignant aléatoirement les individus à l'un ou l'autre groupe. Le présent mémoire s'attarde uniquement à la validité interne, l'échantillon utilisé n'étant pas construit pour être représentatif.

1.1. Les sources de biais considérés

Le biais est la différence entre la valeur d'un paramètre populationnel et sa valeur au sein d'un échantillon. La notion de biais fait référence à la capacité d'estimation de la valeur d'un paramètre au sein d'une population à partir d'un échantillon de celle-ci. L'estimation d'un paramètre est dite non biaisée si la valeur prédite par l'estimation est égale à la valeur réelle. L'analyse de l'effet d'une intervention à l'aide de courbes de croissance n'utilise pas un plan d'échantillonnage aléatoire, elle reste tout de même sujette à différents biais. Deux principales sources de biais peuvent influencer les estimations : le biais dû à l'assignation aléatoire et le biais dû aux données manquantes (McKnight, 2007).

L'attention de ce mémoire est portée principalement sur le biais dû aux données manquantes.

L'ensemble de la littérature consultée est unanime pour affirmer qu'une analyse portant uniquement sur les données observées, menée et interprétée comme s'il n'y avait pas de données manquantes, comporte de fortes possibilités de biais. De nombreuses études faites à partir de simulation de données, où les données manquantes sont également prédéterminées, soulignent que dans la majorité des cas, l'utilisation des cas complets entraîne une estimation biaisée de la valeur des paramètres (Allison, 2001; Enders, 2001; Schafer & Graham, 2002). Lors de l'analyse de données réelles, la vraie valeur du paramètre à estimer n'est pas connue, il est alors très difficile de déterminer si les estimations sont biaisées ou de déterminer l'ampleur de ce biais. Puisque chaque situation est différente, les experts de l'analyse des données manquantes s'entendent sur la nécessité d'appliquer une approche systématique et transparente plutôt que de chercher une règle d'or qui pourrait s'appliquer à toutes les situations. Le succès de l'opération de traitement des données manquantes est dépendant (A) des raisons possibles expliquant les données manquantes et (B) de la fiabilité des conclusions obtenues selon les différentes explications évoquées (Carpenter, Kenward, 2008).

Le présent mémoire est composé de cinq sections : la théorie sur les données manquantes, les méthodes de traitement des données manquantes, la description des données utilisées, les analyses et les résultats.

Chapitre 2 : Théorie sur les données manquantes

Une donnée manquante peut être définie comme une donnée qui était visée par le processus de collecte, mais qui n'a pu être obtenue. Différents éléments peuvent être à l'origine de données manquantes, ils peuvent être associés aux participants, au plan de l'étude ou à une interaction entre les deux. Une donnée peut être manquante au moment du recrutement, de la collecte ou du traitement des données. La donnée manquante peut être un participant, un item d'un questionnaire ou une vague de collecte (McKnight, 2007). Toutes ces caractéristiques permettent de décrire les données manquantes pour chacune des analyses menées.

Les données manquantes qui nécessitent d'être traitées sont celles qui sont incluses dans le modèle statistique utilisé. Le principe est d'assurer que les relations entre les variables ne soient pas affectées par les données manquantes et que les tests d'hypothèses reflètent bien l'incertitude inhérente aux données manquantes.

Souvent, le producteur des données et l'utilisateur final ne sont pas les mêmes personnes. Deux approches sont décrites dans la littérature : les données manquantes sont traitées avant que le chercheur produise des analyses ou le chercheur traite les données manquantes comme une étape lors du processus d'analyse. Diverses options sont offertes selon le choix de la personne qui traitera les données manquantes. Chacune d'elles comporte ses avantages et inconvénients.

La nature du traitement des données manquantes suppose que cette étape soit effectuée par la personne qui produit l'analyse finale des données (Carpenter et Kenward, 2008). La présente étude suppose que la même personne traite les données manquantes et produit l'analyse. Toutefois, dans certaines conditions, les méthodes présentées pourraient être appliquées à un fichier de données avant que ce dernier ne soit fourni à l'analyste.

Les données manquantes peuvent avoir un impact sur la validité, la fiabilité et la généralisation des estimations effectuées. Le type et la quantité de données manquantes peuvent également avoir un impact différent sur chacun de ces aspects.

La proportion de données manquantes n'est pas suffisante pour déterminer si leur présence est problématique ou non. C'est plutôt la question de recherche, l'information contenue dans les données observées et les raisons à l'origine des données manquantes qui déterminent leur impact (Carpenter & Kenward, 2008). Les raisons à l'origine des données manquantes constituent le volet sur lequel le chercheur a le moins de contrôle. Leur présence apporte une incertitude supplémentaire dans l'analyse de données, incertitude bien distincte de la notion d'erreur d'échantillonnage.

2.1. Conséquences de la présence de données manquantes

La présence de données manquantes peut avoir un impact à différents niveaux, au niveau de la validité interne, de la validité des associations entre variables, au niveau de la validité d'un construit latent (dans le cas d'une échelle de type Likert où plusieurs items sont utilisés pour définir un construit) et au niveau de la généralisation d'une association entre deux variables.

Les données manquantes peuvent menacer la validité interne à différents moments du processus de recherche : au moment de la sélection de l'échantillon, de l'assignation aléatoire au groupe expérimental ou au groupe contrôle, de la collecte (non-réponse complète ou partielle, attrition) et de l'analyse statistique des données.

Au moment de la sélection de l'échantillon à analyser, le choix d'utiliser seulement les participants avec des données complètes par rapport aux participants avec données manquantes peut avoir un impact sur la validité interne. L'assignation au groupe expérimental ou au groupe contrôle doit être faite de manière aléatoire pour assurer que ces groupes soient statistiquement équivalents. Si les groupes se différencient par rapport au type ou à la quantité de données manquantes, le principe d'équivalence statistique peut être affecté. Les données manquantes sur une ou plusieurs covariables peuvent également empêcher de valider l'assignation aléatoire. L'attrition différentielle entre les groupes

expérimental et contrôle peut également affecter la validité interne pour les mêmes raisons (Raudenbush, 2001).

Les données manquantes ont généralement un impact sur la taille de l'échantillon à partir duquel les analyses statistiques sont effectuées, elles peuvent ainsi affecter la puissance statistique des analyses. Une faible puissance statistique augmente la probabilité de commettre des erreurs de type II. La présence de données manquantes peut également affecter certains postulats des méthodes d'analyses : la distribution normale des variables ou des résidus. Dans le cas d'analyses multivariées, la quantité de données manquantes peut croître rapidement : par exemple, dans le cas où quatre variables incluses comportent chacune 5 % de données manquantes et que ce faible pourcentage ne provient pas des mêmes participants, l'analyse peut impliquer jusqu'à 20 % de données manquantes.

Les erreurs de type 1 et 2, ou erreurs alpha et bêta, font référence à la décision prise concernant l'hypothèse nulle dans l'échantillon par rapport à la réalité de cette hypothèse nulle. Lorsqu'un chercheur rejette l'hypothèse nulle alors qu'elle est vraie en réalité, il commet une erreur de type 1, ou erreur alpha. À l'inverse, lorsque le chercheur accepte l'hypothèse nulle alors qu'elle est fausse, il commet une erreur de type 2, ou erreur bêta (Fox, 1999).

2.2. Description des données manquantes

Toute analyse de données commence généralement par une analyse descriptive des variables. Le chercheur décrit les variables, leurs distributions, leurs significations. Il en va de même pour les données qui devaient être récoltées et qui ne l'ont pas été. La description des données manquantes constitue un point de départ pour évaluer l'impact que ces dernières peuvent avoir sur les analyses prévues. Une description des données manquantes permettra également au lecteur de juger de la validité des résultats, comme la déclaration du nombre de cas inclus dans les analyses est une information importante pour que le lecteur puisse juger d'un certain niveau de signification. La déclaration des problèmes de données

manquantes contenus dans les données utilisées permet également au lecteur de comparer plus adéquatement les résultats publiés à partir de données différentes (Carpenter & Kenward, 2008).

2.2.1. Quantité de données manquantes

La quantité de données manquantes peut être définie de différentes manières selon ce à quoi on réfère. On peut se référer aux cas (lignes), aux variables (colonnes) ou à l'échantillon complet. Selon qu'on réfère à l'un ou à l'autre, différentes statistiques peuvent être calculées pour informer sur la quantité de données manquantes.

Les cas complets et les cas complets par paires sont lorsqu'on réfère aux cas. On calcule alors la proportion d'individus avec des données complètes. Le calcul des cas complets par paires est similaire aux cas complets, mais il est utilisable uniquement dans le cadre d'une analyse bi- ou multivariée. Les cas complets sur toutes les variables utilisées dans l'analyse sont alors additionnés. Si on réfère aux variables, on calcule la proportion de variables avec des données complètes. En se référant à l'ensemble de l'échantillon, on parle de matrice clairsemée. Il s'agit de calculer la proportion d'observations manquantes sur l'ensemble des observations effectuées.

Posons l'exemple d'un échantillon de dix individus pour lesquels cinq variables furent mesurées. Six d'entre eux ont des valeurs manquantes. De plus, elles sont manquantes sur la même variable. Le tableau I à la page suivants décrit cet exemple : les cases noires représentent les données observées et les blanches, les manquantes. Les trois équations qui suivent montrent le calcul de ces trois statistiques à partir des données de l'exemple précédent.

Tableau I : Illustration d'une matrice de données manquantes

ID	Y	X1	X2	X3	X4
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Équation 1 : Proportion d'individus avec données complètes

$$\frac{Nombre\ d'individus\ avec\ données\ complètes}{Nombre\ total\ d'individus} = \frac{4}{10} = 40\ \%$$

Équation 2 : Proportion de variables avec données complètes

$$\frac{Nombre\ de\ variables\ avec\ données\ complètes}{Nombre\ total\ de\ variables} = \frac{4}{5} = 80\ \%$$

Équation 3 : Proportion d'observations manquantes

$$\frac{Nombre\ d'observations\ manquantes}{Nombre\ total\ d'observations\ effectuées} = \frac{6}{50} = 12\ \%$$

Il est également possible de calculer des proportions basées sur ces différentes statistiques. Ces proportions fournissent une estimation de la densité des données manquantes au niveau des répondants ou au niveau des variables. Ces proportions indiquent la proportion moyenne de variables manquantes pour les répondants ayant des données manquantes; ou la proportion moyenne de répondants manquants pour les variables contenant des réponses manquantes. La première est le rapport de la proportion obtenue par la matrice clairsemée à la proportion d'observations avec données manquantes (le même calcul peut être effectué en rapport à la proportion de variables avec données

manquantes). Avec les données de l'exemple précédent, la proportion moyenne de variables manquantes pour les répondants ayant des données manquantes serait de :

Équation 4 : Proportion d'observations manquantes

$$\frac{Proportion\ d'observations\ manquantes}{Proportion\ d'individus\ avec\ données\ manquantes} = \frac{12}{60} = 20\ \%$$

2.3. Schéma et mécanisme

Une première catégorisation des données manquantes a été proposée par Little et Rubin en 1987 (Little & Rubin, 1987). Les auteurs distinguent le schéma et le mécanisme relatif à l'absence d'une donnée. Un schéma de données manquantes désigne une séquence de valeurs observées et manquantes dans une matrice de données. Quant à lui, le mécanisme renvoie à la relation entre les valeurs contenues dans la matrice de données et le fait qu'une donnée soit observée ou non. Tant le schéma que le mécanisme de données manquantes peuvent avoir une influence sur l'estimation des paramètres et sur le choix d'une méthode de traitement des données manquantes. Une description appropriée des données manquantes devrait contenir la quantité, les dimensions, le schéma et le mécanisme.

L'analyse des schémas des données manquantes sert à identifier si certaines caractéristiques sont associées à un schéma particulier. Un schéma de données manquantes est constitué de toutes les variables et de tous les cas qui partagent une même séquence de valeurs observées et manquantes. Par exemple, dans un plan longitudinal, le fait de quitter l'enquête produit une séquence de valeurs observées puis de valeurs manquantes. Cette séquence particulière décrit le phénomène de l'attrition. Si le fait de quitter l'enquête est associé à certaines autres variables mesurées, il pourrait être déterminé, par exemple, que les gens quittant l'enquête avaient systématiquement un niveau d'éducation différent de ceux qui sont restés. Dans le cas d'un plan synchronique, il pourrait être déterminé que la

combinaison de données manquantes sur des variables comme la croyance religieuse, le revenu et l'orientation politique est systématiquement reliée à l'âge du répondant.

L'analyse du mécanisme des données manquantes renvoie quant à lui à la description des causes relatives à l'absence d'une donnée pour une variable particulière. L'identification de variables associées à ce mécanisme permet de contrôler pour les données manquantes lors d'une analyse incluant cette variable. Le schéma est constitué de plusieurs variables alors que le mécanisme renvoie à l'explication d'une seule variable.

2.3.1. Schéma

Chaque matrice de données contient généralement plusieurs schémas de données manquantes. Posons une matrice de données, nommée *matrice R*, contenant le même nombre de lignes et de colonnes que la matrice de données originale que nous nommons *Y*. Pour identifier un schéma, il s'agit d'attribuer la valeur de 1 aux cellules de la matrice *R* si la cellule correspondante de *Y* est manquante et 0 si la cellule correspondante de *Y* est observée. La matrice *R* permet de définir les différents schémas de données manquantes. Les colonnes ou rangées de cette matrice peuvent être déplacées afin de structurer les schémas. Un schéma est une séquence individuelle, qui s'observe sur une ligne dans la matrice *R*, de valeurs observées et manquantes. La figure 1 à la page suivante contient des exemples de certains schémas. Certaines méthodes de traitement des données manquantes s'appliquent à une structure de schémas particulière alors que d'autres peuvent être appliquées à tous les types de schémas. La taille d'un schéma renvoie au nombre d'observations qui partagent la même séquence de 0 et de 1.

Figure 1 : Exemples de schémas de données manquantes

Non-réponse univariée

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

Deux schémas, multivariés

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	0	0
0	0	0	0	0
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1

Schéma monotone

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	1	1
0	0	1	1	1
0	1	1	1	1

Schémas lors d'une fusion de matrices de données

Y_1	Y_2	Y_3
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Schéma d'une analyse factorielle

Y_1	X
0	1
0	1
0	1
0	1
0	1

Schémas désordonnés

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	1	0
0	0	1	0	1
0	1	0	0	0
0	0	0	1	1
0	1	1	0	0

Une méthode permettant de décrire la quantité de schémas présents ainsi que l'analyse de ces schémas est ici présentée. L'analyse des schémas peut permettre de tester si un schéma est associé à un sous-groupe particulier ou à quelques autres variables présentes dans la matrice de données. Cette méthode consiste à ajouter une ligne et une colonne à la matrice R.

La nouvelle ligne contient une valeur différente pour chaque colonne, ou variable, présente. Cette valeur remplacera les 1 dans chacune des colonnes. McKnight et al. (2007) suggèrent d'utiliser un nombre premier (2, par exemple) et de l'élever à une puissance égale au rang de chacune des colonnes. Dans le cas où 2 est choisi comme nombre, les 1 de la première colonne seront remplacés par des 2, ceux de la seconde colonne par des 4 (2^2), ceux de la cinquième colonne par des 32 (2^5), etc.

La nouvelle colonne, que nous identifions MISPAT, contient la somme de chaque ligne. De cette manière, la valeur 2 dans MISPAT indique que cette ligne contient une seule donnée manquante et qu'elle est située dans la première colonne. De la même manière, une valeur de 36 indiquerait que la ligne contient deux valeurs manquantes et qu'elles sont situées aux deuxièmes et cinquièmes colonnes (4+32=36). Le fait d'utiliser un nombre élevé à une puissance permet d'obtenir des valeurs éloignées les unes des autres pour chaque colonne, résultant dans le fait que chaque valeur de MISPAT est associée à un schéma unique de données manquantes.

MISPAT peut ensuite être analysée à l'aide de tests statistiques traditionnels (test-T, ANOVA, etc.) afin d'établir les différences initiales entre les divers schémas de données manquantes. Les schémas présentés plus haut tels que monotone, général (désordonné) ou univarié pourraient également être identifiés de cette manière.

Les données longitudinales posent un problème particulier lors de l'identification des schémas des données manquantes. Pour une étude comportant T temps de mesure, la littérature suggère qu'il peut y avoir jusqu'à 2^T schémas de données. La taille de l'échantillon peut également limiter les possibilités d'analyse des schémas. Une stratégie

permettant de limiter le nombre de schémas à analyser avec des données longitudinales est proposée par Yang et consiste à distinguer le schéma monotone de tous les autres schémas. Les schémas qui ne sont pas monotones sont qualifiés d'intermittents. Le schéma intermittent regroupe tous les schémas où des données sont manquantes à un certain temps de mesure et que des données ont été observées à un temps de mesure antérieur et à un temps de mesure postérieur (Yang & Shoptaw, 2005). Le schéma monotone indique que lorsqu'une donnée est manquante à un certain temps de mesure pour un certain individu, les données pour ce même individu sont également manquantes pour tous les temps de mesure suivants.

Certains auteurs (Schafer & Graham, 2002; Yang & Shoptaw, 2005) soutiennent que les schémas désordonnés sont de bons indicateurs que les données sont manquantes de manière aléatoire, D'autres (Graham, 2009; McKnight, 2007) suggèrent de voir le niveau de désordre comme un continuum, le ratio de désordre servant à décrire où se situent les données le long de ce continuum. Le ratio de désordre est le ratio du nombre de schémas par rapport au nombre de lignes dans la matrice de données (le nombre de participants). Deux schémas identifiés au sein d'un échantillon de 10 personnes fournissent un ratio de 0,2 (2/10=0,2). Plus le ratio est près de 1, plus les schémas de données manquantes sont désordonnés (McKnight, 2007).

Pour avoir une idée rapide, il est également possible d'utiliser des techniques graphiques pour chercher les schémas de données manquantes. Une première technique consiste à afficher l'ensemble de la matrice R en ajoutant une couleur en fonction des valeurs (noir=0=valeur observée; blanc=1=valeur manquante). Il est ensuite possible de trier la matrice de différentes manières afin de mettre en évidence un schéma ou un autre. Cette technique permet également d'avoir une idée rapide de la quantité de données manquantes.

2.3.2. Mécanisme

Pour définir les différents mécanismes de données manquantes, nous utiliserons les expressions anglophones qui sont largement répandues.

Trois mécanismes de données manquantes ont été décrits par Little et Rubin : *missing completely at random (MCAR), missing at random (MAR) et missing not at random (MNAR)*. La catégorisation des mécanismes est fondée sur la probabilité qu'une donnée soit manquante. Le mécanisme est dit *MCAR* lorsque la probabilité qu'une donnée soit manquante est indépendante des valeurs de la variable d'intérêt. En d'autres termes, le mécanisme est *MCAR* lorsque les données observées constituent un échantillon représentatif de l'ensemble des données visées par la collecte (Graham, 2009). Dans cette situation, chacun des schémas des données manquantes peut fournir de l'information adéquate pour estimer un paramètre ou tester une hypothèse. L'analyse des cas complets dans cette situation permet donc d'obtenir un estimé non biaisé. Le mécanisme est *MAR* lorsque la probabilité qu'une donnée soit manquante sur notre même variable d'intérêt est dépendante des valeurs observées sur une autre variable à la disposition du chercheur. Le terme « random » dans l'expression réfère au fait que, si l'analyse contrôle pour les variables que l'on croit associées à la probabilité qu'une valeur soit manquante, les données manquantes peuvent être considérées aléatoires ou *MCAR*. En d'autres termes, même l'analyse des cas complets peut générer des estimations non biaisées si le modèle d'analyse contrôle pour les variables associées aux données manquantes.. Graham utilise l'expression *conditionally missing at random* pour faciliter la compréhension de l'expression *MAR*. Il est à noter qu'il ne tient pas à proposer un nouveau terme remplaçant *MAR*, mais uniquement à mieux le faire comprendre (Allison, 2000; Graham, 2009). Enfin, il est *MNAR* quand la probabilité qu'une donnée soit manquante est dépendante des valeurs non observées de la variable d'intérêt.

La distinction entre ces mécanismes est faite en analysant la relation entre la variable où les données sont manquantes et les autres variables que l'on désire inclure dans

l'analyse. Le mécanisme décrit donc les causes potentielles des données manquantes. Graham mentionne que le terme mécanisme des données manquantes, tel qu'utilisé par les statisticiens, renvoie à une notion distincte de ce que les chercheurs en sciences sociales entendent généralement en utilisant le terme de mécanisme. Le terme mécanisme, pour les chercheurs en sciences sociales, renvoie fréquemment à la notion de mécanisme causal, alors que tel qu'utilisé par les statisticiens, il renvoie à la description de la manière dont les données sont manquantes (causes relatives à l'absence des données). L'objectif est de déterminer si les causes relatives à l'absence des données peuvent être conditionnelles à certaines variables observées. C'est pourquoi seules les variables accessibles au chercheur doivent être considérées pour déterminer le mécanisme des données manquantes. En d'autres termes, le mécanisme des données manquantes réfère aux relations entre les variables de l'échantillon et non pas à des causes extérieures et non mesurées (Allison, 2001; Graham & Donaldson, 1993).

La principale conséquence des données manquantes *MCAR* est la perte de puissance statistique. Cette situation engendre néanmoins des estimations non biaisées. Les données manquantes *MAR* engendrent également des estimations non biaisées lorsque les analyses contrôlent pour le mécanisme des données manquantes. La troisième situation, *MNAR*, est la plus problématique parce qu'elle engendre des estimations biaisées (Graham, 2009).

Pour décrire les mécanismes de données manquantes, posons le vecteur y_i représentant une variable avec données manquantes. Ce vecteur est formé de deux composantes, $y_i^{(o)}$ et $y_i^{(m)}$, tel que $y_i = y_i^{(o)} + y_i^{(m)}$. Le premier étant le vecteur contenant les valeurs observées et le second contenant les valeurs manquantes. Le vecteur $y_i^{(m)}$ est purement théorique, il est concrètement impossible de lui attribuer des valeurs puisque ces données ne sont pas disponibles. Enfin, un dernier vecteur (r_i) identifiant l'absence/présence d'une donnée est posé. Le vecteur r_i prend la valeur 0 si la donnée est observée et 1 si elle est manquante.

Les données manquantes sur le vecteur y_i sont dites *MCAR* si la probabilité qu'une donnée soit manquante est indépendante des valeurs de y_i et des covariables. Elles sont *MAR* si la probabilité qu'une donnée soit manquante est conditionnelle à une ou plusieurs covariables. Enfin, elles sont *MNAR* si la probabilité qu'une donnée soit manquante est conditionnelle à une ou plusieurs covariables, aux données observées et aux données manquantes. Le tableau II présente les équations théoriques représentant ces trois mécanismes.

Tableau II : Mécanismes de données manquantes

Mécanisme	**Représentation théorique**
MCAR	$P(r_i \mid x_i, y_i, \psi) = P(r_i \mid \psi)$
MAR	$P(r_i \mid x_i, y_i, \psi) = P(r_i \mid x_i, \psi)$
MNAR	$P(r_i \mid x_i, y_i, \psi) = P(r_i \mid x_i, y_i^{(o)}, y_i^{(m)}, \psi)$

Où r_i = indicateur binaire [0, 1] de la présence/absence d'une observation
x_i = covariable(s)
y_i = variable d'intérêt (dépendante)
ψ = paramètre associé à la probabilité de r_i

Le diagnostic du mécanisme pose certains problèmes. Prenons d'abord la situation *MCAR*. La situation *MCAR* se présente lorsqu'il n'y a aucune relation entre r_i et y_i, donc aucune relation entre r_i et $y_i^{(o)}$ ni entre r_i et $y_i^{(m)}$. En d'autres termes, la présence d'une donnée n'est aucunement liée aux valeurs de y_i, peu importe qu'elles aient été effectivement observées ou non. Il appert clairement qu'une telle conclusion ne pourra jamais être affirmée hors de tout doute puisqu'il est impossible d'évaluer la relation entre r_i et $y_i^{(m)}$. Il est toutefois possible de tester la relation entre r_i et $y_i^{(o)}$. Advenant qu'une relation soit trouvée entre r_i et $y_i^{(o)}$, la situation *MCAR* pourrait être réfutée.

La méthode décrite dans la littérature pour tester si le mécanisme *MCAR* peut être rejeté est d'utiliser des tests-t. Toutes les covariables sont utilisées dans un test de différence des moyennes selon que r_i=0 ou r_i=1. Si des différences significatives de moyennes entre les deux groupes sont trouvées, le mécanisme *MCAR* peut être rejeté (Graham, 2009; Little & Rubin, 2002).

Dans le cas où la matrice de données contient de nombreuses variables, Little et Rubin ont développé un test visant à évaluer si les données manquantes d'une matrice complète de données peuvent être considérées *MCAR*. À notre connaissance, ce test est pour le moment uniquement implanté dans le module *Missing Value Analysis (MVA)* disponible pour SPSS.

Si le mécanisme *MCAR* a été réfuté, il reste à vérifier si le mécanisme est *MAR* ou *MNAR*. Ces deux mécanismes sont très difficiles à distinguer puisque l'indépendance entre r_i et y_i ne peut être testée, la portion $y_i^{(m)}$ du vecteur y_i étant inconnue (Allison, 2001).

Ignorabilité

Il est théoriquement impossible de discriminer entre *MAR* et *MNAR*, donc de déterminer avec certitude le mécanisme des données manquantes (la cause relative à l'absence des données).

Sur ce sujet, Graham (2009) écrit :

> *The major three missingness mechanisms are MCAR, MAR, and MNAR. These three kinds of missingness should not be thought of as mutually exclusive categories of missingness, despite the fact that they are often misperceived as such. In particular, MCAR, pure MAR, and pure MNAR really never exist because the pure form of any of these requires almost universally untenable assumptions. The best way to think of all missing data is as a continuum between MAR and MNAR. Because all missingness is MNAR (i.e., not purely MAR), then whether it is MNAR or not should never be the issue. Rather than focusing on whether the MI/ML [MAR] assumptions are violated, we should answer the question of whether the violation is big enough to matter to any practical extent. (Graham, 2009, p.567)*

Little et Rubin (2002) présentent un concept pour comprendre la distinction entre les mécanismes *MAR* et *MNAR*. Le concept est celui de l'ignorabilité. Le mécanisme *MAR* est dit ignorable alors que *MNAR* est non ignorable. Comme mentionné, aucune procédure ne peut définitivement distinguer les deux mécanismes, certains auteurs utilisent donc uniquement la notion d'ignorabilité, c'est-à-dire établir si le mécanisme est ignorable ou non, plutôt que discriminer entre *MAR* et *MNAR*. Dans le cas où elles sont ignorables, et supposant l'utilisation de stratégies34 d'analyse appropriée, il y a peu de chances que les estimations soient biaisées. Lorsque les données manquantes sont non ignorables, il devient nécessaire de modéliser le mécanisme des données manquantes à même l'analyse utilisée (Little, 1995; Little & Wang, 1996; Little & Rubin, 2002). Il devient toutefois difficile de déterminer si les estimations sont biaisées ou non.

Graham et Donaldson présentent un exemple d'analyse montrant bien la distinction entre *MAR* et *MNAR*. L'exemple fictif concerne l'analyse de l'effet d'un programme de prévention du tabagisme où des individus ont été assignés aléatoirement à un groupe contrôle ou intervention. L'assignation au groupe est représentée par une variable binaire, *programme*. Le plan d'échantillonnage compte deux temps de mesure de la fréquence de consommation de tabac : une mesure pré-test, puis une mesure prise après l'intervention. Supposons ensuite que le second temps de mesure comporte plusieurs données manquantes, et que ces dernières sont associées aux mesures prises au temps un. Si, lors de l'analyse de l'effet du programme à l'aide de méthodes adaptées (imputation multiple ou maximum de vraisemblance), l'analyste inclut à la fois la variable *programme* et la mesure prétest de la fréquence de consommation de tabac, alors la relation entre smoking2 et *programme* est évaluée en contrôlant pour smoking1. L'utilisation de la variable contrôle smoking1 dans l'analyse rend les données manquantes sur smoking2 *MAR* (ignorable), alors que si smoking1 n'avait pas été incluse dans l'analyse, les données manquantes sur smoking2 auraient été *MNAR* (non ignorable) puisque l'analyste n'aurait pas inclus la cause relative à l'absence des données dans l'analyse (Graham, 2009). Pour revenir sur la spécification de la cause relative à l'absence des données, rappelons que l'expression ne renvoie pas à

l'explication causale d'un phénomène, c'est-à-dire que la même enquête pourrait être reprise avec un nouvel échantillon et qu'il serait possible que d'autres variables soient associées à la cause relative à l'absence des données. Théoriquement, deux enquêtes identiques (même questionnaire), menées à partir de deux échantillons aléatoires différents (A et B), pourraient avoir des causes relatives à l'absence des données distinctes. Il pourrait être inapproprié d'inclure dans les analyses de l'échantillon A, les causes relatives à l'absence des données identifiées dans l'échantillon B.

Schafer a également proposé des lignes directrices indépendantes de l'analyse de données pour juger si une situation peut raisonnablement être considérée ignorable.

Situations pouvant être présumées ignorables :

- Sélection aléatoire de non-répondants qui seront contactés pour un suivi intensif ;
- Essais cliniques avec assignation aléatoire.

Situations ne pouvant pas être présumées ignorables :

- Les non-répondants ne font pas l'objet d'un suivi;
- Les plans expérimentaux où les données sont manquantes de manière non intentionnelle;
- Dans les études observationnelles où le chercheur ne peut contrôler le fait que les données visées ne soient pas collectées. Un exemple serait une étude sur la délinquance à l'école à l'adolescence où les plus délinquants sont perdus parce qu'ils quittent l'école, sont suspendus ou décrochent. (Schafer, 1997)

L'identification du mécanisme des données manquantes est une tâche fortement influencée par la nature des données. Dans les études à plan longitudinal-expérimental, le schéma intermittent peut généralement être considéré comme ignorable si la proportion de données manquantes est faible. Mais le mécanisme des données manquantes dû à l'attrition peut rarement être considéré comme ignorable. Puisqu'il est impossible de distinguer en pratique les mécanismes *MAR* et *MNAR*, différents auteurs utilisent une approche où le mécanisme *MNAR* est postulé et la cause relative à l'absence de données est modélisée au sein de l'analyse. Différentes versions de la cause relative à l'absence de données sont

appliquées et une analyse de sensibilité est ensuite effectuée en considérant les différentes modélisations du mécanisme des données manquantes.

Chapitre 3 : Méthodes de traitement des données manquantes

Les méthodes de traitement des données manquantes se distinguent selon deux approches, les méthodes supprimant les données manquantes et les méthodes utilisant toute l'information disponible. Dans la première catégorie, l'on retrouve les techniques connues sous l'appellation analyse des cas complets (listwise deletion) et analyse des cas complets par paires (pairwise deletion). Certains auteurs suggèrent également que le choix de ne pas inclure dans les analyses les variables qui comportent des données manquantes constitue une forme de suppression des données. Parmi les méthodes utilisant toute l'information disponible, notons l'ajustement par variable binaire, toutes les variantes de l'imputation, le maximum de vraisemblance, l'algorithme EM, le Markov Chain Monte Carlo (MCMC), la pondération ainsi que l'imputation multiple. Il est à noter que cette liste de méthodes n'est pas exhaustive. Les méthodes utilisant toute l'information disponible sont généralement préférées. Ce chapitre couvre d'abord certaines méthodes classiques et se termine par les méthodes récentes qui seront appliquées par la suite.

3.1. Méthodes classiques

Parmi les méthodes classiques, figurent l'analyse des cas complets, l'ajustement par variable binaire et l'imputation simple. Les deux premières sont rapidement décrites ici et la section 3.1.1 est consacrée à l'imputation simple. La méthode des cas complets (listwise ou pairwise) consiste à utiliser dans les analyses uniquement les cas qui ne comportent aucune donnée manquante. Cette méthode entraîne généralement une diminution de la taille de l'échantillon, entraînant une diminution de la puissance statistique et possiblement un biais. L'ajustement par variable binaire peut uniquement être utilisé dans des modèles de régression. Elle consiste à remplacer une variable contenant des données manquantes par deux variables créées de la manière suivante :

1. Une première variable, la variable binaire, est créée en codant 1 pour indiquer une donnée manquante et 0 pour indiquer une donnée observée;

2. La seconde variable est une copie de la variable originale où les données manquantes ont été remplacées par une constante (par exemple, la moyenne des valeurs observées).

3.1.1. L'imputation simple

L'imputation est une méthode où les données manquantes sont remplacées par des valeurs probables. L'identification des valeurs probables à utiliser est faite selon une panoplie de méthodes. Les méthodes d'imputation sont soit l'imputation simple ou l'imputation multiple. L'imputation simple remplace les données manquantes par une seule valeur. Ces méthodes entraînent généralement une diminution de la variance des variables et peuvent altérer la force des associations identifiées.

Les méthodes d'imputation simple sont généralement celles générant les moins bons résultats. Bien qu'un nombre important de méthodes d'imputation simple aient été développées, l'imputation multiple procure des résultats généralement plus adéquats. Les méthodes d'imputation simple peuvent être regroupées en trois catégories, les méthodes d'imputation par une constante, les méthodes d'imputation par une valeur aléatoire et les méthodes d'imputation par une valeur non aléatoire. Le tableau III de la page suivante résume différentes méthodes d'imputation simple et le type de données auxquelles elles s'appliquent.

Tableau III : Méthodes d'imputation simple

Procédures générales	***Procédure spécifique***	***Type de données***
Constante	Substitution par la moyenne	Normale continue
	Substitution par la moyenne obtenue par maximum de vraisemblance	Normale continue
	Substitution par la médiane	Continue
	Imputation de zéro	Catégorielle ou Continue
Aléatoire		
Basée sur les données	Hot deck	Tous
	Cold deck	Tous
Basée sur un modèle	Bayesienne (MCMC)	Tous
	Maximum de vraisemblance	Normale continue
Non-aléatoire		
Une condition	Moyenne du groupe	Continue avec groupes
	Médiane du groupe	Continue avec groupes
	Dernière observation passée en avant (LOCF)	Longitudinale
	Prochaine observation passée en arrière (NOCB)	Longitudinale
Plusieurs conditions	Moyenne des observations précédentes	Longitudinale
	Moyenne des observations suivantes	Longitudinale
	Moyenne de l'obs. précédente et de l'obs. suivante	Longitudinale
	Régression	Continue multivariée
	Régression avec erreur	Continue multivariée

Source : (McKnight, 2007)

Imputation d'une constante

Ces méthodes remplacent les valeurs manquantes par la moyenne arithmétique (ou obtenue par maximum de vraisemblance) ou par d'autres statistiques de tendance centrale. Ces méthodes diminuent la variance des données et produisent donc généralement des estimations de paramètres de faible qualité.

Imputation d'un nombre aléatoire

Ces méthodes remplacent les données manquantes par une valeur choisie aléatoirement. La valeur peut être choisie au sein de l'échantillon ou dans une autre enquête similaire. Elle peut également être sélectionnée en fonction des caractéristiques de l'observation manquante. Par exemple, si une femme a une donnée manquante, la valeur sélectionnée pour l'imputation serait choisie aléatoirement à partir des données observées des autres femmes.

Imputation d'un nombre non aléatoire

Une condition

L'imputation d'un nombre non aléatoire est également effectuée en fonction d'une ou de plusieurs caractéristiques de l'observation manquante. Contrairement aux méthodes précédentes, celles-ci utilisent la moyenne ou la médiane d'un sous-groupe comportant la ou les mêmes caractéristiques que l'observation manquante. Dans le cas de données longitudinales, deux méthodes sont également utilisées : la dernière observation passée en avant ou la prochaine observation passée en arrière. Il s'agit d'utiliser la valeur de la dernière ou de la prochaine observation non manquante pour l'imputation des temps de mesure manquants. Cette méthode suppose que la trajectoire de changement soit considérée comme fixe où des données sont manquantes.

Plusieurs conditions

Toujours applicable aux données longitudinales, la moyenne de toutes les observations antérieures ou postérieures peut être utilisée comme valeur à imputer.

Une dernière méthode dans le même ordre d'idée est d'utiliser la moyenne de l'observation précédente et de l'observation suivante comme valeur à imputer. Cette méthode tend généralement à conserver la forme de la trajectoire des données observées. Engels et Diehr ont démontré que cette méthode entraîne peu de biais dans l'estimation des paramètres et de leur variance (Engels et Diehr, 2003).

Régression

La méthode de la régression utilise des variables associées à la variable comportant des données manquantes. La régression est effectuée en utilisant les cas complets. Le coefficient obtenu par la régression des covariables sur la variable avec données manquantes est utilisé pour l'imputation. L'imputation par régression est la méthode sur laquelle est basée l'imputation multiple.

3.2. Méthodes récentes

3.2.1. Le maximum de vraisemblance

Le maximum de vraisemblance est une méthode paramétrique, utilisant une distribution sous-jacente, afin d'obtenir l'estimation des paramètres inconnus qualifiant la distribution des données observées. Le principe du maximum de vraisemblance est que l'estimation obtenue est celle qui maximise la probabilité d'observer ce qui a effectivement été observé (Allison, 2001). En d'autres termes, dans un échantillon où il n'y a aucune donnée manquante, c'est la probabilité des données observées qui est modélisée. Elle est définie comme conditionnelle à la fois aux données observées et à un ou plusieurs paramètres non connus.

Posons que l'on inclut dans une analyse, une VD (y) et une VI (x); la probabilité des données observées est conditionnelle à y, x et un paramètre inconnu, pour l'ensemble des données. Lorsque l'on est en présence de données manquantes, la probabilité des données observées est définie comme conditionnelle à deux fonctions de vraisemblance, une pour les données complètement observées et une seconde pour les données partiellement observées (Allison, 2001). La fonction de maximum de vraisemblance peut être résolue par différents algorithmes. Le plus utilisé en présence de données manquantes est l'*expected maximization* (EM).

3.2.2. L'imputation multiple

L'imputation multiple est une technique d'imputation visant à corriger la sous-estimation de la variance qui est caractéristique des méthodes d'imputation simple. C'est une méthode de plus en plus utilisée dans une variété de disciplines puisqu'elle génère des estimations sans biais lorsque l'analyse est *MAR* et que le modèle d'imputation tient compte de la cause des données manquantes. Les méthodes d'imputation multiple ajoutent un facteur de correction à la variance des estimations afin de tenir compte de l'incertitude de la bonne valeur à imputer. Ce facteur de correction est calculé à partir de la variance inter-imputation et appliqué à l'étape de l'agrégation des paramètres. L'imputation multiple est une expression générale qui comprend plusieurs manières de procéder. Généralement tous les processus d'imputation multiple ont en commun quatre étapes :

1. L'imputation (rappelons que différentes méthodes peuvent être appliquées à cette étape);
2. Les analyses
3. L'agrégation des paramètres estimés;
4. Le calcul du taux d'information manquante.

L'imputation

Différentes implantations de cette technique peuvent être retrouvées dans différents logiciels. La procédure implantée dans SAS v9.1.3 est ici décrite. L'imputation effectuée par PROC MI utilise la méthode de la régression pour identifier les valeurs à imputer. La

régression effectuée utilise le maximum de vraisemblance afin d'optimiser l'utilisation de l'information disponible. Cette régression par maximum de vraisemblance est résolue en utilisant l'algorithme EM. Les estimations obtenues de cet algorithme, une matrice de variance et covariance avec un vecteur de moyenne, sont utilisées comme valeurs de départ pour la procédure d'augmentation de données.

L'augmentation de données

L'augmentation de données est également une procédure itérative en deux étapes. La première étape est une imputation où les données manquantes sont simulées en fonction des estimations actuelles des paramètres. La seconde étape simule de nouveaux paramètres en fonction des données actuelles (observées et imputées). C'est une méthode appartenant à la famille des chaînes de Markov Monte Carlo (MCMC), puisque toute l'information d'une itération d'augmentation de données se trouve également dans l'itération précédente. Cette situation engendre le fait que les paramètres estimés lors de deux itérations consécutives sont très similaires, en fait, trop pour pouvoir être considérés comme issu de tirages aléatoires d'une population. La sélection de paramètres éloignés de plusieurs itérations ressemble plus à un tirage aléatoire des paramètres au sein de la population. L'étape d'imputation est effectuée à partir de la procédure PROC MI de SAS v.9.1.3. Par défaut, les premiers paramètres sélectionnés par la procédure sont ceux obtenus après 200 itérations et les suivants le sont à intervalle de 100 itérations. Des outils de diagnostic de la chaîne sont également disponibles. Ces outils sont la fonction d'autocorrélation et le graphique de la série chronologique. Le premier permet d'assurer que les différents tirages sont bien indépendants et le second permet d'assurer que la chaîne est restée à l'intérieur de certaines bornes.

Dans le cas des données longitudinales, la procédure d'imputation multiple basée sur une distribution multivariée normale n'est pas idéale puisque les données mesurées à différents temps de mesure ne sont pas indépendantes. Les modèles d'analyse de ce type de données permettent, en incluant des paramètres aléatoires, à la moyenne d'une variable

(ordonnée aléatoire) ainsi que la covariance entre VD et VI (pente aléatoire) d'être différente pour chaque individu. Graham (2009) décrit une méthode d'imputation qui inclut, dans le modèle d'imputation, une variable factice pour chaque groupe afin de simuler les ordonnées et les pentes aléatoires. Cette méthode permet d'utiliser l'imputation multiple sous le modèle normal multivarié. Dans le cas des données longitudinales, l'utilisation de cette méthode requiert un nombre très important de variables factices, réduisant d'autant la puissance de l'analyse et la possibilité d'utiliser d'autres variables réellement informatives.

Allison (2001) propose une autre méthode pour utiliser le modèle normal. Elle est similaire à celle de Graham (2009) au sens où la dépendance entre les observations est modélisée dans le modèle d'imputation. La méthode proposée par Allison demande d'exécuter l'imputation avec la base de données formatée de manière à ce que chaque temps de mesure de la variable dépendante soit dans une variable distincte (c.-à-d., la matrice de donnée contient une ligne par personne). Tous les temps de mesure sont utilisés comme variables explicatives des données manquantes aux autres temps de mesure. Une fois les données imputées, les données peuvent être reformatées de manière à permettre l'utilisation des méthodes d'analyse des données longitudinales (c.-à-d. une seule variable dépendante contenant tous les temps de mesure et une ligne par temps de mesure, par personne). Cette méthode permet à la procédure d'imputation d'estimer des moyennes et covariances distinctes pour chaque individu et donc de prendre en compte l'état niché des données.

Suite à l'obtention de plusieurs bases de données complètes, les analyses sont menées sur chacune des bases de données en utilisant les méthodes statistiques usuelles (régression simple, multiple, logistique, poisson, multiniveau, etc.). Suite aux analyses, plusieurs estimations des paramètres sont obtenues. L'agrégation des paramètres se fait simplement en exécutant la moyenne des paramètres estimés.

3.2.3. La pondération

La pondération est une méthode corrigeant l'importance de chacune des observations lors d'une analyse. Dans le cas de l'analyse d'un échantillon avec données manquantes, la pondération peut être utilisée pour donner plus d'importance aux valeurs observées en fonction de certaines caractéristiques associées à l'absence des données. Ces poids peuvent être calculés par l'inverse de la probabilité qu'une donnée soit manquante.

L'utilisation de ces poids est fondée sur la théorie du score de propension développé par Rosenbaum et Rubin (Rosenbaum & Rubin, 1983). Le score de propension a été développé pour niveler les différences entre les groupes lors de l'analyse de l'effet d'un traitement dans une enquête expérimentale. Le principe est que la distribution d'un vecteur de covariables, conditionnelle au score de propension, est équivalente dans le groupe contrôle et dans le groupe intervention (D'Agostino, 1998). Le même principe peut également être appliqué au traitement des données manquantes. Il s'agit de niveler les différences entre les individus ayant des valeurs complètement observées et ceux ayant des valeurs manquantes (D'Agostino & Rubin, 2000; Rosenbaum & Rubin, 1984). Carpenter et Kenward (2008) ont démontré que cette méthode produisait des résultats valides dans certaines conditions, mais qu'elle était très sensible au choix du modèle de pondération (Carpenter, Kenward, & Vansteelandt, 2006; Carpenter & Kenward, 2008). Le score de propension, ou le poids, est calculé à l'aide de régression logistique. En utilisant un indicateur binaire (0=donnée présente; 1=donnée manquante) comme variable dépendante et une série de covariables comme variables indépendantes, l'on obtient la probabilité qu'une donnée soit manquante, conditionnelle aux covariables utilisées dans la régression logistique. L'inverse de cette probabilité est utilisé pour pondérer les observations.

L'exemple suivant illustre l'utilisation cette technique de pondération. Posons que les données complètes sont :

Groupe	A			B			C		
Réponse	1	1	1	2	2	2	3	3	3

La moyenne des données complètes est : 18/9=2.

Posons maintenant que les données ont été partiellement observées et que nous obtenons :

Groupe	A			B			C		
Réponse	1	?	?	2	2	2	?	3	3

L'estimation de la moyenne à partir de ces données est biaisée : 13/6=2,17.

Notons que la probabilité de répondre n'est pas la même dans tous les groupes. La probabilité de répondre dans le groupe A est 1/3, elle est de 1 dans le groupe B et de 2/3 dans le groupe C. Reprenons le calcul de la moyenne à partir des données incomplètes, mais en pondérant par un facteur correspondant à l'inverse de la probabilité de répondre dans chaque groupe:

$$\frac{1 * \frac{3}{1} + (2+2+2) * 1 + (3+3) * \frac{3}{2}}{\frac{3}{1} + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}} = 2$$

L'estimation de la moyenne à partir de données incomplètes est maintenant non biaisée grâce à la pondération.

Le maximum de vraisemblance, l'imputation multiple et la pondération sont des méthodes qui utilisent toute l'information disponible. Le prochain chapitre présente les objectifs, l'échantillon et les variables.

Chapitre 4 : Méthodologie

4.1. Objectifs de la recherche

Les analyses menées visent deux objectifs différents qui ne doivent pas être confondus. Le premier objectif vise à comparer différentes méthodes d'analyse permettant de prendre en compte les données manquantes. Sachant que ces méthodes supposent un mécanisme de données manquantes différent, elles sont considérées comme différents moyens qui peuvent être utilisés pour évaluer l'effet d'une intervention en présence de données manquantes. Le second objectif est de tester différentes hypothèses sur le mécanisme des données manquantes pour évaluer la sensibilité des conclusions. Ces hypothèses sont testées au sein de la procédure d'imputation multiple. L'intérêt d'évaluer différents mécanismes au sein de la même méthode d'analyse réside dans le fait que ces méthodes ne peuvent toutes être menées sous le même postulat sur le mécanisme des données manquantes. Ainsi, deux préoccupations guident l'interprétation des résultats. L'une concerne la méthode d'estimation et l'autre, le mécanisme des données manquantes.

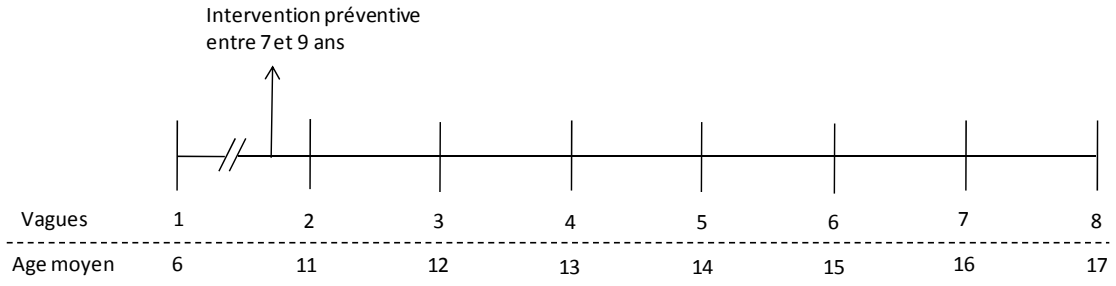
4.2. Échantillon

L'enquête longitudinale-expérimentale de Montréal (ELEM) a débuté en 1984. Les participants sont des garçons qui fréquentaient 53 écoles publiques francophones et défavorisées de Montréal. Lors de l'année préscolaire (maternelle), les enseignants ont évalué les garçons : 87 % des enseignants contactés ont évalué un total de 1161 garçons. Dans le but de créer un échantillon culturellement homogène, uniquement les garçons dont les parents étaient nés au Canada et dont la langue maternelle était le français ont été sélectionnés. En retirant les garçons qui ne répondaient pas aux deux critères mentionnés ainsi que ceux dont les parents ont refusé de participer à l'enquête ou qui n'ont pu être rejoints suite à la vague initiale de collecte, l'échantillon a été ramené à 1037 garçons, soit 89 % des garçons initialement évalués (Tremblay, Pihl, Vitaro, & Dobkin, 1994).

Les données furent récoltées à l'aide du *Social Behavior Questionnaire* (Tremblay, McCord, et al., 1991). Le questionnaire contient 38 items regroupés selon quatre composantes : comportements perturbateurs (13 items), anxiété (5 items), inattention (4 items) et prosocialité (10 items). L'échelle de comportements perturbateurs (*Alpha de Cronbach*=0,93*)* (Boisjoli, Vitaro, Lacourse, Barker, & Tremblay, 2007) comprend trois sous-catégories : agressivité, comportements d'opposition et hyperactivité. Les garçons ayant obtenu un résultat situé au-dessus du 70^{e} percentile (N=250) sur l'échelle de comportements perturbateurs ont été considérés comme les plus à risque de présenter des comportements antisociaux et de décrocher de l'école (Tremblay, 1994). Ces garçons ont été assignés aléatoirement à l'un des trois groupes suivants : groupe expérimental (N=69), groupe contrôle (N=58) ou groupe témoin (N=123). Les groupes témoin et contrôle furent ensuite fusionnés puisqu'ils ne présentaient pas de différences significatives sur l'ensemble des variables mesurées (Boisjoli et coll., 2007).

La première vague de collecte a eu lieu en 1984 alors que les garçons étaient âgés en moyenne de 6,1 ans (e.-t.=0,32). Le programme de prévention fut administré entre 1985 et 1987 alors que les garçons avaient entre 7 et 9 ans (Boisjoli et coll., 2007). Les données des vagues subséquentes ont été collectées selon un intervalle annuel régulier entre 1989 et 1995, soit de 11 à 17 ans.

Figure 2 : Vagues de collecte de données utilisées



4.2.1. Programme de prévention

Le programme de prévention visait à diminuer l'utilisation de la violence chez les garçons. L'intervention visait à la fois les garçons, leurs parents et leurs enseignants. Les participants assignés au groupe expérimental ont participé à des séances de développement des compétences sociales. Ces séances étaient tenues à l'école en petits groupes de quatre à sept participants. Un ratio de trois enfants prosociaux pour un enfant à risque fut maintenu. Le volet parental visait à développer des pratiques parentales efficaces. Ce volet était basé sur le *Oregon Social Learning Center Model* (Patterson, 1975). Enfin, de l'information et du support concernant les garçons à risque furent transmis aux enseignants.

4.3. Données

Les données utilisées portent sur huit temps de mesure et comportent 13 variables, dont sept sont les différentes mesures de la violence dans le temps et les six autres sont des covariables mesurées à six ans. Seule la variable dépendante violence a été utilisée longitudinalement. Les covariables utilisées ont été mesurées à l'âge de six ans, avant que l'intervention préventive n'ait lieu.

4.3.1. Variable dépendante - Violence

La mesure de violence utilisée aux vagues deux à huit était différente de la sous-catégorie de l'échelle de comportements perturbateurs utilisée pour identifier les garçons à risque. Il s'agit d'une mesure autodéclarée de la fréquence de sept comportements lors des 12 derniers mois. Les comportements sont : menacer d'attaquer une personne, se batailler à coup de poing, attaquer une personne innocente, participer à une bataille au sein d'un groupe, lancer des objets vers d'autres personnes, porter une arme et faire usage d'une arme lors d'une bataille. La fréquence fut mesurée sur une échelle à quatre niveaux (jamais, quelques fois, fréquemment, toujours). La somme de la fréquence des sept comportements a permis la constitution de l'échelle de violence (*Alpha de Cronbach*=0,77) (Lacourse et

coll., 2002). La distribution asymétrique de l'échelle de violence a exigé l'utilisation du logarithme naturel de l'échelle dans les analyses présentées.

4.3.2. Covariables

Le temps

La variable mesurant le temps qui a été utilisée dans les analyses longitudinales représente l'âge des garçons au moment de la mesure. Pour que l'ordonnée à l'origine représente l'estimation de la violence à 11 ans, soit le premier temps de mesure suivant l'intervention, la variable a été ramenée sur une échelle de 0 à 6, où 0 représente la mesure à 11 ans et 6, la mesure à 17 ans.

Toutes les covariables ont été mesurées avant l'implantation du programme de prévention. Elles sont utilisées à titre de variables contrôles. La variable *intervention* sert à distinguer les garçons appartenant au groupe contrôle (*intervention*=0; N=181) de ceux ayant reçu l'intervention (*intervention*=1; N=69).

La mesure de l'adversité familiale a été constituée à partir de différentes caractéristiques obtenues lors d'entrevues téléphoniques avec la mère des garçons. Les caractéristiques utilisées ont été mesurées à la fois auprès de la mère (ou du père dans le cas de familles monoparentales où la mère est absente) et, s'il y a lieu, du père. Elles sont : l'âge du parent à la naissance du premier enfant, le nombre d'années d'éducation, la profession (qui fut transformé en index de statut socioéconomique) et la structure familiale. Les garçons se voyaient attribuer la valeur de 0 sur les trois premiers critères s'appliquant aux parents si ceux-ci se trouvaient au-dessus du 30^{e} percentile et de 1 s'ils étaient au dessous. Le rang percentile a été calculé à partir de l'échantillon complet (N=1161). Pour la structure familiale, une valeur de 0 était attribuée aux garçons qui vivaient avec leurs deux parents biologiques et de 1 pour tous les autres types de structure familiale. Puisque le maximum de l'échelle diffère selon qu'un garçon est dans une famille avec deux parents ou monoparentale, le score total obtenu a été divisé par sept pour les garçons ayant deux

parents et par quatre pour ceux d'une famille monoparentale. Ainsi, une valeur élevée représente un haut niveau d'adversité familiale (Tremblay, Loeber et al., 1991; Vitaro, Brendgen & Tremblay, 2001).

La prosocialité, l'agressivité physique, l'opposition, l'inattention et l'hyperactivité ont été mesurées à partir du SBQ par l'enseignant qui connaissait le mieux l'enfant. Chaque comportement a été évalué selon différents items à partir d'une échelle à trois niveaux indiquant la fréquence de chacun de ces items. Les enseignants devaient indiquer si chacun des critères s'appliquait souvent, parfois ou jamais.

La prosocialité a été mesurée à partir de dix items : essaie d'arrêter les querelles, invite les autres à se joindre, essaie d'aider quelqu'un qui est blessé, aide à ramasser les objets échappés par quelqu'un d'autre, encourage le travail des enfants moins capables, montre la sympathie envers quelqu'un qui a fait une erreur, aide les enfants qui ont de la difficulté avec une tâche, aide les enfants qui sont malades, réconforte les enfants en pleurs ou fâchés, et aide nettoyer le désordre causé par quelqu'un d'autre (*Alpha de Cronbach*=0,92). L'agressivité physique a été mesurée à partir de trois items : se bat avec les autres enfants, donne des coups de pied, mord ou frappe les autres enfants et intimide d'autres enfants (*Alpha de Cronbach*=0,84). L'opposition a été mesurée à partir de cinq items : ne partage pas son matériel, est irritable, est désobéissant, blâme les autres et est peu attentif aux autres (*Alpha de Cronbach*=0,84). L'inattention a été mesurée à partir de deux items : est facilement distrait et a une faible concentration (*Alpha de Cronbach*=0,74). L'hyperactivité a été mesurée à partir de deux items : remue continuellement et est très agité (*Alpha de Cronbach*=0,87) (Nagin & Tremblay, 2001; Pagani, Boulerice, Tremblay, & Vitaro, 1997).

4.3.3. Statistiques descriptives et aperçu des données manquantes

Les données ont été complètement observées pour 143 (57,2 %) des enfants; 107 garçons comportent donc une ou plusieurs données manquantes. Selon la méthode de la matrice clairsemée, 377 observations sont manquantes sur un total de 3250, représentant

11,6 % (377/(250*13)) d'observations manquantes. Cette proportion calculée par groupe est de 10,6 % pour le groupe intervention et de 11,7 % pour le groupe contrôle. Dix des treize (71,4 %) variables utilisées dans les analyses comportent des données manquantes, dont les sept temps de mesure de la violence. Les données manquantes se situent uniquement au niveau des individus. Pour les garçons comportant des données manquantes, la proportion moyenne d'observations manquantes est de 25,2 % (377/(107*14)). Calculée en fonction des variables, la proportion moyenne d'observations manquantes est de 15,1 % (377/(250*10)). Ces proportions indiquent que la densité des données manquantes est plus importante au niveau des individus que des variables.

Les données manquantes concernant la variable dépendante (violence) sont les plus nombreuses, leurs proportions varient de 11,2 % à 30,4 %. Le phénomène d'attrition est perceptible ici puisque la proportion de données manquantes pour la variable dépendante augmente systématiquement de 11 à 17 ans. La proportion de données manquantes par groupe semble équivalente. L'ampleur du phénomène d'attrition ainsi que les possibles différences par groupes seront évaluée plus précisément dans la section sur l'analyse du schéma monotone des données manquantes. Le tableau IV de la page suivante contient le nombre de données observées et manquantes selon le temps de mesure et le groupe. `

Les figures 3 et 4 présentent l'évolution de la moyenne de violence par groupe selon que l'on utilise les cas complets ou l'ensemble des observations disponibles. Les deux groupes semblent suivre une trajectoire cubique. Les moyennes de violence du groupe intervention sont légèrement inférieures à celles du groupe contrôle. Il semble que les moyennes des deux groupes estimées à partir de l'ensemble des données observées soient plus près que celles estimées à partir des cas complets.

Tableau IV : Nombre et proportion de données manquantes pour la mesure de *violence* selon le temps et le groupe

	Âge		**6**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Contrôle	Observées	N	181	160	154	149	138	136	130	124
	Manquantes	N	0	21	27	32	43	45	51	57
		%	0	11,6	14,9	17,7	23,8	24,9	28,2	31,5
Intervention	Observées	N	69	62	62	56	55	53	50	50
	Manquantes	N	0	7	7	13	14	16	19	19
		%	0	10,1	10,1	18,8	20,3	23,2	27,5	27,5
Total	Observées	N	250	222	216	205	193	189	180	174
	Manquantes	N	0	28	34	45	57	61	70	76
		%	0	11,2	13,6	18,0	22,8	24,4	28,0	30,4

Figure 3 : Moyenne de violence par groupe - Données observées

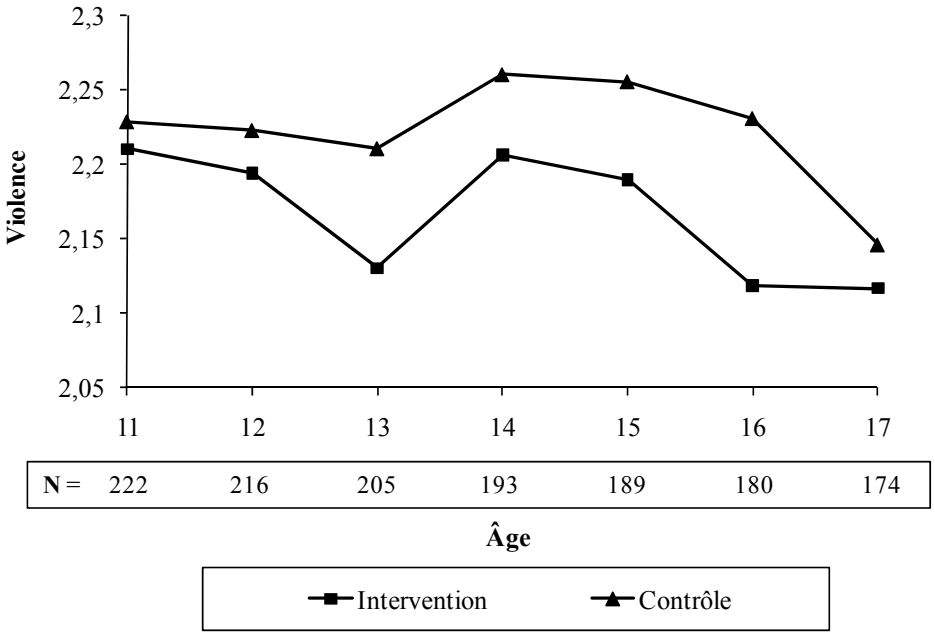
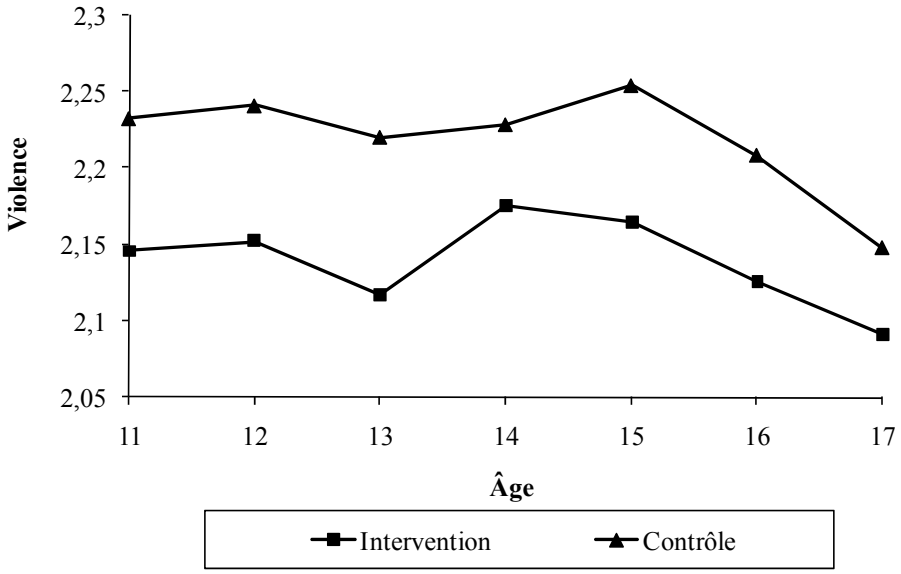


Figure 4 : Moyenne de violence par groupe - Cas complets (N=146)



Les données manquantes concernant les covariables mesurées à 6 ans sont peu nombreuses, leurs proportions varient de 0,4 % à 1,2 %. Considérant la très faible fréquence des données manquantes sur les covariables, elles ne feront pas l'objet d'une analyse approfondie. Lorsque ces variables seront incluses dans les analyses, le N disponible pourra très légèrement varier, mais par souci de simplicité, nous ne porterons pas notre attention sur ces 6 cas. Aussi, nous considérerons les données comme complètement observées pour 146 garçons.

Chapitre 5 : Analyses

Ce chapitre présente les analyses exécutées. La première section (5.1) décrit comment peut être effectuée l'estimation de l'effet d'une intervention dans le cadre d'un devis expérimental. La méthode est ensuite étendue pour s'appliquer à des données mesurées longitudinalement. Une description du processus d'analyse des données récoltées lors d'un devis expérimental où certaines observations sont manquantes est ensuite présentée (Carpenter & Kenward, 2008). La deuxième section (5.2) explique les différentes hypothèses posées concernant le mécanisme des données manquantes et comment les nouvelles méthodes d'estimation en présence de données manquantes (présentées à la section 3.2) ont été appliquées aux données de l'ELEM.

5.1. Analyse de l'effet d'un programme de prévention

Cette section montre comment estimer l'effet d'une intervention dans le cadre d'un devis expérimental. La section 5.1.1 étend cette stratégie aux données longitudinales et la section 5.1.2 présente une stratégie proposée par Carpenter et Kenward pour l'analyse d'un devis longitudinal-expérimental avec données manquantes (Carpenter & Kenward, 2008).

Posons que l'on désire déterminer l'effet d'une intervention sur une variable d'intérêt (Y). L'approche classique consiste à constituer deux groupes, un premier qui recevra l'intervention (I) et un second qui sera utilisé comme groupe contrôle (C). Rausenbaum et Rubin ont décrit une définition statistique très simple de l'effet causal d'une intervention, connu comme la théorie RRH (Rausenbaum & Rubin, 1983). Dans l'équation 5, Δ_i représente l'effet du traitement, elle est la suivante :

Équation 5 : Effet d'un traitement dans une étude transversale

$$\Delta_i = Y_i(I) - Y_i(C)$$

Où : $Y_i(I)$ = mesure si le participant est assigné au groupe expérimental

$Y_i(C)$ = mesure si le participant est assigné au groupe contrôle

L'effet causal d'une intervention est la différence entre le résultat d'un participant subissant un traitement et celui qui aurait été obtenu par ce même participant sans le traitement. Cette définition entraîne : [I] que l'effet causal est défini comme étant propre à chaque participant (représenté par les sous-indices *i* dans l'équation 5) et [II] que l'effet causal ne peut être directement observé puisque chaque participant appartient soit au groupe expérimental, soit au groupe contrôle. La théorie RRH est ainsi fondée sur une approche contrefactuelle.

L'effet causal étant défini individuellement, il peut théoriquement varier d'un participant à l'autre. Il ne peut toutefois être directement observé puisque chaque participant ne peut qu'être dans un seul groupe. Il est alors impossible de mesurer directement l'effet causal individuel. Si un participant est assigné au groupe expérimental, $Y_i(C)$ est conceptualisé comme l'indicateur contrefactuel et ne peut être observé. Ainsi, pour chaque participant, l'indicateur contrefactuel est toujours manquant, une seule des deux situations possibles ayant été mesurée. Cette situation peut être conceptualisée comme un problème de données manquantes. Si le principe de l'assignation aléatoire au groupe (intervention ou contrôle) a été respecté, il est possible d'affirmer que les indicateurs contrefactuels (données manquantes) sont manquants de manière aléatoire (*MCAR*). L'effet causal moyen peut alors être estimé sans biais à l'aide de l'équation suivante :

Équation 6 : Effet causal moyen d'un traitement dans une étude transversale

$$\Delta = E[Y_i(I) - Y_i(C)]$$

La résolution de l'équation 6 permet d'établir l'effet causal moyen de l'intervention ainsi que la variabilité de cet effet. En obtenant l'effet moyen ainsi que sa variance, il est possible de tester si l'effet de l'intervention est significatif.

Les hypothèses à tester peuvent être formulées de la manière suivante :

H_0 : Le groupe contrôle et le groupe intervention ne se distinguent pas significativement; Δ n'est pas significativement différent de zéro.

H_1 : Le groupe intervention est significativement différent du groupe contrôle; Δ est significativement différent de zéro.

Pour tester si Δ est significativement différent de zéro, des tests unicaudaux sont généralement utilisés puisque l'intervention vise explicitement une direction de l'effet.

5.1.1. Effet d'un programme à partir de données longitudinales

Les données longitudinales font référence à des mesures répétées dans le temps. Les mesures répétées peuvent être interprétées de différentes manières. Il peut s'agir (1) de la probabilité qu'un événement survienne après un certain laps de temps, on parle d'analyse de survie; (2) de la probabilité d'être dans un état en fonction de l'état précédent, cela renvoie également à l'analyse de survie; (3) de la séquence des valeurs de plusieurs variables d'intérêt que l'on suppose reliées, ceci renvoie aux analyses structurelles; finalement, (4) de l'évolution temporelle d'une caractéristique donnée, on parle alors d'analyse de courbe de croissance. Les analyses de courbes de croissance peuvent être paramétriques ou non paramétriques (Nagin, 1999; Raudenbush, 2001). Les analyses présentées ici portent sur l'évolution dans le temps de la violence, nous faisons alors référence à des courbes de croissance.

Les principes d'évaluation de l'effet d'une intervention dans les enquêtes transversales peuvent également être appliqués à l'analyse de courbes de croissance et se traduisent en ces termes : une seule des deux trajectoires possibles pour chaque individu sera observée (la trajectoire contrôle ou la trajectoire expérimentale); si les individus ont été assignés aléatoirement à un ou l'autre des groupes et qu'il n'y a pas d'attrition, alors l'effet du traitement pourra être estimé sans biais. S'il y a présence d'attrition, le mécanisme des données manquantes doit être analysé afin de déterminer si les analyses peuvent être menées en contrôlant pour les causes potentielles des données manquantes (Graham, 2009; Raudenbush, 2001).

Le fait que les mesures répétées sont prises auprès d'une même unité implique que les données récoltées ne sont pas toutes indépendantes. Cette caractéristique des mesures

répétées implique que les méthodes de régression classique postulant l'indépendance des observations ne peuvent être appliquées adéquatement. Des méthodes d'analyse des données basées sur la régression et qui tiennent compte de la non-indépendance des observations ont été développées au milieu des années 70. Elles sont connues sous les expressions modèles hiérarchiques, modèles multiniveaux ou modèles mixtes. Les équations hiérarchiques intègrent des termes aléatoires permettant d'estimer la variation intra- et inter-groupe d'un paramètre décrivant la distribution générale des données. Dans le cas de l'analyse de courbes de croissance, ces méthodes permettent de modéliser la variation intra- et inter-individuelle. La variation intra-individuelle permet de modéliser les données récoltées auprès d'une même personne et la variation inter-individuelle, de postuler que cette variation peut être différente selon les individus.

La littérature présente le modèle de deux manières, soit en écrivant plusieurs équations décrivant chacun des niveaux d'analyse ou en écrivant une seule équation combinant l'ensemble des paramètres à estimer. Nous privilégions la première approche et utilisons la notation présentée par Singer et Willet dans *Applied Longitudinal Data Analysis* (Singer & Willet, 2003). Pour évaluer l'effet d'une intervention, une variable indiquant l'appartenance au groupe (X_i) doit être insérée au deuxième niveau ($\beta_{01} et\, \beta_{11}$ dans l'équation 7). C'est ce paramètre qui représente la différence entre les groupes sur l'évolution des courbes de croissance.

Équation 7 : Modèle hiérarchique de courbe de croissance avec paramètre associé à l'appartenance au groupe intervention ou contrôle

$$Y_{it} = \pi_{0i} + \pi_{1i}\alpha_{it} + \varepsilon_{it} \qquad \varepsilon_{it} \sim (0, \sigma^2)$$
$$\pi_{0i} = \beta_{00} + \beta_{01}X_i + \zeta_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}X_i + \zeta_{1i}$$

$$Avec \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}\, \tau_{01} \\ \tau_{10}\, \tau_{11} \end{pmatrix}\right]$$

Et où : β_{00} est le paramètre associé à l'ordonnée à l'origine;

β_{01} est le paramètre indiquant l'effet de l'intervention sur l'ordonnée à l'origine;

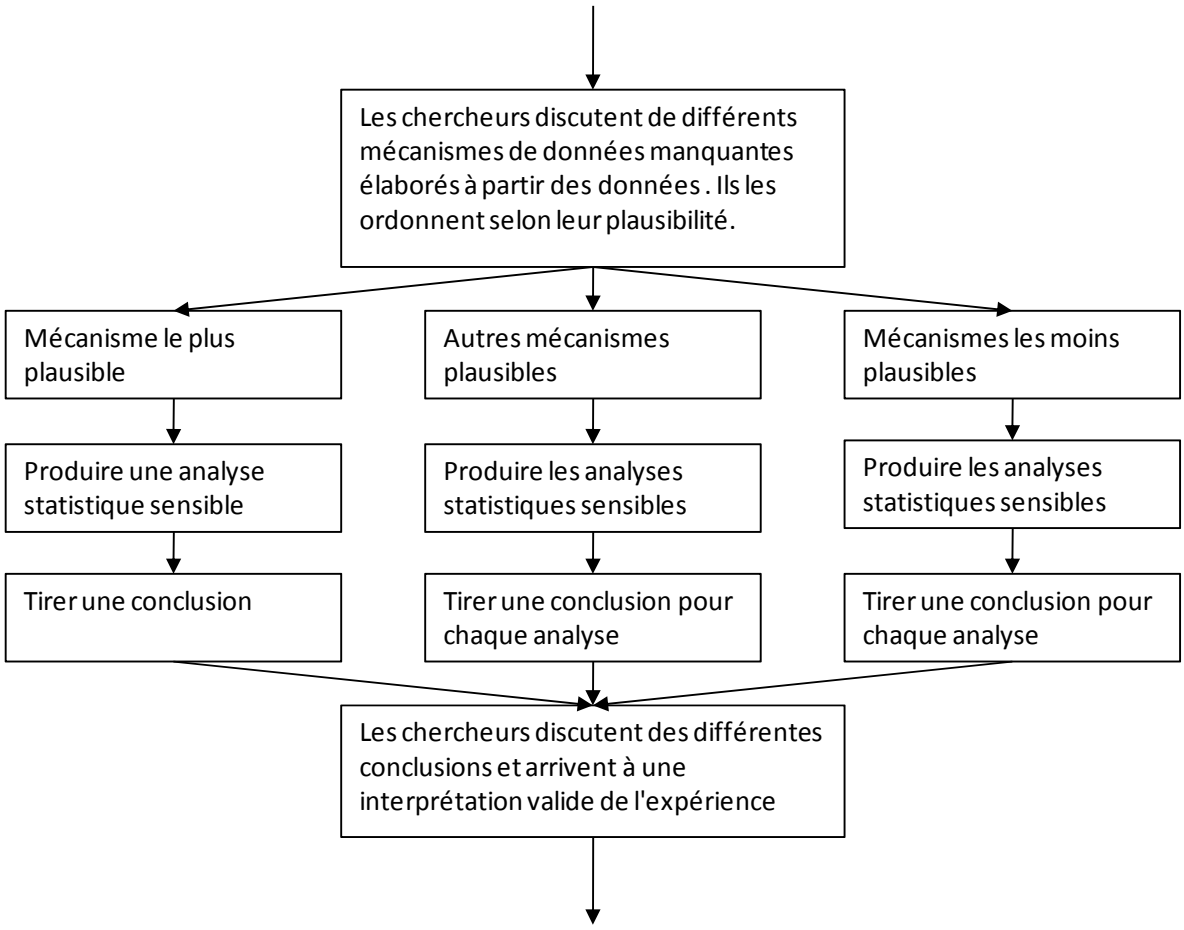
β_{10} est le paramètre associé au temps linéaire (la pente);

β_{11} est le paramètre indiquant l'effet de l'intervention sur la pente.

5.1.2. Analyse de l'effet d'un programme avec données manquantes

De nombreux auteurs ont déjà décrit comment l'estimation d'un paramètre en présence de données manquantes peut être biaisée (voir section 2.1). Toutefois, on ne peut pas directement conclure que l'estimation sera biaisée en présence d'une quantité importante de données manquantes. L'effet de l'intervention pourra être estimé sans biais si la méthode d'analyse utilisée tient compte des raisons ayant mené à ce que des données manquent et que ces raisons postulées sont les bonnes. Puisque la ou les raisons possibles ne peuvent être identifiées hors de tout doute, Carpenter et Kenward (2008) proposent d'utiliser une approche transparente et systématique. Il s'agit de poser différentes hypothèses sur la cause des données manquantes et de les ordonner de la plus plausible à la moins plausible. Ensuite, effectuer les analyses statistiques appropriées à partir de méthodes d'estimation tenant compte de ces hypothèses et comparer les résultats de ces différentes analyses avant de tirer une conclusion sur l'effet de l'intervention. Ce processus, illustré par la figure 5, constitue une *analyse de sensibilité*.

Figure 5 : Processus d'analyse d'un devis expérimental avec données manquantes



Traduction libre de : (Carpenter & Kenward, 2008)

Comme il a déjà été mentionné, le mécanisme des données manquantes ne peut être établi avec certitude. Les auteurs proposent donc de considérer le mécanisme retenu comme un postulat sous lequel l'analyse est menée. Carpenter et Kenward (2008) écrivent :

> *This approach is fundamentally different from common practice where the analyst regards missing data as a 'problem' and casts around for a 'solution', usually a computationally simple procedure. Once the data have been analysed using this procedure the problem is regarded as being 'solved'. [...]*

> *Our proposed approach may give sharper conclusions than the practice of replacing each missing observation with either a best or worst value, seeking that combination which gives the smallest estimated intervention effect. We do not believe this makes it misleading, though. Rather, best/worst values, which often represent extremely implausible, degenerate probability distribution for the missing data, are more likely to mislead. (Carpenter et Kenward, 2008, p. 12-13)*

5.2. Hypothèses sur le mécanisme des données manquantes et méthodes d'estimation

Cette section contient les six hypothèses posées sur le mécanisme des données manquantes ainsi que la description de l'application aux données de l'ELEM des nouvelles méthodes d'analyse de données en présence de données manquantes.

5.2.1. Hypothèses sur le mécanisme des données manquantes

Rappelons que le mécanisme des données manquantes renvoie à la cause relative à l'absence des données. Il est impossible d'assurer hors de tout doute que les causes ont bien été identifiées. Dans le cadre de la théorie sur les données manquantes, toute analyse effectuée à partir d'observations incomplètes suppose un certain mécanisme de données manquantes. Nous appliquons le processus d'analyse suggéré par Carpenter et Kenward (2008) et posons différentes hypothèses sur le mécanisme des données manquantes afin de tester leurs impacts sur les conclusions de l'évaluation de l'effet de l'intervention. Le choix des hypothèses sur le mécanisme des données manquantes a été guidé par la littérature sur les données manquantes et par la littérature sur l'analyse de l'effet d'une intervention au sein d'un devis expérimental. Six hypothèses sont posées et testées.

La première hypothèse que nous posons est que les données manquantes sont *MCAR*. L'approche classique utilisant la méthode des cas complets est appliquée pour tester cette hypothèse. Les autres hypothèses sont différentes versions de *MAR* et sont

testées à partir de différents modèles d'imputation incorporés au sein de la procédure d'imputation multiple.

Menant à la seconde hypothèse, on trouve dans la littérature sur le traitement des données manquantes, une technique nommée l'analyse par mélange de schémas (pattern mixture) (Schafer, 1997; Schafer & Graham, 2002). Théoriquement, on suppose que les mécanismes des données manquantes sont distincts pour chaque schéma de données manquantes. En pratique, cela implique que les raisons pour lesquelles un participant quitte l'étude et ne revient pas (schéma monotone) seraient différentes de celles qui expliqueraient qu'un participant ait manqué, disons, uniquement la seconde vague de collecte. En pratique, en raison de la faible taille de certains schémas, cette méthode d'analyse ne s'applique pas aisément.

Une stratégie permettant de limiter le nombre de schémas à analyser avec des données longitudinales est proposée par Yang et Shoptaw (2005). Elle consiste à distinguer le schéma monotone de tous les autres schémas. Les schémas qui ne sont pas monotones sont qualifiés d'intermittents. Le schéma intermittent regroupe tous les schémas où des données sont manquantes à un certain temps de mesure et où des données ont été observées à un temps de mesure antérieur et à un temps de mesure postérieur (Yang & Shoptaw, 2005). Le schéma monotone indique que lorsqu'une donnée est manquante à un certain temps de mesure pour un certain individu, les données pour ce même individu sont également manquantes pour tous les temps de mesure suivants.

Ainsi, pour la seconde hypothèse sur le mécanisme des données manquantes, la stratégie suggérée par Yang sera appliquée, soit l'analyse séparée du schéma intermittent et du schéma monotone. L'analyse des schémas des données manquantes de l'ELEM portera sur un schéma strictement monotone (N=63) et un schéma aléatoire (N=41). Notons que les 146 cas complets sont utilisés dans les deux modèles d'imputation afin d'ajouter de l'information au processus d'imputation.

Pour la troisième hypothèse sur le mécanisme des données manquantes, il sera supposé que le mécanisme expliquant les données manquantes est distinct selon le groupe intervention ou contrôle. Lors de l'analyse de l'effet d'une intervention, tout comme il est essentiel que les groupes soient statistiquement équivalents au début de l'enquête, leur évolution durant la période de collecte doit pouvoir se faire de manière indépendante. Il est donc tout à fait cohérent d'étendre cette logique au mécanisme des données manquantes et de tester s'il est différent dans les deux groupes.

La stratégie idéale aurait été de combiner les seconde et troisième hypothèses pour évaluer l'impact de considérer les schémas à l'intérieur des groupes comme ayant des causes différentes. Toutefois, en segmentant l'échantillon ELEM de cette manière, les analyses du mécanisme seraient menées sur très peu de cas, ce qui, à notre avis, ne permet pas d'obtenir des résultats suffisamment valides.

Enfin, trois autres hypothèses sur le mécanisme des données manquantes ont été testées, elles supposent toutes que le mécanisme est unique pour tout l'échantillon (sans distinction de schéma ou de groupe). D'abord, l'hypothèse que les données manquantes sont *MAR* et qu'aucune covariable n'est associée aux données manquantes. Ensuite, l'hypothèse que les données manquantes sont *MAR* et que le mécanisme est associé à certaines covariables identifiées à l'aide de régressions logistiques, et enfin, l'hypothèse que les données manquantes sont *MAR* et que le mécanisme est associé à toutes les covariables. Ces trois mécanismes sont considérés comme les moins plausibles. Le tableau V contient un résumé des différentes hypothèses posées sur le mécanisme des données manquantes.

Les analyses postulants *MAR* sont effectuées en définissant différents modèles d'imputation au sein de la procédure d'imputation multiple. La fonction SEED de la procédure PROC MI de SAS a été utilisée. Elle fournit à la procédure le point de départ de l'algorithme générant les nombres aléatoires qui sont utilisés lors de l'étape d'augmentation de données de l'imputation multiple. En utilisant une valeur de SEED identique pour

chaque modèle d'imputation multiple, cela permet d'empêcher que les différences observées entre ces différents modèles soient attribuables au générateur de nombre pseudoaléatoire qui est lancé à chaque exécution de cette procédure. La valeur « 051109 » a été utilisée pour la fonction SEED.

Tableau V : Postulats sur le mécanisme des données manquantes

Modèle	**Ensemble de données**	**Hypothèse sur le mécanisme**	**Covariables postulées comme associées aux données manquantes**
A	Cas complets (N=146)	*MCAR*	---
B[2]	Schéma intermittent (N=187)	*MAR*	Base[1] + Variables associées à r_{ij}
	Schéma monotone (N=209)	*MAR*	Base[1] + Variables associées à r_{ij}
C	Groupe contrôle (N=181)	*MAR*	Base[1] + Variables associées à r_{ij}
	Groupe intervention (N=69)	*MAR*	Base[1] + Variables associées à r_{ij}
D	Toutes les données (N=250)	*MAR*	Base[1]
E	Toutes les données (N=250)	*MAR*	Base[1] + Variables associées à r_{ij}
F	Toutes les données (N=250)	*MAR*	Base[1] + Toutes les covariables

[1] Le terme « Base » indique tous les temps de mesures de la violence.

[2] Pour l'analyse des schémas monotone et intermittent, les 146 cas complets ont été laissé dans chaque matrice de données.

5.2.1.1. Notes sur le schéma monotone

Notons que lors de l'utilisation de données longitudinales, le schéma monotone correspond à ce qui est généralement décrit comme le phénomène d'attrition. Dans le cadre de la théorie sur les données manquantes, l'attrition est un des nombreux schémas de données manquantes qui peuvent être présents au sein d'une matrice de données. L'impact de l'attrition sur les estimations est un problème spécifique qui a souvent été abordé de

manière isolée, sans tenir comptes des autres schémas ni de la théorie sur les données manquantes.

Selon Hansen (Hansen, Collins, Malotte, Johnson, & Fielding, 1985), dans le cas d'une étude à plan longitudinal-expérimental, deux questions sont d'intérêt pour déterminer si l'attrition peut créer un biais. Premièrement, est-ce que le taux d'attrition est différent entre les groupes, et deuxièmement, est-ce que les individus qui quittent l'étude ont des valeurs différentes sur les covariables mesurées avant l'intervention par rapport à ceux qui restent jusqu'à la fin de l'étude (Hansen, et al., 1985)? Une troisième question pourrait y être ajoutée, advenant des différences entre les participants et les non-participants, est-ce que ces différences sont les mêmes à travers les groupes? Réexaminons ces questions en utilisant la nomenclature propre à la théorie sur les données manquantes.

Un taux d'attrition différent selon le groupe suppose que, proportionnellement à la taille des groupes, la taille du schéma monotone est plus élevée dans un groupe que dans l'autre. La taille d'un schéma n'est pas une question d'intérêt dans la théorie sur les données manquantes. Comme démontré lors de simulations, même un nombre important de données manquantes ne cause pas nécessairement de biais si le mécanisme des données manquantes est intégré dans les analyses. La question est plutôt de savoir si le mécanisme des données manquantes associé à un schéma est le même selon le groupe.

Il a tout de même évalué si le taux d'attrition est différent selon le groupe. Pour chaque temps de mesure, une régression logistique bivariée est exécutée où la variable dépendante binaire indique si la donnée est présente ($r_{ij}=0$) ou absente ($r_{ij}=1$) (référence est $r_{ij}=0$), la variable indépendante indiquant l'appartenance au groupe.

La seconde question est d'identifier si les individus qui quittent l'étude ont des valeurs différentes sur les covariables par rapport aux individus qui restent jusqu'à la fin. Si les individus qui quittent l'étude ont des valeurs différentes sur les covariables mesurées avant l'intervention, cela suppose que les données manquantes puissent être prédites à partir de ces variables. Cela correspond assez bien à ce qu'est le mécanisme de données

manquantes *MAR*. Reste à savoir si les différences sont les mêmes à travers les groupes. Cela peut être conceptualisé au sein de la théorie sur les données manquantes de la manière suivante : est-ce que le mécanisme des données manquantes associé au schéma monotone du groupe intervention est le même que le mécanisme associé au schéma monotone du groupe contrôle?

Différentes stratégies peuvent être utilisées pour répondre à ce deuxième volet d'analyse du schéma monotone. L'une d'entre elles consiste à utiliser l'analyse de survie et de considérer les individus n'ayant aucune donnée manquante comme censurés à droite, c'est-à-dire qu'ils n'ont pas connu l'événement. Nous appliquerons une méthode plus simple qui utilise la régression multiple. Une variable binaire est créée indiquant si un individu fait partie du schéma monotone, donc qui a quitté l'étude (battrition=1), ou non (battrition=0). Sept régressions multiples sont ensuite effectuées au sein du schéma monotone, une pour chaque covariable mesurée avant l'intervention. Les variables indépendantes incluses sont « intervention », « battrition » ainsi que l'interaction entre les deux. Si le paramètre associé à intervention est significativement différent de zéro, cela indique que les participants et les non-participants se différencient au niveau des covariables mesurées avant l'intervention. De même, si le paramètre associé à l'interaction entre intervention et battrition est significativement différent de zéro, ces différences ne seront pas les mêmes selon le groupe.

5.2.1.2. Identification de covariables associées aux données manquantes

Pour guider le choix des covariables associées au mécanisme des données manquantes ayant été postulés, une série de régressions logistiques multivariées sur les variables binaires r_{ij} sont effectuées. Pour chaque temps de mesure, ces variables binaires indiquent donc si une donnée est observée ou manquante. Les covariables insérées dans ces modèles de régression logistique sont les covariables mesurées avant l'intervention. Les covariables ont été retenues comme cause possible de l'absence des données si le niveau de signification obtenu est inférieur à $p = 0{,}10$.

5.2.2. Méthodes d'estimation

Cette section contient la description des méthodes d'estimation telles qu'elles ont été appliquées aux données de l'ELEM. Pour chacun d'elle, les mécanismes de données manquantes supposés sont aussi décrits.

5.2.2.1. Analyse des cas complets

L'analyse des cas complets comprend tous les garçons qui n'ont aucune donnée manquante sur tous les temps de mesure (N=146). L'analyse porte sur 1022 mesures de violence prises à différents temps (7 temps de mesure pour 146 garçons). Ceci représente 58,4 % (1022/1750) des mesures visées par la collecte. Le modèle exécuté postule un mécanisme *MCAR*.

5.2.1.2. Maximum de vraisemblance

L'analyse par maximum de vraisemblance comprend toutes les observations qui ne sont pas des données manquantes. L'analyse porte donc sur 1379 mesures de violence prises à différents temps. Ceci représente 78,8 % (1379/1750) des mesures visées par la collecte. L'estimation par maximum de vraisemblance avec des données manquantes implique deux postulats : le mécanisme est ignorable et les données suivent une distribution normale multivariée. Un mécanisme *MAR*, ignorable, qui dépend uniquement des valeurs de la variable dépendante est postulé lors de cette analyse. Ce postulat suppose que l'information contenue sur tous les temps de mesure de la violence est suffisante pour assurer des résultats non biaisés.

5.2.1.3. Pondération

Le calcul des poids est basé sur la méthode de la probabilité inverse qu'une donnée soit manquante. Une variable (*GRPPRES*) à quatre catégories a été créée représentant la combinaison de l'appartenance au groupe et de la présence de données manquantes sur la variable violence, peu importe le temps de mesure manquant.

Tableau VI : Distribution de la variable de groupement utilisée pour le calcul des poids

Groupe	Données	GRPPRES	N	%
Contrôle	Complètes	0	107	42,8
	Manquantes	1	74	29,6
Intervention	Complètes	2	39	15,6
	Manquantes	3	30	12,0

Le calcul des poids a été effectué à partir d'une régression logistique multinomiale. La catégorie de référence est le fait d'être dans le groupe contrôle et d'avoir des données complètes (*GRPPRES*=0). Ces poids sont ensuite standardisés pour maintenir l'équilibre dans la taille des groupes. Les quelques cas contenant des données manquantes sur les variables indépendantes incluses dans la régression multinomiale se sont vu attribuer la moyenne des poids de leur valeur sur *GRPPRES*.

Comme Carpenter (2006) l'a souligné, les résultats obtenus par cette méthode sont très sensibles au choix du modèle de pondération, c'est-à-dire aux hypothèses sur le mécanisme des données manquantes. Le mécanisme *MAR* est postulé par ce modèle de pondération. Les covariables associées à la cause des données manquantes identifiées au sein de l'échantillon complet ont été utilisées dans le modèle de pondération. L'estimation est également effectuée par maximum de vraisemblance, mais en pondérant à partir de l'ensemble de covariables associées au mécanisme des données manquantes. L'estimation est ici conditionnelle à la probabilité, elle aussi, conditionnelle, qu'une donnée soit manquante.

5.2.1.4. Imputation multiple

Plusieurs modèles d'imputation sont testés. Chacun des modèles contient tous les temps de mesure de la variable dépendante en plus de certaines variables postulées comme étant associées au mécanisme des données manquantes (tableau V). Tous les modèles d'imputation testés supposent un mécanisme *MAR* qui dépend de différentes covariables associées au mécanisme.

5.3 Comparaison des modèles

Les analyses menées visent deux objectifs différents qui ne doivent pas être confondus. Le premier vise à comparer différentes méthodes d'analyse tenant compte des données manquantes. En sachant que ces méthodes supposent un mécanisme de données manquantes distinct, elles sont considérées comme étant différents moyens qui peuvent être utilisés pour évaluer l'effet d'une intervention en présence de données manquantes. Le second objectif est de tester différentes hypothèses sur le mécanisme des données manquantes pour évaluer la sensibilité des conclusions. Ces hypothèses sont testées au sein de la procédure d'imputation multiple. L'intérêt d'évaluer différents mécanismes au sein de la même méthode d'analyse repose sur le fait que ces méthodes ne peuvent pas toutes être menées sous le même postulat portant sur le mécanisme des données manquantes.

Ainsi, ces deux objectifs ont guidé l'interprétation des résultats. Autrement dit, l'une concerne la méthode d'estimation et l'autre, le mécanisme des données manquantes. Les mêmes difficultés sont partagées pour ces deux objectifs quant à la comparaison des résultats obtenus.

La première difficulté est liée au fait que les analyses sont menées à partir de différents ensembles de données. Il devient donc impossible d'utiliser les statistiques classiques, comme le *Bayesian Information Criterion* (*BIC*) ou le logarithme du maximum de vraisemblance, pour juger de l'adéquation d'un modèle aux données. De plus, puisque de vraies données sont utilisées, il est impossible d'établir l'ampleur du biais car la valeur réelle du paramètre estimé n'est pas connue. L'accent sera porté sur la description des différences s'observant entre les résultats obtenus, en évaluant l'impact de différentes hypothèses sur le mécanisme des données manquantes et en utilisant différentes méthodes d'estimation.

Le prochain chapitre présente les résultats des analyses menées.

Chapitre 6 : Résultats

Cette section comprend les résultats de l'analyse de l'effet du programme de prévention visant à diminuer l'expression de violence. Les résultats sont présentés suivant l'ordre exposé au chapitre précédent. Ainsi, la modélisation des courbes de croissance est d'abord présentée, suivie des résultats des analyses multivariées des variables binaires r_{ij} associées à chaque temps de mesure. Rappelons que ces analyses servent à identifier des covariables permettant de prédire qu'une observation est manquante. À partir de ces résultats, les cinq hypothèses sur le mécanisme des données manquantes sont spécifiées et testées à l'aide de différents modèles d'imputation définis au sein de la procédure d'imputation multiple. Finalement, les quatre méthodes d'estimation précédemment discutées sont appliquées aux données de l'ELEM. Il est à noter que toutes les analyses présentées dans ce mémoire ont été produites à l'aide du logiciel SAS/STAT, version 9.1.3 du SAS System pour Windows.

6.1. Modélisation des courbes de croissance

Pour identifier la forme idéale des trajectoires, un modèle multiniveau sans autre covariable que le temps a été exécuté. Les cas complets furent utilisés (N=146) pour évaluer l'évolution dans le temps de la violence. Il fut identifié que c'est une fonction cubique qui représente le mieux les données. Ainsi, le modèle d'analyse de l'effet de l'intervention initialement testé est le suivant :

Équation 8 : Modèle d'analyse initial pour décrire les courbes de croissance

$$y_{ij} = \pi_{0i} + \pi_{1i} age + \pi_{2i} age^2 + \pi_{3i} age^3 + \varepsilon_{ij}$$
$$\pi_{0i} = \beta_{00} + \beta_{01} intervention + \zeta_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11} intervention + \zeta_{1i}$$
$$\pi_{2i} = \beta_{20} + \beta_{21} intervention + \zeta_{2i}$$
$$\pi_{3i} = \beta_{30} + \beta_{31} \text{intervention}$$

Les paramètres $\beta_{01}, \beta_{11}, \beta_{21}$ et β_{31} sont ceux représentant l'effet de l'intervention sur l'ordonnée à l'origine et les différents paramètres décrivant l'évolution temporelle. Pour

conclure que le programme d'intervention a eu effectivement un impact sur les trajectoires de violence des garçons, cela suppose que certains de ces paramètres doivent être significativement différents de zéro. Des tests unicaudaux seront utilisés lors de l'estimation de l'effet de l'intervention.

L'intérêt de garder dans le modèle les termes d'interaction entre le temps et l'appartenance au groupe a ensuite été testé. Ces termes d'interaction permettent d'identifier si les trajectoires au sein des groupes intervention et contrôle sont similaires. Les termes d'interaction ont tous été insérés dans un premier modèle décrit par l'équation 8, puis β_{31}, β_{21} et β_{11} ont été retirés un à un. Les résultats sont présentés au tableau VII.

Tableau VII : Modélisation des courbes de croissance

	Modèle 1 (BIC=-185,0)		**Modèle 2 (BIC=-189,8)**		**Modèle 3 (BIC=-194,7)**		**Modèle 4 (BIC=-199,6)**	
Effets	**Estimé**	**Err type**	**Estimé**	**Err type**	**Estimé**	**Err type**	**Estimé**	**Err type**
Effets fixes								
Ordonnée	2,141***	0,041	2,137***	0,040	2,141***	0,038	2,142***	0,034
age	-0,042	0,045	-0,031	0,031	-0,036	0,024	-0,037	0,023
age2	0,024	0,017	0,019**	0,009	0,020**	0,009	0,020**	0,009
age3	-0,003*	0,002	-0,003***	0,001	-0,003***	0,001	-0,003***	0,001
intervention	-0,098**	0,048	-0,102**	0,046	-0,098**	0,043	-0,100**	0,037
age*intervention	-0,007	0,052	0,008	-0,028	0,001	-0,010	--	--
age2*intervention	0,005	0,020	-0,001	-0,004	--	--	--	--
age3*intervention	-0,001	0,002	--	--	--	--	--	--
Effets aléatoires								
Ordonnée	0,035***	0,007	0,035***	0,007	0,035***	0,007	0,035***	0,007
age	0,009***	0,003	0,009***	0,003	0,009***	0,003	0,009***	0,003
age2	0,0001**	0,000	0,0001**	0,0001	0,0001**	0,0001	0,0001**	0,0001
Résidu	0,027***	0,002	0,027***	0,002	0,027***	0,002	0,027***	0,002

*p<0,1; **p<0,05; ***p<0,01; N=146.

Les résultats de l'analyse des interactions entre le temps et le groupe indiquent que les coefficients décrivant les courbes de chacun des groupes ne se distinguent pas d'un groupe à l'autre. Cela dit, les termes d'interaction n'ont pas été repris dans les analyses subséquentes. L'équation utilisée pour tester l'effet de l'intervention est la suivante :

Équation 9 : Modèle d'analyse visant à estimer l'effet de l'intervention sur les courbes de croissance de la violence

$$y_{ij} = \pi_{0i} + \pi_{1i} age + \pi_{2i} age^2 + \pi_{3i} age^3 + \varepsilon_{ij}$$
$$\pi_{0i} = \beta_{00} + \beta_{01} intervention + \zeta_{0i}$$
$$\pi_{1i} = \beta_{10} + \zeta_{1i}$$
$$\pi_{2i} = \beta_{20} + \zeta_{2i}$$
$$\pi_{3i} = \beta_{30}$$

Les hypothèses sur l'effet de l'intervention testées dans le cadre des analyses exécutées sont formulées de la manière suivante :

H_0 : Le groupe contrôle et le groupe intervention ne se distinguent pas significativement au niveau de la violence initiale (l'ordonnée à l'origine); β_{01} n'est pas significativement différent de zéro.

H_1 : Le groupe intervention présente un niveau de violence initial significativement plus faible que le groupe contrôle; β_{01} est négatif et significativement différent de zéro.

6.2. Hypothèses sur le mécanisme des données manquantes

Rappelons que cinq hypothèses *MAR* sur le mécanisme des données manquantes ont été testées (voir tableau V, p.50). Les résultats de l'identification de covariables associées à la probabilité qu'une donnée soit manquante sont d'abord présentés. Ensuite, la section 6.2.1 contient les résultats liés à l'hypothèse que le mécanisme *MAR* est différent selon le schéma de données manquantes. Les analyses menées spécifiquement auprès des schémas monotone et intermittent sont également dans cette section. Ces dernières visent à répondre

aux interrogations classiques de l'analyse de l'attrition. Premièrement, est-ce que le taux d'attrition est différent entre les groupes, et deuxièmement, est-ce que les individus qui quittent l'étude ont des valeurs différentes sur les covariables mesurées avant l'intervention par rapport à ceux qui restent jusqu'à la fin de l'étude? C'est dans la section suivante que sont traités les résultats de l'hypothèse que le mécanisme *MAR* diffère selon le groupe. La section 6.2.3 présente les résultats de l'hypothèse selon laquelle le mécanisme *MAR* est unique pour l'ensemble de l'échantillon. Les variables identifiées sont finalement utilisées dans différents modèles d'imputation définis au sein de la procédure d'imputation multiple, ce qui est présenté à la section 6.2.4.

6.2.1. Mécanisme distinct par schéma de données manquantes

La procédure PROC MI de SAS, avec l'option NIMPUTE=0, permet d'obtenir une description des schémas des données manquantes. Pour les données de l'ELEM, la procédure a trouvé 30 schémas différents dont le nombre de participants varie entre 1 et 20 (en excluant le schéma où les données sont complètes, N=146). La faible taille de chacun des schémas ne permet pas d'appliquer directement la méthode de la comparaison de moyenne ou de la probabilité de la présence d'un certain schéma. La méthode suggérée par Yang (2005) pour analyser les schémas permet de distinguer un schéma monotone et un schéma intermittent. Les résultats de l'analyse de ces schémas sont présentés dans les sections suivantes.

6.2.1.1. Schéma monotone (attrition)

Les analyses présentées ici portent uniquement sur les cas créant un schéma monotone, qui contient aussi les cas complets et portant le nombre de cas disponible pour ces analyses à 209. Elles sont présentées au tableau VIII.

Tableau VIII : Mécanisme des données manquantes – schéma monotone (N=209)

r_{ij}	**Variables associées à la probabilité qu'une donnée soit manquante (ref : r_{ij} =0)**
11	---
12	---
13	---
14	---
15	---
16	Hyperactivité (OR : 1,357)
17	---

* Le niveau de signification utilisé est $p<0,1$

Ces régressions logistiques montrent que l'hyperactivité serait associée à la probabilité qu'une donnée soit manquante dans le schéma monotone. Ainsi, un haut niveau d'hyperactivité est associé à une probabilité plus élevée qu'une donnée soit manquante à 16 ans.

Une analyse spécifique du schéma monotone a également été effectuée afin de répondre aux questions ayant été soulevées par l'analyse classique de l'attrition. Est-ce que le taux d'attrition est différent entre les groupes? Est-ce que les individus qui quittent l'étude ont des valeurs différentes sur les covariables mesurées avant l'intervention par rapport à ceux qui restent jusqu'à la fin de l'étude? Advenant des différences entre les participants et les non participants, est-ce que ces différences sont les mêmes à travers les groupes?

Pour déterminer si le taux d'attrition est différent entre les groupes, une régression logistique bivariée a été effectuée à chaque temps de mesure. Les régressions utilisent la variable dépendante binaire indiquant si la donnée est présente (r_{ij}=0) ou absente (r_{ij}=1) (référence est r_{ij}=0) et comme variable indépendante, celle indiquant l'appartenance au groupe. De cette manière, un coefficient associé à la variable intervention significativement différent de zéro indiquerait une probabilité plus élevée qu'une donnée soit manquante dans l'un ou l'autre groupe. Les résultats présentés dans le tableau IX montrent que la probabilité qu'une donnée soit manquante n'est pas différente entre les groupes, et ce, pour chaque temps de mesure.

Tableau IX : Régressions logistiques bivariées indiquant la probabilité qu'une donnée soit manquante (réf: r_{ij}=0)

Variables r_{ij}	***Paramètres***	***β***	***Err.-Type***	***p***
Violence à 11 ans	Ordonnée	-2,322	0,469	<0,0001
	intervention	-0,103	0,542	0,8490
Violence à 12 ans	Ordonnée	-2,322	0,469	-2,3224
	intervention	-0,538	0,522	0,5382
Violence à 13 ans	Ordonnée	-1,946	0,404	<0,0001
	intervention	-0,360	0,458	0,4322
Violence à 14 ans	Ordonnée	-1,653	0,364	<0,0001
	intervention	-0,283	0,416	0,4963
Violence à 15 ans	Ordonnée	-1,299	0,326	<0,0001
	intervention	0,031	0,382	0,9355
Violence à 16 ans	Ordonnée	-1,196	0,317	0,0002
	intervention	-0,089	0,368	0,8089
Violence à 17 ans	Ordonnée	-0,830	0,291	0,0043
	intervention	0,014	0,340	0,9675

La seconde question cherche à déterminer si les individus ayant quitté l'étude ont des valeurs différentes sur les covariables mesurées avant l'intervention par rapport à ceux qui sont restés. Advenant des différences, est-ce qu'elles sont les mêmes selon le groupe?

Une stratégie utilisant la régression multiple a été appliquée. (Graham & Donaldson, 1993). Sept régressions multiples ont été effectuées, soit une pour chaque covariable. Une variable binaire indiquant si un individu a quitté l'étude (battrition=1) ou non (battrition=0) a été créée. Les variables indépendantes incluses sont donc battrition et l'interaction entre battrition et intervention. Le paramètre associé à battrition permet de déterminer si les individus ayant quitté l'étude ont des valeurs différentes sur les variables mesurées avant

l'intervention de ceux étant restés jusqu'à la fin. Le paramètre associé à l'interaction entre battrition et intervention permet quant à lui de déterminer si les différences sont les mêmes à travers les groupes contrôle et intervention. Les résultats sont présentés au tableau X.

Tableau X : Régressions multiples évaluant si les individus ayant quitté l'étude ont des valeurs différentes sur les variables pré-test

Variables dépendantes	***Paramètres***	***β***	***Err.-Type***	***p***
Adversité familiale	Ordonnée	0,417	0,020	<0,0001
	battrition	-0,011	0,062	0,8575
	battrition*intervention	0,011	0,068	0,8686
Prosocialité	Ordonnée	6,918	0,389	<0,0001
	battrition	0,906	1,203	0,4525
	battrition*intervention	1,432	1,333	0,2838
Agressivité physique	Ordonnée	3,356	0,129	<0,0001
	battrition	0,350	0,399	0,3812
	battrition*intervention	0,293	0,441	0,5078
Opposition	Ordonnée	5,486	0,176	<0,0001
	battrition	0,161	0,544	0,7678
	battrition*intervention	0,017	0,602	0,9780
Inattention	Ordonnée	2,228	0,113	<0,0001
	battrition	0,022	0,359	0,9503
	battrition*intervention	0,050	0,397	0,8999
Hyperactivité	Ordonnée	2,701	0,103	<0,0001
	battrition	0,0633	0,318	0,8425
	battrition*intervention	-0,170	0,352	0,6296

Dans les sept régressions, aucune des variables indépendantes n'est significative.

Cette stratégie ne tient pas compte des moments où chaque participant a quitté l'étude. Il aurait d'ailleurs été intéressant d'appliquer l'analyse de survie qui comporte cet avantage. En conclusion, il appert que les individus qui on quitté l'étude avant sa fin ne sont pas différents de ceux étant restés jusqu'à la fin. Ceci reste vrai à l'intérieur des deux groupes.

6.2.1.2. Schéma intermittent

Les résultats de l'analyse du mécanisme des données manquantes au sein du schéma intermittent sont présentés.

Tableau XI : Mécanisme des données manquantes – schéma intermittent (N=41)

r_{ij}	**Variables associées à la probabilité qu'une donnée soit manquante (ref : r_{ij} =0)**
11	Agressivité physique (OR : 0,363)
12	---
13	---
14	Adversité familiale (OR : 0,034), Opposition (OR : 0,520), Hyperactivité (OR : 2,516)
15	---
16	Prosocialité (OR : 0,769), Inattention (OR : 2,852) et Hyperactivité (OR : 0,163)
17	---

* Le niveau de signification utilisé est $p<0,1$

Les résultats des régressions logistiques menées au sein du schéma intermittent ne montrent aucune tendance claire. Toutes les covariables sont significatives à un temps de mesure ou l'autre. En outre, les résultats concernant l'hyperactivité vont dans deux directions différentes. Un score élevé d'hyperactivité est associé à une probabilité plus élevée d'avoir une donnée manquante à 14 ans, alors que la relation est inverse sur la probabilité d'avoir une donnée manquante à 16 ans. Comme le suggère Yang (2005), le mécanisme associé à ce schéma semble ignorable puisqu'aucune tendance claire ne semble émerger.

Somme toute, l'analyse du schéma intermittent n'a pas montré de différences entre les groupes. Enfin, l'analyse du schéma monotone indique que le taux d'attrition est similaire entre les groupes et que les garçons qui quittent l'étude n'ont pas de valeurs différentes sur les variables mesurées avant l'intervention par rapport à ceux qui ont des données complètes. Il ne semble pas y avoir de raison de croire que les schémas de données manquantes affectent la validité interne de l'étude.

6.2.2. Mécanisme distinct par groupe

6.2.2.1. Groupe contrôle

Le tableau XII présente les résultats de l'analyse du mécanisme des données manquantes au sein du groupe contrôle.

Tableau XII : Mécanisme des données manquantes – groupe contrôle (N=181)

r_{ij}	Variables associées à la probabilité qu'une donnée soit manquante (ref : r_{ij} =0)
11	Hyperactivité (OR : 1,587)
12	---
13	---
14	Hyperactivité (OR : 1,421)
15	---
16	---
17	Hyperactivité (OR : 1,436) et Inattention (OR : 0,765)

* Le niveau de signification utilisé est p<0,1

Les résultats au sein du groupe contrôle permettent d'affirmer qu'un haut niveau d'hyperactivité est associé à une probabilité plus élevée qu'une donnée soit manquante pour trois des sept temps de mesure, alors que la relation est inverse pour l'inattention.

6.2.2.2. Groupe intervention

Ensuite, le tableau XIII présente les résultats de l'analyse du mécanisme des données manquantes au sein du groupe intervention.

Tableau XIII : Mécanisme des données manquantes – groupe intervention (N=69)

r_{ij}	Variables associées à la probabilité qu'une donnée soit manquante (ref : r_{ij}=0)
11	---
12	---
13	---
14	Inattention (OR : 0,589)
15	---
16	---
17	---

* Le niveau de signification utilisé est p<0,1

Les résultats au sein du groupe intervention suggèrent que seule l'inattention est associée à la probabilité qu'une donnée soit manquante. Ceci dit, un haut niveau d'inattention est associé à une plus faible probabilité qu'une donnée soit manquante. Le fait que l'hyperactivité ne soit pas associée à la probabilité qu'une donnée soit manquante suggère donc que le mécanisme des données manquantes serait différent dans les deux groupes.

6.2.3. Mécanisme unique pour l'ensemble de l'échantillon

Le tableau XIV présente enfin les résultats de l'analyse du mécanisme des données manquantes au sein de l'échantillon complet.

Tableau XIV : Mécanisme des données manquantes – échantillon complet (N=250)

r_{ij}	**Variables associées à la probabilité qu'une donnée soit manquante (ref : r_{ij} =0)**
11	---
12	---
13	Hyperactivité (OR :1,340)
14	Hyperactivité (OR :1,488)
15	Hyperactivité (OR :1,380) et Inattention (OR :0,803)
16	---
17	Hyperactivité (OR :1,251)

* Le niveau de signification utilisé est $p<0,1$

À la lumière de ces analyses, il est possible d'affirmer que plus le niveau d'hyperactivité augmente, plus les probabilités sont élevées qu'une donnée soit manquante pour quatre des sept temps de mesure. L'effet est par contre inverse pour l'inattention : plus le niveau d'inattention augmente, plus les probabilités qu'une donnée soit manquante sont faibles. L'inattention est reliée à la variable binaire r_{ij} pour seulement un temps de mesure. Ces deux variables, l'hyperactivité et l'inattention, sont alors considérées comme associée à la cause de l'absence de données pour les analyses utilisant l'échantillon complet.

Les analyses des mécanismes des données manquantes ont permis de trouver des variables qui seraient liées à la cause de l'absence de données. Ces variables permettent de

poser différentes hypothèses sur le mécanisme *MAR*. La prochaine section porte sur la comparaison des résultats de différentes estimations faites sur l'effet de l'intervention sur les trajectoires de violence observées chez les garçons.

6.2.4. Comparaison des hypothèses sur le mécanisme des données manquantes

Les résultats présentés à la section précédente ont permis de trouver des covariables associées aux cinq mécanismes *MAR* qui ont été postulés. Le tableau XV résume les différences entre ces versions de *MAR* et le tableau XVI contient les résultats des estimations faites sur l'effet de l'intervention. Ceux-ci ont été obtenus en appliquant la procédure d'imputation multiple sous chacune de ces cinq hypothèses.

Tableau XV : Hypothèses *MAR* sur le mécanisme des données manquantes

Modèle	**Ensemble de données**	**Covariables postulées comme associées aux données manquantes**
B[2]	Schéma intermittent (N=187)	Base[1] + Toutes les covariables
	Schéma monotone (N=209)	Base[1] + Hyperactivité
C	Groupe contrôle (N=181)	Base[1] + Hyperactivité et inattention
	Groupe intervention (N=69)	Base[1] + Inattention
D	Toutes les données (N=250)	Base[1]
E	Toutes les données (N=250)	Base[1] + Hyperactivité et inattention
F	Toutes les données (N=250)	Base[1] + Toutes les covariables

[1] Le terme « Base » indique tous les temps de mesures de la violence.

[2] Pour l'analyse des schémas monotone et intermittent, les 146 cas complets ont été laissés dans chaque matrice de données.

Tableau XVI : Comparaison d'hypothèses sur le mécanisme des données manquantes

Effet	*Hypothèse MAR*	*Modèle*	*β*	*Err.-Type*	*p*
Ordonnée	Mécanisme distinct par schéma	B	2,192	0,028	<0,0001
	Mécanisme distinct par groupe	C	2,188	0,027	<0,0001
	Mécanisme unique – Aucune covariable	D	2,191	0,027	<0,0001
	Mécanisme unique – Hyp. et In.	E	2,197	0,028	<0,0001
	Mécanisme unique – Toutes les covariables	F	2,197	0,032	<0,0001
Age	Mécanisme distinct par schéma	B	-0,047	0,021	0,0161
	Mécanisme distinct par groupe	C	-0,043	0,019	0,0106
	Mécanisme unique – Aucune covariable	D	-0,046	0,019	0,0094
	Mécanisme unique – Hyp. et In.	E	-0,048	0,022	0,0167
	Mécanisme unique – Toutes les covariables	F	-0,047	0,019	0,0060
Age2	Mécanisme distinct par schéma	B	0,026	0,009	0,0040
	Mécanisme distinct par groupe	C	0,024	0,007	0,0003
	Mécanisme unique – Aucune covariable	D	0,027	0,009	0,0018
	Mécanisme unique – Hyp. et In.	E	0,026	0,009	0,0027
	Mécanisme unique – Toutes les covariables	F	0,026	0,007	0,0003
Age3	Mécanisme distinct par schéma	B	-0,003	0,001	0,0015
	Mécanisme distinct par groupe	C	-0,003	0,001	<0,0001
	Mécanisme unique – Aucune covariable	D	-0,004	0,001	0,0012
	Mécanisme unique – Hyp. et In.	E	-0,003	0,001	0,0007
	Mécanisme unique – Toutes les covariables	F	-0,003	0,001	<0,0001
Intervention	Mécanisme distinct par schéma	B	-0,052	0,032	0,0520
	Mécanisme distinct par groupe	C	-0,045	0,029	0,0649
	Mécanisme unique – Aucune covariable	D	-0,044	0,029	0,0683
	Mécanisme unique – Hyp. et In.	E	-0,041	0,030	0,0852
	Mécanisme unique – Toutes les covariables	F	-0,041	0,033	0,1088

Les résultats des estimations faites sur l'effet de l'intervention présentent peu de variabilité. L'hypothèse que le mécanisme des données manquantes est distinct selon le schéma des données manquantes produit un résultat à la limite du niveau de signification généralement utilisé en sciences sociales, soit $p \leq 0{,}05$. Toutes les autres hypothèses produisent des résultats non significatifs.

Les résultats de l'estimation de l'effet du programme en postulant que le mécanisme est différent selon le groupe ou le schéma montrent qu'une très faible variation du modèle

d'imputation peut affecter les estimations. En effet, la seule différence entre les modèles C et E, présentés au tableau XVI, est que lors de l'imputation distincte par groupe seule l'inattention est incluse dans le modèle d'imputation pour le groupe intervention. Ainsi, le fait de modéliser séparément le mécanisme des données manquantes pour les 69 garçons du groupe intervention a fait passer l'estimation de l'effet du programme de β=0,041 à β=0,045 et la signification de p=0,085 à p=0,065. Par contre, lors de l'imputation distincte par schéma, le mécanisme des données manquantes du schéma monotone a été défini comme étant dépendant uniquement sur l'hyperactivité. De plus, toutes les covariables ont été laissées dans le modèle d'imputation pour le schéma intermittent. En comparant les modèles B et F, nous pouvons constater que l'estimation portant sur l'effet du programme est passée de β=0,040 à β=0,052 et la signification de p=0,109 à p=0,052. En définissant séparément le mécanisme des données manquantes pour les 63 garçons formant un schéma monotone, l'effet du programme est passé de non significatif à significatif.

En terminant, il appert que l'erreur type croit avec l'augmentation du nombre de variables postulées comme associées au mécanisme des données manquantes, et ce, en comparant les modèles D, E et F. Le modèle D, où aucune covariable n'était incluse dans le modèle d'imputation, a généré une erreur type de 0,029. Le modèle E, où l'hyperactivité et l'inattention ont été ajoutées, a produit une erreur type de 0,030. Enfin, le modèle F, où toutes les covariables étaient incluse comme cause possible des données manquantes, présente une erreur type de 0,033. Le niveau de signification est ainsi passé de p=0,07 pour le modèle D à p=0,11 pour le modèle F.

6.3. Comparaison des méthodes d'estimation de l'effet d'une intervention avec données manquantes

Outre l'hypothèse sur le mécanisme des données manquantes, différentes méthodes d'estimation peuvent être utilisées pour évaluer l'effet d'une intervention en présence de données manquantes. Certaines méthodes d'estimation utilisées pour déterminer l'effet

d'une intervention, comme la pondération et l'imputation multiple, permettent de spécifier un certain type de mécanisme *MAR*. D'autres méthodes supposent un mécanisme qu'il n'est pas possible de définir, par exemple l'analyse des cas complets ou le maximum de vraisemblance. La méthode d'analyse des cas complets suppose que le mécanisme est *MCAR*. Quant au maximum de vraisemblance, il implique que le mécanisme est *MAR* et qu'il n'est pas associé à des covariables. La comparaison des méthodes d'estimation doit donc être effectuée en considérant les spécificités de chacune d'entre elles.

Le tableau XVII contient les résultats de l'analyse des cas complets, du maximum de vraisemblance, de l'imputation multiple ainsi que de la pondération. Pour l'imputation multiple et la pondération, un mécanisme unique est postulé dépendant de l'hyperactivité et de l'inattention.

L'analyse des cas complets, postulant *MCAR*, produit les résultats qui sont les plus différents par rapport à ceux obtenus par n'importe quelles autres méthodes. Il s'agit de l'analyse menant à la conclusion la plus franche, montrant que le programme d'intervention a eu un impact significatif (β=-0,08; p=0,01). Puisque certaines covariables ont pu être identifiées comme associées au mécanisme des données manquantes, le postulat *MCAR* ne devrait pas être accepté sans en évaluer la sensibilité.

En postulant un mécanisme *MAR*, les résultats obtenus par maximum de vraisemblance, par pondération ou après imputation multiple sont très similaires, et ce, peu importe le bloc de covariables postulé comme associé au mécanisme. L'effet du programme d'intervention est non significatif dans tous les cas (β=-0,04; p>0,05).

Tableau XVII : Comparaison de quatre méthodes d'estimation en présence de données manquantes

Effet	*Méthode d'estimation*	*Modèle*	*β*	*Err.-Type*	*p*
Ordonnée	Cas complets	A	2,158	0,034	<0,0001
	Imputation multiple	E	2,197	0,028	<,0001
	Maximum de vraisemblance	G	2,195	0,028	<0,0001
	Pondération	H	2,196	0,028	<,0001
Age	Cas complets	A	-0,029	0,023	0,2068
	Imputation multiple	E	-0,048	0,022	0,0167
	Maximum de vraisemblance	G	-0,046	0,019	0,0185
	Pondération	H	-0,047	0,020	0,0180
Age2	Cas complets	A	0,017	0,009	0,0531
	Imputation multiple	E	0,026	0,009	0,0027
	Maximum de vraisemblance	G	0,025	0,008	0,0010
	Pondération	H	0,025	0,008	0,0011
Age3	Cas complets	A	-0,002	0,001	0,0100
	Imputation multiple	E	-0,003	0,001	0,0007
	Maximum de vraisemblance	G	-0,003	0,001	<0,0001
	Pondération	H	-0,003	0,001	<0,0001
Intervention	Cas complets	A	-0,082	0,037	0,0142
	Imputation multiple	E	-0,041	0,030	0,0852
	Maximum de vraisemblance	G	-0,043	0,030	0,0776
	Pondération	H	-0,043	0,030	0,0768

Chapitre 7 : Discussion

Il est impossible d'identifier hors de tout doute le mécanisme des données manquantes. Pour contourner le problème, différentes hypothèses sur la cause relative à l'absence des données ont été comparées. Cette approche est nommée analyse de sensibilité. Spécifiquement, nous avons premièrement cherché à identifier des covariables associées à la probabilité qu'une donnée soit manquante. L'identification de ces covariables a permis de postuler différentes versions d'un mécanisme *MAR*. Rappelons que le mécanisme *MAR* stipule que la probabilité qu'une donnée soit manquante est associée à des covariables, mais pas aux données de la variable avec valeurs manquantes.

Une analyse du mécanisme a été exécutée pour l'échantillon complet, pour chaque groupe puis pour chaque schéma. Les covariables associées à la cause relative à l'absence des données ont été utilisées pour comparer différentes hypothèses sur le mécanisme des données manquantes.

Les analyses étaient centrées autour de deux volets. Un premier volet visait à comparer différentes hypothèses sur le mécanisme des données manquantes à partir de la même méthode d'analyse. Le second volet visait à comparer différentes méthodes d'estimation en présence de données manquantes. Les analyses exposées ont montré que les données manquantes peuvent avoir un impact important sur les conclusions d'une analyse.

Considérant son utilisation directe, le maximum de vraisemblance semble présenter un avantage indéniable sur les autres méthodes. Il est toutefois impossible de poser différentes hypothèses sur le mécanisme des données manquantes en utilisant cette méthode. Son principal avantage sur l'utilisation des cas complets réside dans le fait que la méthode considère toutes les données observées.

La pondération par l'inverse de la probabilité qu'une donnée soit manquante peut être située dans la famille des scores de propension. Cette méthode présente des avantages certains sur les autres puisqu'en plus d'utiliser toute l'information disponible, comme le maximum de vraisemblance, elle permet de séparer les étapes de traitement des données manquantes et d'analyse des données, contrairement à l'imputation multiple. Une fois les

poids produits, un fichier de données peut être transféré à un analyste qui pourra effectuer des analyses d'une manière habituelle. Cette caractéristique peut rendre la méthode très attrayante pour de nombreux contextes de recherche.

Nous avons également pu observer que le postulat le plus inclusif sur le mécanisme des données manquantes (c.-à-d. modèle F : imputation multiple avec toutes les covariables incluses dans le modèle d'imputation) a généré les résultats les moins significatifs. Bien qu'il nous soit impossible à l'affirmer hors de tout doute à partir des données analysées, il semble que l'inclusion de variables non associées au mécanisme des données manquantes dans le modèle d'imputation augmente inutilement la valeur de l'erreur type. En outre, la spécification du mécanisme par groupe et par schéma semble générer de plus grandes différences dans les paramètres que celles observées entre le maximum de vraisemblance, la pondération et l'imputation multiple. Il appert que la modélisation du mécanisme des données manquantes ait un impact important sur l'estimation de l'effet d'une intervention.

De plus, l'imputation distincte par schéma produit la plus forte estimation de l'effet du programme de toutes les analyses menées en traitant les données manquantes. Il s'agit de la seule estimation de l'effet de l'intervention qui ait identifié une différence statistiquement significative entre les groupes contrôle et intervention.

Il semble que le fait de préciser le mécanisme des données manquantes pour différents sous-groupes de l'échantillon ait une influence importante sur l'estimation de l'effet de l'intervention. Rappelons que le mécanisme est défini, pour une variable donnée, comme la probabilité qu'une donnée soit manquante, conditionnelle à différentes covariables et/ou aux valeurs de la variable elle-même. Ainsi, tout comme la littérature le suggère, les données manquantes de certains sous-groupes de l'échantillon semblent être mieux qualifiées par des mécanismes distincts. Ceci implique de conditionner le modèle d'imputation sur des ensembles distincts de variables. Parmi les méthodes d'analyse utilisées, la pondération et l'imputation multiple sont les seules à permettre une certaine forme de modélisation du mécanisme.

Il a été démontré que pour qualifier les données manquantes de l'ELEM, le mécanisme *MCAR* était possible, mais peu probable. Par conséquent l'analyse des cas complets comporte des possibilités non négligeables de biais. L'estimation de l'effet du programme de prévention, basée sur les cas complets, indique un effet négatif et significatif. La majorité des autres analyses traitant les données manquantes d'une manière ou d'une autre, donc impliquant un postulat *MAR*, produisent une estimation de l'effet du programme plus faible, découlant sur une interprétation d'un effet non significatif du programme.

Les analyses présentées ont été centrées autour de la notion de biais engendré par les données manquantes. Toute analyse de données comportant des valeurs manquantes, que ces valeurs manquantes soient prises en compte ou non, suppose un postulat sur la cause relative à l'absence de ces observations. En effet, par l'analyse des cas complets, postulant que les données manquantes sont *MCAR*, l'estimation de l'effet du programme est significative, alors que toutes les méthodes de traitement des données manquantes présentées mènent à une interprétation plus nuancée.

Les résultats obtenus selon différents postulats sur le mécanisme des données manquantes convergent avec ceux présentés par d'autres auteurs, à savoir (I) que différentes méthodes d'estimation permettent d'arriver à des résultats similaires; (II) que ces méthodes d'estimation sont généralement robustes à différents postulats sur le mécanisme des données manquantes, et (III) que l'analyse des cas complets comporte des chances de biais lorsque la quantité de données manquantes est importante. Analyser les données manquantes et appliquer les méthodes d'analyse qui optimisent l'utilisation de l'information contenue dans les données permet d'augmenter la validité interne d'une étude.

Le fait d'utiliser des données réelles, par opposition à des données simulées, ne nous a pas permis d'établir la performance des méthodes présentée au niveau de la précision de l'estimation d'un paramètre. Cette situation n'est pas unique à notre analyse et la plupart

des chercheurs ayant à traiter des données ne connaissent pas la valeur réelle des paramètres. Mais en s'appuyant sur des analyses antérieures de simulation de données, il semble adéquat d'affirmer que le fait d'appliquer une technique d'analyse de données tenant compte des valeurs manquantes apporte une plus grande précision aux estimations. De plus, la taille de l'échantillon utilisé a limité les possibilités d'analyse du mécanisme des données manquantes. La littérature suggère d'analyser le mécanisme selon chacun des schémas à l'intérieur de chacun des groupes. Lors de l'analyse des données manquantes dans un échantillon de plus forte taille, un mécanisme propre à chaque schéma à l'intérieur de chaque groupe pourrait permettre une estimation plus adéquate de l'effet d'une intervention. En fait, beaucoup d'autres possibilités de regroupement existent et il pourrait être fort adéquat dans certains cas de postuler un mécanisme distinct selon sexe par exemple.

Les nouvelles méthodes d'analyse de données comportant des postulats sur la cause relative à l'absence des données sont de plus en plus accessibles. Bien que certaines méthodes nécessitent plus d'apprentissages pour être appliquées, comme l'imputation multiple, d'autres peuvent être utilisées relativement simplement, comme le maximum de vraisemblance. L'imputation multiple comporte toutefois l'avantage de permettre une certaine forme de modélisation du mécanisme des données manquantes lors de la spécification d'un modèle d'imputation. La littérature sur les données manquantes contient également de nombreuses autres méthodes de spécification du mécanisme des données manquantes. Indiquons par exemple les modèles de sélection de Diggle et Kenward (Diggle & Kenward, 1994).

De nombreux articles ont été publiés depuis les cinq dernières années qui visent à évaluer les possibilités de modélisation du mécanisme des données lors d'analyse statistique. Ces recherches sont généralement publiées dans des revues couvrant le champ de la statistique. Le transfert des connaissances d'un champ à un autre demande du temps, mais il appert indispensable que les chercheurs en sciences sociales puissent utiliser à leur tour les avancées produites dans d'autres champs de recherche.

Bibliographie

Allison, P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological methods & research, 28*(3), 301.

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol, 51*(6), 1173-1182.

Boisjoli, R., Vitaro, F., Lacourse, E., Barker, E. D., & Tremblay, R. E. (2007). Impact and clinical significance of a preventive intervention for disruptive boys: 15-year follow-up. *Br J Psychiatry, 191*, 415-419. doi:191/5/415 [pii]

Carpenter, J., Kenward, M., & Vansteelandt, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A, 169*(3), 571-584.

Carpenter, J. R., & Kenward, M. G. (2008). *Missing data in randomised controled trials - a practical guide*.

D'Agostino, J., & Rubin, D. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association, 95*(451).

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med, 17*(19), 2265-2281.

Diggle, P., & Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 43*(1), 49-93.

Enders, C. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*(5), 713.

Fox, W. (1999). *Statistiques sociales, 3e édition*. (3e^{e} éd.). Québec: De Boeck et Les Presses de l'Université Laval.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review Psychology, 60*.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78*(1), 10.

Hansen, W. B., Collins, L. M., Malotte, C. K., Johnson, C. A., & Fielding, J. E. (1985). Attrition in prevention research. *J Behav Med, 8*(3), 261-275.

Lacourse, E., & al. (2002). A Longitudinal-experimental Approach to Testing Theories of Antisocial Behavior Development. *Development and Psychopathology, 14*(4), 909-924.

Little, R. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association, 90*(431).

Little, R. J., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics, 52*(1), 98-111.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Hoboken, N.J.: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data (2nd edition)*. (2nd[e] éd.). Hoboken, N.J.: Wiley.

McKnight, P. E. (2007). *Missing data : a gentle introduction*. New York: Guilford Press.

Nagin, D. S. (1999). Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach. *Psychological Methods, 4,* 139-157.

Nagin, D. S., & Tremblay, R. E. (2001). Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. *Arch Gen Psychiatry, 58*(4), 389-394.

Pagani, L., Boulerice, B., Tremblay, R. E., & Vitaro, F. (1997). Behavioural development in children of divorce and remarriage. *J Child Psychol Psychiatry, 38*(7), 769-781.

Patterson, G. R. R., J. B.; Jones, R. R. (1975). *A Social Learning Approach to Family Intervention. Vol. 1. Families with Aggressive Children*. Castalia.

Raudenbush, S. W. (2001). Comparing Personal Trajectories and Drawing Causal Inferences from Longitudinal Data. *Annual Review of Psychology, 52*, 501-525.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

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

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7,* 147-177.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *J Abnorm Child Psychol, 19*(3), 285-300.

Tremblay, R. E., McCord, J., Boileau, H., Charlebois, P., Gagnon, C., Le Blanc, M., et al. (1991). Can disruptive boys be helped to become competent? *Psychiatry, 54*(2), 148-161.

Tremblay, R. E., Pihl, R. O., Vitaro, F., & Dobkin, P. L. (1994). Predicting early onset of male antisocial behavior from preschool behavior. *Arch Gen Psychiatry, 51*(9), 732-739.

Vitaro, F., Brendgen, M., & Tremblay, R. (2001). Preventive intervention: Assessing its effects on the trajectories of delinquency and testing for mediational processes. *Applied Developmental Science, 5*(4), 201-213.

Yang, X., & Shoptaw, S. (2005). Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug and Alcohol Dependance, 77*, 12.