IMPUTATION MULTIPLE DES DONNEES MANQUANTES POUR LE TRAITEMENT DE DONNEES D'ENQUETE *

Coralie Lupo [1], **Sophie Le Bouquin** [1], **Virginie Michel** [1], **Pierre Colin** [2] et **Claire Chauvin** [1]

RESUME

Le traitement classique des données manquantes consiste à éliminer les observations incomplètes de l'analyse statistique, limitant ainsi sa puissance statistique. De plus, un biais de sélection peut être introduit si le sous-échantillon analysé n'est plus représentatif de l'échantillon initial. Une autre méthode consiste à attribuer des valeurs à ces données manquantes. Une revue de la méthode d'imputation multiple des données manquantes est ici présentée. Cette méthode permet de conserver toute l'information des données et d'effectuer des analyses sans perte de puissance ni introduction de biais. Une illustration de l'application de cette méthode est proposée au travers d'une étude de recherche de marqueurs de risque de la saisie des carcasses de poulets en abattoir, à partir de données collectées lors d'une enquête épidémiologique.

Mots-clés : Donnée manquante, imputation multiple, poulet, saisie sanitaire, enquête.

SUMMARY

Typical management of missing data consists in excluding incomplete observations from the statistical analysis -listwise deletion- which limits its statistical power. A selection bias can also occur if the sample analyzed is no longer representative of the initial sample. Another approach to handling missing data is reviewed. Multiple imputation of missing data allows to keep all data information and to conduct the analysis without loss of power or introduction of bias. An example of application of this method is given with the analysis of risk markers for chicken carcass condemnation at the slaughterhouse.

Keywords : Missing value, Multiple imputation, Chicken, Sanitary condemnation, Survey.

* Texte de la communication orale présentée lors des Journées AEEMA, 22-23 mai 2008

[1] Agence française de sécurité sanitaire des aliments – site de Ploufragan, BP 53, 22440 Ploufragan, France

[2] Université de Bretagne Occidentale - Technopôle Brest-Iroise, 29280 Plouzané, France

I - INTRODUCTION

En épidémiologie, les données sont souvent collectées lors d'enquêtes dites « sur le terrain ». Et dans le cas des enquêtes d'observation, l'épidémiologiste est très régulièrement confronté au problème des données manquantes. L'exemple le plus classique dans les enquêtes est le refus de participation des personnes sélectionnées. Dans le cas de questionnaires auto-administrés, les gens peuvent oublier de répondre à certaines questions. De même, des enquêteurs professionnels peuvent occasionnellement négliger de poser une des questions. Plus fréquemment, les répondants ne sont pas en mesure d'apporter une réponse, soit parce qu'ils ne savent pas répondre à la question soit parce qu'ils n'ont pas accès aux sources pour y répondre (documents perdus ou illisibles). Ces non-réponses limitent le nombre d'observations utilisable pour chaque variable étudiée. Ainsi, un faible pourcentage de valeurs manquantes pour chaque variable peut rapidement aboutir à un grand nombre d'observations dont l'une des variables au moins comporte une valeur manquante. Ces observations sont appelées des observations incomplètes. L'information contenue dans les observations incomplètes est alors fréquemment perdue pour l'analyse statistique car, souvent, les logiciels ne permettent pas de manipuler les données manquantes : ils ne travaillent que sur des observations complètes. Le jeu de données récolté peut ainsi être diminué de façon considérable, se traduisant par une perte de puissance et de précision des analyses statistiques (modélisation ou estimation). Un biais de sélection dans le traitement des données peut également être introduit si le mécanisme conduisant aux données manquantes n'est pas complètement aléatoire, c'est-à-dire si le sous-échantillon d'observations complètes analysé n'est pas représentatif de l'échantillon initial.

L'imputation des données manquantes a été envisagée depuis plus de 50 ans comme traitement de la non-réponse. Imputer c'est combler les valeurs manquantes par des valeurs prédites ou simulées. Cette technique permet de conserver toute l'information des données et d'effectuer des analyses valides en utilisant des logiciels standards. Plusieurs méthodes ont été développées, distinguant l'imputation simple de l'imputation multiple. Les méthodes courantes d'imputation simple sont le remplacement de la valeur manquante par la moyenne, par une valeur observée tirée au hasard ou encore par la réponse issue d'un modèle de régression. L'imputation multiple se définit par la création de plusieurs valeurs plausibles d'une donnée manquante. Elle permet de prendre en compte l'incertitude de prédiction des valeurs manquantes. Ce présent article a choisi de s'intéresser à l'imputation multiple car son utilisation dans le contexte des enquêtes est encore relativement rare.

Après une présentation synthétique de la méthode d'imputation multiple des données manquantes, son application lors de la recherche de marqueurs de risque de la saisie sanitaire des carcasses de poulets en abattoir à partir de données collectées lors d'une enquête épidémiologique est présentée. Plusieurs outils informatiques ont été développés pour réaliser l'imputation multiple des données manquantes. Le logiciel SAS [SAS, 2007] a été utilisé dans la présente illustration.

II - L'IMPUTATION MULTIPLE DES DONNEES MANQUANTES

1. OBJECTIF

L'objectif de l'imputation multiple est de refléter correctement l'incertitude des données manquantes, sans altérer les relations importantes entre variables ou leur distribution. Il ne s'agit pas de prédire les données manquantes avec la plus grande précision. Cette méthode remplace chaque valeur manquante par un jeu de valeurs plausibles, qui représente l'incertitude de la réelle valeur à imputer.

L'épidémiologiste peut alors appliquer à son jeu de données complété des procédures d'analyses statistiques standard, qui utilisent des observations complètes.

2. PRINCIPE

Développée par Rubin dès 1976, la théorie de l'imputation multiple des données manquantes est issue d'une stratégie d'analyse Bayésienne. Il s'agit de définir 1) un modèle d'imputation pour les variables comportant des valeurs manquantes, lié aux variables observées intégralement et 2) un modèle du mécanisme conduisant à la non-réponse. Le modèle d'imputation, donne lieu à une distribution prédictive *a posteriori* des valeurs manquantes, liée aux valeurs observées. Les imputations sont un échantillon aléatoire de valeurs extrait de cette distribution *a posteriori*.

L'imputation est donc avant tout un exercice de modélisation pour décrire la distribution de la variable d'intérêt Y conditionnelle à des variables auxiliaires X, et pour générer ensuite les imputations. Bien que cette méthode mène à la création d'un jeu de données complété, les estimateurs ne seront valides que si les hypothèses sous-jacentes au mécanisme qui aboutit à la non-réponse et au modèle d'imputation sont satisfaites.

3. HYPOTHESES

Souvent la structure des données manquantes dépend de la variable Y considérée. On dira que le mécanisme des données manquantes pour la variable Y est de type uniforme (*Missing completely at random*), si le fait de ne pas avoir de réponse est totalement indépendant de la valeur de Y elle-même ou de n'importe quelle autre variable auxiliaire X du jeu de données. L'échantillon des répondants reste donc une représentation non-biaisée de la population d'origine. Si le mécanisme des données manquantes n'est pas uniforme, il s'agit de savoir si les différences entre les caractéristiques des non-répondants et des répondants peuvent être expliquées par des variables communes aux répondants et aux non-répondants. Le mécanisme est dit « ignorable » (*Missing at random*), si la probabilité de réponse sur la variable Y peut dépendre d'autres variables auxiliaires X, mais pas de la valeur de Y elle-même. Ces deux mécanismes, uniforme et « ignorable », impliquent que l'épidémiologiste n'a pas besoin de se préoccuper du processus de survenue de la non-réponse dans la spécification du modèle d'imputation. Malheureusement, il n'existe pas de moyen pour tester le caractère « ignorable » du mécanisme de non-réponse. En effet, sans connaître la valeur de la donnée manquante, il est impossible de la comparer aux valeurs de cette donnée chez les répondants, pour vérifier qu'elles ne diffèrent pas systématiquement. En revanche, s'assurer que le mécanisme de non-réponse n'est pas associé aux paramètres à estimer (par exemple la variable d'intérêt à modéliser ou à estimer) peut renforcer l'hypothèse de mécanisme « ignorable ».

Le cas échéant, le mécanisme est dit « non-ignorable » (*Missing not at random*) : la probabilité de réponse sur la variable Y dépend de Y. En présence d'un mécanisme « non-ignorable », il y aura inévitablement un biais dû à la non-réponse que le modèle d'imputation devra prendre en compte. L'élimination de ce biais requiert généralement des techniques sophistiquées et spécifiques à chaque application car une très bonne connaissance *a priori* du mécanisme des données manquantes est nécessaire. Globalement trois stratégies sont décrites. Les modèles par sélection (*Selection models*) décrivent conjointement la probabilité de réponse à une variable donnée et cette variable [Heckman, 1976 ; Little, 1995 ; Verbeke *et al.*, 2000]. Les modèles par mélange (*Mixture pattern models*) classent les variables incomplètes dans plusieurs groupes de non-réponse selon leur distribution et approchent le mécanisme global de non-réponse par une combinaison de ces distributions [Little, 1993 ; Verbeke *et al.*, 2000]. Les modèles à variable partagée (*Shared parameter models*) ajoutent un effet aléatoire commun à une variable et à la probabilité de réponse à cette variable [Wu *et al.*, 1988].

L'adéquation de l'imputation des données manquantes est également cruciale pour aboutir à des estimations valides. Idéalement, le modèle d'imputation devrait être construit pour représenter les caractéristiques spécifiques du jeu de données. Son adéquation avec les données repose sur la disponibilité d'information auxiliaire. On appelle « information auxiliaire » un ensemble de variables disponibles pour toutes les unités échantillonnées ou pour toutes les unités de la population (pas de non-réponse). Cette information auxiliaire servira à construire des valeurs imputées en reflétant les relations entre les variables.

La robustesse du modèle d'imputation est également déterminée par le type, la distribution et la proportion des données manquantes. En général, l'approximation de la distribution prédictive *a posteriori* des données manquantes par une distribution multi-normale est utilisée : chaque variable peut être

représentée par une fonction linéaire de toutes les autres variables. Ce modèle d'imputation multi-normal s'applique à toute variable continue et normalement distribuée, ce qui en pratique n'est pas toujours la situation rencontrée.

4. REALISATION PRATIQUE

La réalisation pratique d'inférence par l'imputation multiple des données manquantes comporte trois étapes [Rubin, 1976] :

- les données manquantes sont imputées *m* fois pour générer *m* jeux de données complétés. Chaque valeur manquante est ainsi remplacée par *m* valeurs simulées, en utilisant un modèle approprié qui incorpore la variabilité appropriée entre les *m* imputations ;
- les *m* jeux de données complétés sont analysés séparément et de façon identique par des analyses statistiques standards ;
- les résultats obtenus sur les *m* jeux de données complétés sont combinés pour produire un résultat inférentiel global.

Quelles que soient les analyses de données complètes utilisées (estimation ou modélisation) la façon de procéder reste la même.

4.1. REALISER L'IMPUTATION

L'imputation multiple crée plusieurs (*m*) jeux de données complétés. Pour un jeu de données comportant peu de données manquantes, l'efficacité relative d'une estimation pour un petit nombre d'imputations *m* est élevée [Rubin, 1987]. L'usage a montré que *m* = 5 est suffisant pour produire une estimation satisfaisante [Allison, 2001].

La stratégie consiste à considérer l'imputation variable par variable, mais en fonction de toutes les variables auxiliaires observées. Le modèle d'imputation doit disposer d'un maximum d'information auxiliaire. Globalement, il est conseillé d'inclure autant de variables auxiliaires qu'il est possible dans le modèle d'imputation [Rubin, 1996]. Notamment, pour limiter l'introduction de biais, il est essentiel d'utiliser la variable dépendante pour imputer les valeurs manquantes des variables explicatives [Schafer, 1997].

Le choix du modèle d'imputation adéquat dépend du type des variables à imputer. Le modèle multi-normal est le plus utilisé. L'écart de la multi-normalité est toléré si le nombre de données manquantes est faible [Schafer, 1997]. Mais les données d'enquête comportent souvent de très nombreuses variables ayant différentes distributions et différents formats. Les conditions d'application du modèle multi-normal ne sont donc pas systématiquement réunies. Dans le cas de variables continues dont la distribution n'est pas normale, une transformation est conseillée pour normaliser les distributions et obtenir des imputations correctes. Dans le cas des variables catégorielles, la création de variables indicatrices est recommandée, pour les imputer comme des variables continues puis les arrondir à 0 ou 1, et finalement retrouver le codage initial de la variable.

4.2. ANALYSER LES DONNEES

Les *m* jeux de données complétés sont ensuite analysés avec les procédures standards qui utilisent des observations complètes. La même analyse est conduite séparément sur chacun des *m* jeux de données complétés. Les *m* résultats sont ensuite combinés pour obtenir une estimation globale, simple moyenne des *m* estimations [Rubin, 1987]. Une mesure de précision (erreur standard) est associée à cette estimation, pour représenter l'incertitude des imputations. Son calcul repose sur la formule de la variance définie par Rubin [1987], qui combine la variabilité inter- et intra-imputations.

4.3. VERIFIER DES HYPOTHESES

Une analyse de sensibilité est recommandée pour vérifier le mécanisme de non-réponse [Molenberghs *et al.*, 2007], car les analyses précédemment décrites sont valides si ce mécanisme est « ignorable ». La vérification de cette hypothèse est pratiquement impossible, car si les répondants ne sont pas différents des non-répondants sur les variables récoltées, qu'en est-il des données non récoltées dans l'enquête ?

Dans le cas d'un mécanisme de non-réponse « non-ignorable », sa prise en compte dans le modèle d'imputation est une autre problématique de recherche à part entière (cf. § II.3.).

5. IMPLEMENTATION AVEC LE LOGICIEL SAS

Le logiciel SAS [SAS, 2007] propose une

procédure d'imputation multiple des données manquantes. Il s'agit de la PROC MI, pour imputer les données manquantes et créer *m* jeux de données complétés. Le modèle d'imputation utilisé est un modèle d'imputation multi-normal. Plusieurs options permettent d'adapter la procédure d'imputation aux variables qui ne satisfont pas les conditions d'application du modèle. Les *m* jeux de données complétés sont ensuite analysés séparément avec toute procédure PROC usuelle. Puis les résultats des *m* modèles sont intégrés dans la PROC MIANALYZE, qui les combine et produit un résultat final et unique selon la formule de Rubin [1987].

Réalisé par des utilisateurs de SAS, un programme IVEware, pour Imputation and Variance Estimation software [Raghunathan *et al.*, 2002], (téléchargeable gratuitement sur le site internet : http://www.isr.umich.edu/src/smp/ive/), autorise des imputations plus adaptées aux données d'enquête que la procédure standard MI [Yu *et al.*, 2007]. L'avantage principal du programme IVEware est que la méthode d'imputation n'est pas restreinte à l'hypothèse de multi-normalité. Les modèles de régression utilisés dépendent du type de variables à imputer [Raghunathan *et al.*, 2001]. Cinq types de variables sont considérés : continue, binaire, catégorielle (plus de deux catégories), comptage, mixte (soit valeur nulle, soit continue). L'utilisateur n'est alors pas obligé de transformer et manipuler les variables de son jeu de données. Il peut également intégrer des limites logiques ou des limites de cohérence au processus d'imputation des valeurs manquantes.

III - APPLICATION A L'ANALYSE DE DONNEES D'ENQUETE

1. NATURE DES DONNEES

Les données sont issues d'une enquête conduite en 2005 par l'Agence française de sécurité sanitaire des aliments pour rechercher des facteurs/marqueurs de risque de saisie sanitaire des carcasses de poulets à l'abattoir.

Les données de l'étude se rapportent à 404 lots de poulets de chair abattus dans les régions Bretagne et Pays de la Loire. Chaque lot a été tiré au sort dans 15 abattoirs participant à l'étude, et suivi depuis son arrivée sur l'aire d'attente jusqu'au résultat de l'inspection sanitaire officielle. Puis, pour les lots des 375 éleveurs ayant accepté de participer à l'enquête, une visite d'élevage rétrospective a été réalisée, afin de recenser les conditions particulières auxquelles les animaux avaient été exposés au cours de leur élevage.

Les données recueillies à l'abattoir concernaient les caractéristiques descriptives du lot (taille du lot, age et poids moyen à l'abattage), le transport jusqu'à l'abattoir (densité pendant le transport, temps et conditions d'attente sur le quai,…), les caractéristiques de l'abattage et le résultat de l'inspection sanitaire officielle (*ante mortem* et *post mortem*) des lots de poulets. Les saisies totales de chacun des lots ont été quantifiées par le pourcentage de carcasses saisies par lot.

En élevage, un questionnaire a permis de collecter des données relatives aux caractéristiques de l'exploitation et du bâtiment d'élevage, aux mesures de biosécurité, aux pratiques de conduite d'élevage, à l'historique sanitaire et zootechnique du lot de poulets, aux conditions d'enlèvement et de ramassage.

2. ANALYSE DE CONDUITE

L'objectif était la détermination de variables marqueurs de risque de saisie des carcasses.

La variable d'intérêt était le pourcentage de saisie sanitaire du lot, calculé en divisant le nombre de carcasses saisies par le nombre de poulets abattus dans ce lot. Une modélisation par une régression de Poisson mixte incorporant un effet aléatoire lié à l'abattoir (procédure GLIMMIX, SAS, 2007) a été conduite pour mesurer l'influence des variables explicatives sur le risque de saisie présenté par un lot abattu dans un abattoir donné.

Seules les variables présentant un faible nombre de données manquantes (moins de 20%) ont été candidates pour l'analyse de facteurs de risque. Après vérification d'une variabilité suffisante de chaque variable catégorielle sur notre échantillon (notamment, chaque classe comportant au moins 10% des effectifs pour les variables qualitatives), une

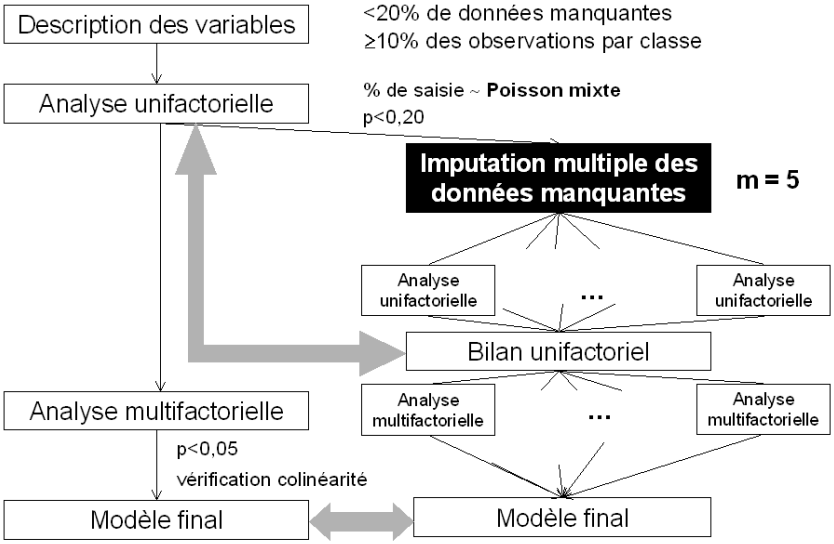
analyse unifactorielle a permis d'identifier les variables les plus associées au pourcentage de saisie, avec un seuil de décision conservateur (test de Wald, p<0,20).

L'imputation multiple des données manquantes des variables sélectionnées à l'issue de l'analyse unifactorielle a été réalisée par le programme IVEware [Raghunathan *et al.*, 2002]. Cinq jeux de données complétés ont ainsi été créés.

La figure 1 illustre les différentes étapes de l'analyse conduite.

Figure 1

Déroulement schématique de l'analyse conduite



L'analyse unifactorielle a été reconduite séparément sur les cinq jeux de données complétés (PROC GLIMMIX), suivie de la procédure MIANALYZE pour combiner les cinq résultats en un résultat final. Ce résultat, obtenu sur les jeux de données complétés a été comparé avec celui du jeu de données initial.

La démarche suivante a été appliquée en parallèle au jeu de données initial, c'est-à-dire uniquement sur les observations complètes, et aux cinq jeux de données complétés, afin de comparer les résultats obtenus.

Un tri a été effectué entre variables potentiellement explicatives afin de ne pas considérer des variables synonymes ou fortement corrélées. En cas de colinéarité avérée (test du Chi², p<0,05), la variable statistiquement la plus associée à la variable d'intérêt a été retenue.

Un modèle de régression multifactorielle a été réalisé selon une procédure descendante pas à pas manuelle. La simplification du modèle sans interaction a éliminé les variables les moins significatives (test de Wald, p<0,05), tout en s'assurant de l'absence de modification majeure des coefficients des variables restantes (<25%). Puis, la simplification du modèle avec des termes d'interaction biologiquement interprétables s'est fondée sur des critères purement statistiques (p<0,05). En particulier, les cinq jeux de données ont été analysés séparément, puis les résultats obtenus combinés par la PROC MIANALYZE, afin d'obtenir un résultat final relatif au jeu de données complété.

Cette modélisation a permis d'estimer un coefficient (e^{β}) équivalent au risque relatif (RR) pour chacun des marqueurs de risque, ajusté sur les autres facteurs du modèle. Ce coefficient représentait l'augmentation proportionnelle du pourcentage de saisie pour un changement d'unité de la variable explicative [Dohoo *et al.*, 2003].

3. RESULTATS

Au total, 404 lots de poulets ont été tirés au sort dans les abattoirs participants, mais les données en élevage n'ont pas été récoltées pour 29 d'entre eux. Ces 29 lots ont été considérés comme les non-répondants à l'étude. Le pourcentage moyen de saisie comme les caractéristiques des lots à l'arrivée à l'abattoir, ne présentaient pas de différence statistiquement significative entre ces 29 lots et les répondants (test de Wilcoxon, $p>0,05$). Ainsi, l'échantillon des répondants semblait être une représentation non-biaisée de la population d'origine. Le mécanisme de non-réponse a été supposé « ignorable » et n'a pas été pris en compte dans la spécification du modèle d'imputation.

Au total, 191 lots (47%) présentaient des données manquantes pour une ou plusieurs des 26 variables explicatives retenues à l'issue de l'analyse unifactorielle. Seulement six variables étaient complètes. Les variables explicatives retenues étaient de format varié : continue (11%), catégorielle binaire (62%), catégorielle à plusieurs classes (19%), comptage (4%), mixte (4%).

L'imputation multiple des données manquantes a été réalisée pour les 20 variables comportant des valeurs manquantes, en utilisant l'information auxiliaire disponible dans les six variables explicatives complètes, la variable d'intérêt (pourcentage de saisie) et la variable structurelle relative à l'abattoir destinataire (tableau 1).

Les résultats de l'analyse unifactorielle réalisée sur les variables complétées étaient comparables (risque relatif et degré de signification p) aux résultats de l'analyse conduite sur les observations complètes.

Le tableau 2 présente le modèle multifactoriel final du risque de saisie pour un lot. La deuxième colonne présente les estimations obtenues en utilisant la méthode d'imputation multiple des données manquantes. La troisième colonne présente les résultats issus du modèle obtenu sur les seules observations complètes, c'est-à-dire après l'élimination des 51 lots incomplets, correspondant à une perte de 13% des observations.

Les résultats des deux modèles étaient très similaires : les six mêmes variables étaient toutes significatives et présentaient des risques relatifs du même ordre de grandeur voire identiques.

IV - DISCUSSION

La gestion des données manquantes est un problème courant pour un épidémiologiste. La prise en compte d'observations incomplètes est notamment intéressante dans les enquêtes analytiques comportant de nombreux facteurs potentiels. La suppression de ces observations incomplètes pour construire un modèle multifactoriel peut entraîner une grande perte de puissance statistique, comme un risque d'introduction d'un biais de sélection.

Dans notre illustration, nous avons observé une grande stabilité des risques relatifs du modèle construit sur les données complétées par rapport aux données initiales. Plusieurs hypothèses peuvent expliquer ces résultats.

Tout d'abord, le fort taux de participation des éleveurs à l'étude (92,8%) a minimisé l'importance attendue des données manquantes. Le modèle multifactoriel obtenu sur les données initiales comportait en effet 353 observations complètes, ce qui ne représente qu'une perte réduite de 13% des observations en comparaison avec la base complétée. L'impact de l'imputation multiple des données manquantes a donc été limité. Par ailleurs, notre échantillon de répondants était comparable à l'échantillon de non-répondants sur les données collectées à l'abattoir. Nous avons donc émis l'hypothèse que le mécanisme de non-réponse pouvait être considéré comme « ignorable » et que les répondants restaient une représentation non-biaisée de notre population initiale. Enfin, seulement trois des six variables du modèle final comportaient des données manquantes : le modèle final était considéré comme robuste. Dans cette illustration, l'application de la méthode de l'imputation multiple des données manquantes a principalement permis d'exploiter l'intégralité de l'information collectée pendant l'enquête, tout en confortant la démarche d'analyse usuelle.

Tableau 1

Variables explicatives et information auxiliaire utilisée pour l'imputation multiple des données manquantes

Description de la variable	Type de variable	Codage	Nombre d'observations
Données collectées à l'abattoir			
Abattoir destinataire du lot	Catégorielle	15 abattoirs	404
% de saisie	Comptage	en %	404
Type de production	Catégorielle	Certifié Lourd Export Standard	404
Poids moyen des oiseaux à l'abattage	Continue	en kg	404
Mortalité pendant le transport	Mixte	en %	404
Chargement des caisses de transport	Continue	en kg/m²	369
Conditions météorologiques d'attente sur le quai	Binaire	Favorables Vent +/- pluie	404
Résultat de l'inspection *ante mortem*	Binaire	Observation de signes cliniques Rien d'anormal	398
Cadence d'abattage usuelle	Binaire	<7000 carcasses/heure ≥7000	404
Nombre de carcasses vues par opérateur de retrait	Binaire	3500 à 5499 carcasses <3500 ou ≥ 5500	339
Positionnement des opérateurs de retrait sur la chaîne d'abattage	Binaire	Sortie plumeuse Sortie plumeuse + éviscération	404
Données collectées en élevage			
Adhésion de l'éleveur à une charte qualité	Binaire	Oui / Non	386
Surface totale des bâtiments dédiée à l'élevage de poulets de chair	Binaire	<1100 m² ≥1100 m²	389
Origine et qualité de l'eau de boisson des oiseaux	Catégorielle	Réseau public éseau privé avec traitement de l'eau adapté Réseau privé sans traitement	366
Acidification de l'eau	Binaire	Oui / Non	384
Désinsectisation du bâtiment	Binaire	Oui / Non	375
Type de bâtiment	Catégorielle	Clair Semi-clair Obscur	382
Accessibilité à l'alimentation	Binaire	<9 mangeoires / 1000 poulets ≥9	363
Nombre de passages quotidiens de l'éleveur au démarrage	Comptage	Nombre de passages	374
Densité à la mise en place	Continue	Nombre de poussins/m^2	388
Homogénéité des poussins à la mise en place	Binaire	Oui / Non	363
% de mortalité au démarrage	Binaire	<0,7% ≥0,7%	364
Tri des animaux	Catégorielle	Non Oui, au démarrage Oui, tout au long du lot	348
Homogénéité des poulets à l'enlèvement	Binaire	Oui / Non	366
% de mortalité cumulé	Binaire	<2,5% ≥2,5%	370
Troubles sanitaires	Catégorielle	Non Anciens (≥1 semaine) Récents (<1 semaine)	376
Survenue d'un stress pendant l'élevage	Binaire	Oui / Non	372
Enlèvement précédent	Binaire	Oui / Non	400

Tableau 2

Comparaison des résultats du modèle multifactoriel final après imputation multiple et sur les cas complets

Variables explicatives	Après imputation multiple (n=404)			Observations complètes (n=353)		
	RR[1]	IC à 95%	P	RR[1]	IC à 95%	p
Type de production						
Certifié	1,71	[1,27-2,31]	0,0004	1,90	[1,41-2,56]	<0,0001
Lourd	1,32	[1,08-1,61]		1,53	[1,25-1,87]	
Export	1,03	[0,82-1,28]		1,06	[0,85-1,32]	
Standard	1[2]			1[2]		
Nombre de passages quotidiens de l'éleveur au démarrage						
	0,93[3]	[0,88-0,99]	0,016	0,90[3]	[0,85-0,96]	0,001
Mortalité cumulée en élevage (en %)						
≥2,5	1,43	[1,23-1,66]	<0,0001	1,43	[1,24-1,65]	<0,0001
<2,5	1[2]			1[2]		
Troubles sanitaires						
Récents	1,29	[1,11-1,50]	0,001	1,29	[1,11-1,51]	0,0001
Anciens	0,99	[0,83-1,17]		0,89	[0,74-1,07]	
Aucun	1[2]			1[2]		
Mortalité pendant le transport (en %)						
	1,40[4]	[1,15-1,71]	0,001	1,37[4]	[1,11-1,69]	0,004
Cadence d'abattage (en carcasses/heure)						
<7000	0,58	[0,45-0,75]	<0,0001	0,57	[0,45-0,72]	<0,0001
≥7000	1[2]			1[2]		

[1] risque relatif ajusté sur les autres variables introduites dans le modèle
[2] classe de référence
[3] pour chaque passage supplémentaire
[4] pour chaque 1% supplémentaire

D'autres publications ont obtenu des modèles différents entre la base initiale et la base complétée. Ainsi, la méthode d'imputation multiple des données manquantes a mis en évidence une tendance à l'augmentation de symptômes urologiques liée à l'âge du patient dans une étude clinique [Taylor *et al.*, 2002], alors que ce résultat n'apparaissait pas lors de l'analyse conduite sur la base de données initiale. Les hommes les plus âgés étaient en effet ceux dont la participation à l'étude était la plus faible et un biais de sélection lié à la non-réponse a été incriminé. A l'inverse, dans le cadre de la recherche de facteurs pronostiques de la signature d'une décharge de l'hôpital pour des hospitalisations psychiatriques infantiles, une association entre une admission le week-end et la signature d'une décharge apparaissait comme statistiquement significative lors de l'analyse conduite sur la base de données initiale [Horton *et al.*, 2007]. Cette association n'était plus significative en appliquant la méthode d'imputation multiple des données manquantes, suggérant un biais de sélection lié à la suppression des observations incomplètes. Une autre étude a observé une différence de 15,3% de l'estimation du risque d'infarctus du myocarde lié à un faible indice de masse corporelle, entre les analyses réalisées sur les données initiales et sur la base de données complétée [Delaney *et al.*, 2007].

L'imputation des données manquantes est une étape de plus dans le traitement classique des données, qui demande du temps et une réflexion supplémentaires à l'épidémiologiste. L'utilisation de logiciels prêts à l'emploi permet un gain de temps et limite les erreurs de spécification des modèles d'imputation. Cette étape supplémentaire nous semble justifiée par un gain de puissance et la valorisation

intégrale permise des données récoltées. Une interrogation demeure cependant sur la manière de procéder dans la recherche de la spécification adéquate d'un modèle. Faut-il imputer directement les données manquantes et construire le modèle, ou au contraire, commencer à travailler sur les données initiales et vérifier le modèle obtenu avec la base complétée ? Aucune solution ne semble encore clairement définie, les deux options étant utilisées dans la littérature. Une récente étude a comparé plusieurs méthodes de sélection des variables candidates à l'analyse multifactorielle lorsque l'imputation multiple des données manquantes est réalisée [Wood *et al.*, 2008]. Estimant que la sélection sur les observations complètes limite la puissance statistique de détection des associations et introduit un potentiel biais de sélection sur les coefficients de régression, les auteurs préconisent de procéder au choix des variables après la phase d'imputation multiple des données manquantes.

L'imputation multiple des données manquantes apparaît comme une avancée technique adaptée à l'exploitation des données d'enquête en épidémiologie, afin d'utiliser l'intégralité des données recueillies et optimiser la puissance des analyses conduites.

BIBLIOGRAPHIE

Allison P.D. - Missing data, 93 pages, Sage Publications, Inc., Thousand Oaks, CA, 2001.

Delaney J.A., Daskalopoulou S.S., Brophy J.M., Steele R.J., Opatrny L., Suissa S. - Lifestyle variables and the risk of myocardial infarction in the general practice research database. *BMC Cardiovasc. Disord.*, 2007, **7**, 38.

Dohoo I., Martin W., Stryhn H. - Modelling count and rate data. In: Veterinary epidemiological research. AVC Inc. (Ed), Charlottetown, Canada, 2003, 391-406.

Heckman J. - The common structure of statistical models of truncation, sample selection and limited dependant variable, and a simple estimator for such models. *Ann. Econ. Soc. Meas.*, 1976, **5**, 475-492.

Horton N.J., Kleinman K.P. - Much ado about nothing : A comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.*, 2007, **61** (1), 79-90.

Little R.J.A. - Pattern-mixture models for multivariate incomplete data. *J. Am. Stat. Assoc.*, 1993, **88**, 125-134.

Little R.J.A. - Modeling the dropout mechanism in repeated-measures studies. *J. Am. Stat. Assoc.*, 1995, **90**, 1112-1121.

Molenberghs G., Kenward M.G. - MNAR, MAR, and the nature of sensitivity. In: Missing Data in Clinical Studies. Wiley. (Ed), Chichester, England, 2007, 284-312.

Raghunathan T.E., Lepkowski J.M., Van Hoewyk J., Solenberger P. - Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 2001, **27** (1), 91-103.

Raghunathan T.E., Solenberger P., Van Hoewyk J. - IVEware : Imputation and Variance Estimation Software, 2002.

Rubin D.B. - Inference and Missing Data. *Biometrika*, 1976, **63** (3), 581-592.

Rubin D.B. - Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., New York, 1987.

Rubin D.B. - Multiple Imputation After 18+ Years. *J. Am. Stat. Assoc.*, 1996, **91**, 473-789.

SAS - SAS OnlineDoc 9.1.3., Cary, NC, SAS Institute Inc., 2007.

Schafer J.L. - Analysis of Incomplete Multivariate Data. Chapman and Hall, New York, 1997.

Taylor J.M., Cooper K.L., Wei J.T., Sarma A.V., Raghunathan T.E., Heeringa S.G. - Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. *Am. J. Epidemiol.*, 2002, **156** (8), 774-782.

Verbeke G., Molenberghs G. - Linear mixed models for longitudinal data. Springer-Verlag, New-York, 2000.

Wood A.M., White I.R., Royston P. - How should variable selection be performed with multiply imputed data? *Stat. Med*., 2008.

Wu M.C., Carroll R.J. - Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 1988, 44, 175-188.

Yu L.M., Burton A., Rivero-Arias O. - Evaluation of software for multiple imputation of semi-continuous data. *Stat. Methods Med. Res.*, 2007, **16** (3), 243-258.

Ouvrages consacrés à la thématique :

Allison P.D. - Missing data, 93 pages. Sage Publications, Inc., Thousand Oaks, CA, 2001.

Schafer J.L., Graham J.W. - Missing data : our view of the state of the art. *Psychol. Methods*, 2002, **7** (2), 147-177.

Numéro spécial sur l'imputation multiple : *Stat. Methods Med. Res.*, 2007, **16**, 95-298.

Remerciements

Les auteurs remercient les Services vétérinaires, les abattoirs et les éleveurs pour leur participation à l'enquête, le Ministère de l'agriculture et de la pêche (Direction générale de l'alimentation) pour son soutien financier. Travaux réalisés dans le cadre de l'aide au développement technologique de l'Office de l'élevage.