L'homogamie des couples – mise en œuvre de diverses méthodes de traitement de la non-réponse et analyse de leurs effets sur la mesure de l'homogamie

Mélanie VANDERSCHELDEN

INSEE, département de la démographie

Introduction

Les enquêtes « Etude de l'histoire familiale », autrefois appelées enquêtes « Famille », sont associées au recensement depuis 1954. Il s'agit d'enquêtes par sondage dont le questionnaire, auto-administré, est à caractère biographique et rétrospectif. Les personnes sont questionnées notamment sur leurs origines, leurs périodes de vie en couple, leurs conjoints, leurs enfants et leur parcours professionnel. L'information ainsi collectée est enrichie par des données provenant du recensement. En 1999, 380 000 hommes et femmes de 18 ans ou plus ont été interrogés.

Une étude sur le choix du conjoint et l'homogamie des couples va être conduite à partir de la dernière enquête « Etude de l'histoire familiale », menée en 1999. Au préalable, il convient de corriger la non réponse qui affecte les variables utiles à cette étude.

L'objectif de cet article est de comparer les avantages et inconvénients de différentes méthodes de correction de la non-réponse et d'étudier leurs effets sur des résultats de l'étude.

1. Les objectifs de l'étude

L'étude sur le choix du conjoint et l'homogamie des couples porte sur les ressemblances et dissemblances entre les deux membres d'un couple, en termes de catégorie socioprofessionnelle, d'âge, d'origine géographique, de nationalité, et de niveau d'études.
Elle vise à actualiser les résultats des études précédentes sur l'homogamie, mais aussi à répondre à un certain nombre de questions qui restent en suspens :

- Les couples non mariés sont-ils plus ou moins homogames que les couples mariés ?
- L'homogamie est-elle plus ou moins marquée parmi les couples qui vivent leur première union que parmi ceux qui ont vécu une autre union par le passé ?
- Les comportements des hommes et des femmes sont-ils très différents en matière de choix du conjoint ?
- Quel est le lien entre l'homogamie et la rupture des unions ?
- Quelle est l'évolution de l'homogamie au fil des générations sur le long terme ?

2. La non-réponse totale

Toutes les personnes ayant déjà formé une union entrent dans le champ de l'étude, qu'elles vivent en couple ou non à la date de l'enquête. La base de données utilisée comporte 322 000 enregistrements correspondant à des personnes déclarant avoir connu au moins une période de vie en couple. Pour les personnes ayant vécu plusieurs unions, seule la dernière est prise en compte, les informations sur le conjoint n'étant données que pour le dernier conjoint.

Certaines des données utiles pour l'étude proviennent directement du questionnaire de l'enquête, d'autres des bulletins individuels du recensement. Or, 2,9 % des personnes recensées ayant rempli un questionnaire de l'enquête « Etude de l'histoire familiale » et ayant déjà vécu en couple n'ont pas été retrouvées dans les fichiers du recensement au moment de l'appariement. Pour ces personnes, les données collectées via l'enquête n'ont donc pas pu être appariées avec celles provenant des bulletins de recensement. Dans ce cas, les données disponibles sont peu nombreuses et il est assez délicat de compléter l'information manquante pour ces individus. Ces cas sont assimilés à de la non-réponse totale.

Lorsque les personnes déclarent avoir déjà vécu en couple, elles doivent ensuite indiquer les principales dates de leurs périodes de vie en couple. Or, 5,2 % d'entre elles n'indiquent aucune date ou donnent des dates erronées de vie en couple (date de début postérieure à la date de fin de vie en couple par exemple). Il est très difficile dans ce cas de compléter l'information manquante, dans la mesure où il faut alors non seulement imputer une date de début d'union, mais également un nombre d'unions, et d'éventuels mariages, ruptures, divorces et décès du conjoint (et le cas échéant, leurs dates), tout en restant cohérent avec le calendrier de naissance des enfants. Ces données sont cependant importantes car elles permettent notamment de distinguer les couples mariés des couples non mariés, les premières unions des autres, les unions en cours des unions rompues et surtout de connaître la date de début et la durée des unions. L'information apportée par ces variables étant capitale pour l'étude, ces observations sont également considérées comme de la non-réponse totale.

Au total, ce sont finalement 7,9 % des observations (soit environ 25 000 sur 322 000) qui sont assimilées à de la non-réponse totale, soit parce que les personnes interrogées n'ont pas été retrouvées dans les fichiers du recensement, soit parce qu'elles n'ont pas indiqué correctement leurs dates de vie en couple.

Pour corriger la non-réponse totale, la méthode privilégiée est la repondération. Il est nécessaire de réduire le biais lié à la non-réponse totale, mais il est important également que l'effectif total, ainsi que quelques uns des effectifs marginaux, soient identiques à ceux donnés par les études déjà réalisées à partir de l'enquête « Etude de l'histoire familiale » de 1999. La répartition des personnes ayant déjà vécu en couple selon le sexe et l'âge, et selon la situation de couple[1] (première union en cours, première union rompue, deuxième union en cours, deuxième union rompue) doit rester inchangée. La technique du calage sur marges, qui permet de faire coïncider la répartition de certaines variables (choisies comme variables de calage) dans l'échantillon et dans la population, est donc employée ici.

Le calage est réalisé sur les marges calculées à partir des 322 000 enregistrements du fichier de l'enquête « Etude de l'histoire familiale » de 1999 correspondant aux personnes ayant déjà vécu en couple.
Les variables de calage doivent être renseignées pour les 322 000 observations du fichier. Parmi les variables qui remplissent cette condition, ce sont celles qui expliquent le mieux la non-réponse qui sont sélectionnées, à l'aide d'une régression logistique. Le modèle dichotomique contenant comme variable expliquée la variable réponse (prenant comme modalités 0 ou 1 selon que

[1] La variable situation de couple est une variable déclarative. Cette information n'est donc pas déduite des dates indiquées de vie en couple. Les informations données par cette variable et les dates de vie en couple concordent cependant presque toujours (dans plus de 98 % des cas) lorsque les dates sont correctement renseignées. Pour l'étude sur l'homogamie, la répartition des personnes selon leur situation de couple sera calculée non pas à partir de la variable déclarative, mais à partir des dates de vie en couple. Elle ne sera donc pas tout à fait calée sur la répartition de la population, qui est calculée à partir de la variable déclarative, mais très proche.

l'observation correspond ou ou non à une observation assimilée à de la non-réponse totale) et comme variables explicatives la combinaison du sexe et de l'âge en 1999, la situation de couple, la région de résidence, le groupe social et l'état matrimonial (célibataire, marié, veuf ou divorcé) de la personne estime correctement la variable réponse dans environ 70 % des cas, ce qui semble satisfaisant. Hormis la combinaison du sexe et de l'âge et la situation de couple, variables d'office sélectionnées comme variables de calage puisque leur répartition ne doit pas être modifiée, les variables qui expliquent le mieux la non-réponse sont donc la région, l'état matrimonial et le groupe social de la personne interrogée. Ces cinq variables sont finalement retenues comme variables de calage.

3. La non-réponse partielle

Outre les principales dates de vie en couple, les variables suivantes sont indispensables à l'étude :

- le groupe social (catégorie socioprofessionnelle) de la personne, celui de son dernier conjoint et celui de son père
- l'année de naissance de la personne et celle de son dernier conjoint[2]
- le pays ou le département de naissance (pour les personnes nées en France) de la personne et celui de son dernier conjoint
- la nationalité de la personne et celle de son dernier conjoint[3]
- le niveau d'études de la personne et celui de son dernier conjoint[4]
- l'état matrimonial antérieur du dernier conjoint de la personne[4].

Les variables groupe social et année de naissance de la personne sont toujours renseignées. Mais parmi les 297 000 observations conservées (c'est-à-dire hormis celles qui ont été assimilées à de la non-réponse totale), certaines sont concernées par la non-réponse partielle, c'est-à-dire qu'elles comptent au moins une valeur manquante sur les dix variables restantes. Une partie des valeurs manquantes (et des incohérences) est corrigée de manière déductive, à l'aide des différentes variables disponibles. Après corrections, 21,5 % de ces observations (soit environ 64 000) comportent toujours au moins une valeur manquante sur ces dix variables. Selon les variables, le taux de non-réponse est plus ou moins élevé : il varie de 0,1 % pour la nationalité de la personne à 8,4 % pour la catégorie socioprofessionnelle du père.

Dans 59 % des cas où au moins une variable est non renseignée, une seule variable est manquante, mais les cas de non-réponses multiples ne sont pas rares : pour 19,6 % des 64 000 observations concernées par la non-réponse, deux de ces variables sont non renseignées, et dans 16,6 % des cas, trois ne sont pas renseignées. Dans environ 4 % des cas, on compte plus de trois non-réponses.
Au total, au moins deux variables sont manquantes pour 8,7 % des 297 000 observations non assimilées à de la non-réponse totale.

La non-réponse partielle est corrigée par imputation. Il reste à choisir une méthode d'imputation adéquate.

Toutes les variables à imputer sont qualitatives. Les méthodes d'imputation par substitution semblent donc les plus adaptées. Les méthodes de hot-deck consistent à remplacer, pour une observation appelée receveur, une valeur manquante sur une variable donnée par une valeur observée sur la même variable pour un individu répondant choisi au hasard et appelé donneur. Si

[2] On ne cherchera pas à connaître, pour les non répondants, l'année de naissance du conjoint, mais l'écart d'âge entre les deux conjoints.
[3] Les variables nationalité et niveau d'études du conjoint proviennent du recensement. Elles ne peuvent donc être renseignées que dans le cas où la personne interrogée réside toujours avec son conjoint, c'est-à-dire dans le cas où l'union est en cours. De même, l'état matrimonial antérieur du conjoint ne doit être indiqué que si le conjoint avait déjà été marié avant le début de l'union. Pour les non répondants, on ne cherchera donc à connaître la nationalité et le niveau d'études du conjoint que si l'union est en cours et l'état matrimonial antérieur du conjoint que dans le cas où il avait déjà été marié auparavant.

la population est trop hétérogène pour que le répondant et le non répondant soient relativement proches, elle peut être décomposée en sous populations plus homogènes. Des classes d'imputation sont alors constituées et le donneur est choisi à l'intérieur de la classe à laquelle appartient le receveur.

3.1. Imputations indépendantes par hot-deck séquentiel

La méthode du hot-deck séquentiel est relativement simple à mettre en œuvre et couramment employée. Il s'agit de déterminer les variables les plus corrélées à la variable à imputer, à partir des données recueillies sur les seuls répondants (variables auxiliaires). Ces variables doivent être renseignées pour les répondants, qui sont utilisés pour repérer les corrélations, mais aussi pour les non répondants, pour lesquelles les variables auxiliaires sont utilisées pour réaliser l'imputation. Elles ne doivent donc finalement être manquantes pour aucune observation. Le fichier des données est ensuite trié selon ces variables. L'échantillon est ainsi ordonné, et pour chaque valeur manquante, la modalité de la variable prise par le répondant précédent est imputée. Cette méthode présente cependant un inconvénient : un même donneur peut être utilisé plusieurs fois, ce qui n'est pas souhaitable. En effet, la duplication des mêmes valeurs augmente la variance, et peut augmenter aussi le biais dans le cas où le donneur est atypique. Le choix de variables auxiliaires peu corrélées à la non-réponse permet toutefois de réduire le nombre de duplications d'une même valeur.

Une première sélection des variables auxiliaires est effectuée à l'aide de tris croisés. Dans un deuxième temps, cette sélection est affinée à l'aide d'une modélisation. Compte tenu de la nature des variables (qualitatives à plus de deux modalités non ordonnées), ce sont les modèles polytomiques non ordonnés qui sont utilisés ici.

L'imputation est réalisée séparément pour chacune des variables. A chaque fois, le fichier est trié dans l'ordre des variables auxiliaires sélectionnées à l'aide des tris croisés et des modèles. L'imputation étant réalisée indépendamment pour chaque variable, les liens qui peuvent exister entre les différentes variables imputées sont négligés. Pour tenir compte de ces liens, deux adaptations de cette méthode peuvent être envisagées. Elles sont citées ici à titre indicatif mais ne sont pas mises en œuvre.

Méthode 1 : Il s'agit de choisir, parmi l'ensemble des variables qui expliquent le mieux chacune des variables à imputer, quelques variables auxiliaires, qui sont utilisées pour réaliser un unique tri du fichier. Lorsqu'il existe une ou plusieurs valeurs manquantes pour une observation donnée, les valeurs de l'observation précédente sont imputées pour ces variables (voir schéma ci-dessous).

Variables de tri			Variables à imputer									
N°1	N°2	N°3	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	25	7	1	3	21	21	01	2	1	01	1	1
1	25	8	**1**	0	59	62	01	3	6	01	2	3
1	30	5	5	-11	87	75	01	2	3	01	3	3
1	30	8	2	1	63	**75**	**01**	1	3	01	1	2
2	25	3	6	2	**63**	**75**	P22	2	5	01	**1**	2

Sur le schéma ci-dessus, les valeurs grisées ont été imputées. Cette méthode présente deux inconvénients. Une fois les valeurs manquantes qu'elle contient imputées, une observation qui correspondait initialement à un receveur est ensuite considérée comme donneur. Or, les valeurs des différentes variables de cette observation n'ont pas le même « statut ». Les valeurs prises par les variables qui sont observées sont « vraies », tandis que les valeurs des variables qui ont été imputées sont « supposées vraies ». C'est le cas par exemple de la quatrième ligne du schéma précédent. La variable lieu de naissance de la personne a été imputée, contrairement à la variable lieu de naissance du conjoint. A la limite, il serait possible de considérer que pour la dernière observation, le donneur est la quatrième observation pour la variable lieu de naissance du conjoint, et la

troisième pour la variable lieu de naissance de la personne. Pour cette dernière observation, il serait préférable de choisir comme donneur la troisième, pour laquelle les valeurs des variables lieu de naissance du conjoint et de la personne ont été observées.

Les inconvénients de l'utilisation indifférenciée de valeurs dites « vraies » et « supposées vraies » sont souvent occultés. Ils doivent pourtant être considérés comme majeurs si les taux de non-réponse sont élevés ou si les variables à imputer sont nombreuses, car l'imputation serait finalement réalisée sur la base d'informations auxiliaires artificielles.

La nécessité de choisir des variables auxiliaires uniques qui soient pertinentes pour toutes les variables à imputer est un autre inconvénient de cette méthode. Ce choix peut s'avérer très difficile si les variables à imputer sont nombreuses.

Méthode 2 : Une deuxième méthode consiste à imputer les variables une par une, dans un ordre à déterminer au préalable. Pour chaque imputation, le fichier est trié dans un ordre différent (selon les variables auxiliaires qui expliquent le mieux la variable imputée). La première variable est imputée comme précédemment, en utilisant comme variables auxiliaires celles qui l'expliquent le mieux, puis la variable imputée est utilisée comme variable auxiliaire (elle peut être associée à d'autres variables) pour imputer la variable suivante et ainsi de suite.
Reprenons l'exemple précédent. Dans un premier temps, seule la variable nationalité de la personne est imputée. Le fichier est trié selon la variable nationalité de la personne (française/étrangère), qui est celle qui explique le mieux la variable nationalité de la personne (détaillée). Le schéma suivant montre le résultat de cette première imputation.

Variable de tri	Variables à imputer									
Nationalité de la personne (française/étrangère)	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
0	5	-11	87	75	01	2	3	01	3	3
0	1	3	21	21	01	2	1	01	1	1
0	.	0	59	62	01	3	6	01	2	3
0	2	1	63	.	**01**	1	3	01	1	2
1	6	2	.	.	P22	2	5	01	.	2

Dans un deuxième temps, une autre variable, le lieu de naissance de la personne, est imputée, mais il est tenu compte des résultats de la première imputation. Le fichier est trié selon les variables lieu de naissance de la personne (France/étranger) et département de résidence de la personne (variables qui expliquent le mieux la variable lieu de naissance de la personne), et selon la variable nationalité de la personne (imputée). La position de cette dernière variable parmi toutes les variables de tri peut-être choisie, par exemple, par une nouvelle modélisation. Le schéma ci-dessous montre le résultat de cette deuxième imputation. La variable nationalité de la personne (en italique) n'est plus à imputer, puisqu'elle l'a été précédemment.

Variables de tri			Variables à imputer								
Lieu naiss. Pers.	Dép. rés. pers	*Nat. pers*	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	59	*01*	.	0	59	62	3	6	01	2	3
1	63	***01***	2	1	63	**62**	1	3	01	1	2
1	77	*01*	5	-11	87	75	2	3	01	3	3
1	92	*01*	1	3	21	21	2	1	01	1	1
0	93	*P22*	6	2	.	**21**	2	5	01	.	2

Cette deuxième méthode permet d'utiliser plus d'information auxiliaire que la première et d'éviter d'avoir à choisir seulement quelques variables auxiliaires pertinentes pour toutes les variables imputées. Cependant, les observations pour lesquelles certaines variables ont été imputées contiennent des valeurs observées, donc « vraies », et d'autres imputées, donc « supposées vraies », qui sont utilisées indifféremment pour imputer les autres variables.

L'inconvénient de la méthode 1 évoqué précédemment est donc aussi un inconvénient de la méthode 2. Il faut de plus choisir un ordre pour l'imputation des variables, ce qui, dans le cas traité ici, est loin d'être évident. En effet, il est possible de commencer par imputer la variable nationalité de la personne comme dans l'exemple ci-dessus, puisqu'elle est très rarement manquante, puis d'imputer le lieu de naissance de la personne, et d'utiliser ensuite ces deux variables, une fois imputées, pour imputer les variables nationalité et lieu de naissance du conjoint. En revanche, il est bien difficile de définir un ordre pour les autres variables.

Les imputations indépendantes par hot-deck séquentiel ne semblent en fin de compte pas adaptées au cas que nous traitons ici puisqu'elles ne permettent pas de prendre en compte de façon satisfaisante le lien entre les différentes variables imputées.

3.2. Imputations simultanées des valeurs prises par un unique donneur choisi au hasard

Afin de tenir compte des éventuelles relations entre les différentes variables à imputer, un même donneur est utilisé pour imputer simultanément toutes les variables non renseignées d'une même observation.

Des classes d'imputation sont constituées. Une observation est affectée à une classe en fonction des modalités prises par certaines variables auxiliaires. Ces variables sont sélectionnées parmi toutes celles qui expliquent le mieux les variables à imputer (cf. 3.1.). Si les variables sexe, groupe social de la personne et âge de la personne au début de l'union (en tranches) sont retenues, les classes obtenues sont de taille assez importante. Les observations sont ainsi réparties en 209 classes. Le donneur est choisi au hasard au sein de la classe à laquelle appartient le receveur. Cette méthode est appelée hot-deck aléatoire par classes.

Pour une observation donnée, la liste des variables à imputer est une combinaison de une à dix variables. Il n'est pas possible de définir comme donneurs potentiels, pour chacun des 1 022 cas possibles ($C_{10}^{1} + C_{10}^{2} + ... + C_{10}^{10}$), les observations pour lesquelles ces variables sont renseignées.

Une solution simple pourrait alors être adoptée : seules les observations pour lesquelles les dix variables sont renseignées seraient considérées comme donneurs potentiels.

Cependant, les variables nationalité et niveau d'études du conjoint ne peuvent être renseignées que pour les personnes dont l'union est encore en cours. De même, la variable état matrimonial antérieur du conjoint n'est renseignée que pour celles dont le conjoint avait déjà été marié avant le début de l'union. De ce fait, les donneurs potentiels ne pourraient correspondre qu'à ce profil.

La solution finalement retenue consiste à classer les observations à imputer en quatre cas :

cas 1 - Ni les variables nationalité et niveau d'études du conjoint, ni la variable état matrimonial du conjoint ne sont à imputer.

cas 2 - L'une (au moins) des variables nationalité et niveau d'études du conjoint est à imputer, mais pas la variable état matrimonial antérieur du conjoint.

cas 3 - Les variables nationalité et niveau d'études du conjoint ne sont pas à imputer, mais la variable état matrimonial antérieur du conjoint l'est.

cas 4 - L'une (au moins) des deux variables nationalité et niveau d'études du conjoint est à imputer et la variable état matrimonial antérieur du conjoint aussi.

Dans chacun de ces quatre cas, l'une ou plusieurs des sept autres variables peuvent être également à imputer.

Les donneurs qui seront utilisés pour imputer les observations correspondant au cas 1 ne devront pas avoir de valeurs manquantes sur les sept variables susceptibles d'être imputées. Pour les

donneurs qui serviront à imputer les observations correspondant au cas 2, ces sept mêmes variables ainsi que les variables nationalité et niveau d'études du conjoint devront être renseignées. Les observations pour lesquelles la nationalité et le niveau d'études du conjoint ne sont pas nécessairement renseignées, mais qui n'ont pas de valeurs manquantes sur la variable état matrimonial antérieur du conjoint et sur les sept autres variables seront des donneurs potentiels pour les observations du cas 3. Enfin, seules les observations pour lesquelles les dix variables sont renseignées pourront être utilisées comme donneurs pour imputer les observations correspondant au cas 4.

Quelques classes sont regroupées avec d'autres de sorte que dans chaque classe, le nombre de donneurs potentiels soit supérieur ou égal au nombre de receveurs. Bien que le nombre de donneurs potentiels dans chaque classe soit suffisant pour utiliser un hot-deck aléatoire sans remise, ce qui est toujours préférable en théorie, c'est un hot-deck aléatoire avec remise qui est réalisé ici, afin de comparer la méthode du hot-deck aléatoire à celle du hot-deck séquentiel, qui se rapproche davantage d'une méthode de tirage avec remise des donneurs.

La condition fixée pour qu'une observation soit considérée comme donneur potentiel (toutes les variables doivent être renseignées) est en fait assez contraignante. Parmi l'ensemble des observations à imputer, plus de 59 % ne comptent qu'une seule valeur manquante. Il est possible de traiter séparément les cas où une seule variable est manquante, pour lesquels sont considérées comme donneurs potentiels toutes les observations pour lesquelles cette variable est renseignée, ce qui revient à décomposer les quatre cas précédents en quatorze cas. La contrainte imposée aux donneurs potentiels est ainsi desserrée pour une bonne partie des cas. Lorsqu'une seule variable est manquante, le donneur est choisi parmi un nombre plus grand de donneurs potentiels. Auparavant, lorsque la variable lieu de naissance de la personne était imputée pour une observation pour laquelle toutes les autres variables étaient renseignées, le donneur était choisi parmi 241 000 observations. Désormais, il est sélectionné parmi 294 000 donneurs potentiels.

Si deux variables au moins sont manquantes, les donneurs doivent remplir les mêmes conditions que précédemment.

Pour former les classes d'imputation, il existe d'autres techniques, telles que les méthodes de classification, qui permettent de prendre en compte davantage d'information auxiliaire. Compte tenu du nombre d'observations et de la nature des variables, cette méthode est cependant relativement lourde à mettre en œuvre car il faut dans un premier temps réaliser une ACM et récupérer les coordonnées des individus sur les axes factoriels puis recourir à une méthode de classification mixte. De plus, le nombre de modalités est très différent d'une variable à l'autre et les effectifs de certaines modalités sont très faibles. Or, pour l'ACM, ce cas de figure pose problème, l'inertie d'une variable croissant avec le nombre de ses modalités et l'inertie d'une modalité dépendant de son effectif. Il est donc nécessaire d'opérer de nombreux regroupements de modalités au préalable, et d'ôter certaines variables auxiliaires.

Les résultats donnés par les méthodes de classification ne sont pas très satisfaisants. En effet, les classes obtenues sont de tailles très inégales. Certaines sont très petites et regroupent des individus un peu « atypiques », et d'autres, très grandes, sont peu homogènes. Dans les classes les plus petites, il pourrait être difficile de trouver des donneurs, ce qui nécessiterait des regroupements de classes, tandis que dans les grandes classes, les donneurs pourraient présenter des caractéristiques assez éloignées de celles des receveurs. De plus, l'information auxiliaire utilisée pour réaliser la classification est fortement appauvrie par les regroupements de modalités et la suppression de variables.

Le choix d'un donneur unique pour imputer simultanément toutes les variables manquantes d'une même observation permet de prendre en compte les liens entre les différentes variables imputées, mais la méthode du hot-deck aléatoire par classes ne permet pas d'exploiter au mieux l'information auxiliaire. Il faut alors chercher une autre méthode d'imputation qui permette d'utiliser un maximum d'information auxiliaire tout en préservant les liens entre variables.

3.3. Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Le hot-deck métrique est une méthode d'imputation qui consiste à remplacer une valeur manquante par la valeur observée pour l'individu le plus proche, au sens d'une « distance »[4] à définir et calculée à partir des variables auxiliaires renseignées pour les répondants et les non répondants.

Puisqu'un donneur unique sert à imputer les différentes variables manquantes d'une observation donnée, les donneurs doivent remplir les mêmes conditions que précédemment : toutes les variables susceptibles d'être imputées dans chacun des cas doivent être renseignées pour qu'une observation puisse être utilisée comme donneur. Il est toujours possible de distinguer les cas où, pour une observation donnée, une seule variable est à imputer, mais pour réduire le temps de calcul, seuls les quatre cas définis au paragraphe précédent sont considérés.

La distance utilisée pour mesurer la proximité entre le receveur et les donneurs potentiels est une somme pondérée et normalisée (comprise entre 0 et 1) de distances partielles.

Les variables auxiliaires retenues pour le calcul de la distance sont celles qui expliquent le mieux chacune des dix variables à imputer (cf. 3.1.). Elles sont au nombre de neuf, et sont toutes qualitatives. Pour chaque variable auxiliaire, une distance partielle est définie. Elle vaut 0 ou 1 selon que la variable sur laquelle elle est calculée prend ou non la même modalité pour le receveur et le donneur potentiel.

Chaque distance partielle est pondérée de façon à ce que l'importance de chaque variable auxiliaire pour l'imputation soit prise en compte. Le poids utilisé correspond à une mesure de la corrélation entre la variable à imputer et la variable auxiliaire, ce qui réduit le caractère arbitraire du choix de la pondération. Il s'agit du V de Cramer, qui est compris entre 0 et 1, et qui ne tient compte ni de la taille de l'échantillon, ni du nombre de modalités des variables. Si une variable auxiliaire explique plus d'une variable à imputer, on calcule le V de Cramer entre la variable auxiliaire et chacune des variables à imputer qu'elle explique et on attribue comme poids à la variable auxiliaire la somme des coefficients.

Enfin, afin d'obtenir une distance normalisée, la somme pondérée des neuf distances partielles est divisée par la somme des poids.

Les neuf variables auxiliaires sont ainsi utilisées. Il n'est pas nécessaire de sélectionner certaines d'entre elles. Supposons maintenant que pour une observation donnée, les variables lieu de naissance du conjoint, nationalité du conjoint et nationalité de la personne soient manquantes. Le lien entre les variables lieu de naissance du conjoint et nationalité du conjoint, toutes deux manquantes, est pris en compte lors de l'imputation puisqu'on choisit toujours un donneur unique pour imputer toutes les variables à blanc d'une observation donnée. Cependant, le lien qui peut exister entre les variables nationalité de la personne (manquante) et lieu de naissance de la personne (renseignée) n'est pas pris en compte. Pour pallier ce problème, dix distances partielles supplémentaires sont ajoutées à la distance définie ci-dessus, chacune d'entre elles étant calculée sur l'une des variables à imputer. Le lieu de naissance de la personne est l'une de ces variables. Si cette variable est renseignée pour le receveur et un donneur potentiel et si tous deux ne sont nés ni dans le même département français, ni dans le même pays (autre que la France), la distance partielle calculée sur la variable lieu de naissance de la personne vaut 1. Dans tous les autres cas, elle vaut 0. Ces dix distances partielles sont pondérées de la même façon que les neuf autres. L'information auxiliaire utilisée est donc enrichie.

[4] On parle ici de distance mais il s'agit en fait d'une mesure de similarité.

Finalement, on obtient la distance suivante :

$$D(r,d) = \frac{\sum_{j=1}^{p} \omega_j \delta_{j,rd} + \sum_{k=1}^{q} \omega_k \delta_{k,rd}}{\sum_{j=1}^{p} \omega_j + \sum_{k=1}^{q} \omega_k}$$

où p est le nombre de variables auxiliaires, q le nombre de variables à imputer, ω_j (resp. ω_k) le poids de la variable auxiliaire j (resp. k), et $\delta_{j,rd}$ (resp. $\delta_{k,rd}$) vaut 0 si la variable j (resp. k) prend la même modalité pour le receveur r et le donneur d, c'est-à-dire si $x_{j,r} = x_{j,d}$ (resp. $x_{k,r} = x_{k,d}$), et 1 sinon.

Afin de choisir le donneur le plus proche, il faut calculer, pour chaque receveur, cette distance entre ce dernier et chacun des donneurs potentiels. Compte tenu du nombre de receveurs et de donneurs potentiels, il est impossible de réaliser simplement le produit cartésien de la table des receveurs et de la table des donneurs. En effet, le résultat de ce produit serait une table de 2 millions à 12 milliards de lignes selon le cas. Le calcul est réalisé receveur par receveur, en évitant les tris de tables, très coûteux en temps. Le temps de calcul reste malgré tout très long. Il est de l'ordre de quelques dizaines d'heures.

Finalement, le hot-deck métrique permet de prendre en compte le lien entre les variables imputées sans qu'il soit nécessaire d'utiliser comme information auxiliaire des variables dont certaines valeurs peuvent avoir elles-mêmes été imputées. Aucune hiérarchie entre les différentes variables imputées n'est induite par la méthode d'imputation puisque toutes les variables auxiliaires qui expliquent les dix variables à imputer sont utilisées. De plus, le choix du donneur est plus « rationnel », et le donneur plus proche du receveur, plus « ressemblant ».

Le principal inconvénient de cette méthode est le temps de calcul qu'elle nécessite. Celui-ci dépend fortement de la taille des tables des receveurs et des donneurs. Afin de réduire la taille de la table des donneurs, il est possible de réaliser un hot-deck métrique par classes. Par exemple, pour un receveur agricultrice (resp. agriculteur), le donneur sélectionné serait le plus proche parmi les donneurs agricultrices (resp. agriculteurs). Il est possible aussi de combiner les méthodes du hot-deck hiérarchisé et du hot-deck métrique. Le donneur serait le plus proche du receveur parmi ceux présentant les caractéristiques X_1, X_2, …, X_k s'il en existe au moins un. Sinon, il s'agirait du plus proche parmi ceux présentant les caractéristiques X_1, X_2, …, X_{k-1} s'il en existe au moins un, et ainsi de suite.

La méthode du hot-deck métrique, telle qu'elle est utilisée ici, est déterministe, une fois la distance définie. Afin de donner une part d'aléa à l'imputation, une variante de cette méthode peut être utilisée. Elle consiste à choisir un donneur aléatoirement parmi les x (la valeur de x restant à fixer – 5 ou 10 par exemple) plus proches du receveur ou parmi les individus pour lesquels la distance est inférieure à un seuil (à fixer également). Cette variante peut conduire au choix de donneurs moins proches des receveurs. De plus, de même que pour un hot-deck métrique sans remise, le temps de calcul est plus long. Cependant, elle donne un caractère moins déterministe à l'imputation et permet en principe de réduire le nombre d'utilisations d'un même donneur.

Les distances partielles choisies ici valent soit 0, soit 1, selon que la variable prend ou non la même modalité pour le donneur et le receveur. La distance partielle peut aussi prendre une valeur comprise entre 0 et y, y étant un nombre fixé arbitrairement. Il est ainsi possible de créer des distances partielles qui prennent plus de deux valeurs, et qui n'ont pas nécessairement toutes le même nombre de modalités. Par exemple, la distance partielle pourrait valoir 0, 5 ou 10 pour la variable groupe social de la personne selon que le donneur et le receveur appartiennent au même groupe social, à des groupes sociaux proches ou éloignés. Pour la variable nationalité de la personne, la distance partielle pourrait valoir 0 si le donneur et le receveur sont tous deux

français ou bien étrangers mais de même nationalité et 10 sinon, etc. Des distances partielles de ce type peuvent être utiles dans les cas où les modalités des variables sont nombreuses et/ou la probabilité que la variable prenne la même valeur pour le receveur et le donneur est faible. Il faut cependant veiller à ce que le nombre y soit commun aux différentes distances partielles, afin de ne pas donner implicitement plus de poids à certaines variables. Les valeurs des distances partielles autres que les valeurs minimale et maximale restent à fixer, et ne peuvent l'être qu'arbitrairement.

Le hot-deck métrique est également utilisable dans le cas où les variables sont toutes quantitatives, mais dans ce cas, la distance euclidienne semble plus adaptée. Si certaines des variables sont qualitatives et d'autres quantitatives, les distances partielles, pour les variables qualitatives, peuvent valoir par exemple soit 0, soit 1, selon que la variable prend ou non la même modalité pour le donneur et le receveur ou bien être définies comme énoncé au paragraphe précédent. Pour les variables quantitatives, elles peuvent être définies comme suit :

$$\delta_{j,rd} = \frac{\left| x_{j,r} - x_{j,d} \right|}{R_j}$$

où R_j est l'étendue de la variable j. Pour une variable quantitative, la distance partielle peut aussi valoir 0 si $\left| x_{j,r} - x_{j,d} \right| \leq T$ où T est un seuil fixé et tel que $T \leq R_j$ et 1 sinon. Les distances partielles sont ainsi du même ordre de grandeur pour toutes les variables.

4. Les résultats

4.1. Utilisation des donneurs

Chacune des trois méthodes employées peut conduire à utiliser plusieurs fois le même donneur. Par exemple, avec le hot-deck métrique, un même donneur est utilisé en moyenne 1,25 fois. Dans 81 % des cas, un même donneur n'est utilisé qu'une seule fois. Dans 15 % des cas, il est utilisé deux fois et dans 4 % des cas, il sert trois fois ou plus. Cependant, dans le cas où un donneur est utilisé plus d'une fois, il ne sert pas toujours à imputer les mêmes variables. Ainsi, pour imputer la variable catégorie socioprofessionnelle du conjoint, un même donneur a été utilisé en moyenne 1,11 fois et 91 % des donneurs utilisés n'ont servi qu'une fois. Pour cette même variable, ces chiffres sont de 1,13 et 89 % avec le hot-deck séquentiel et de 1,08 et 93 % avec le hot-deck aléatoire (14 cas). Les résultats obtenus avec chacune de ces trois méthodes en termes de duplication des mêmes valeurs sont donc assez similaires.

4.2. Proximité des donneurs et des receveurs (hot-deck métrique)

Le hot-deck métrique permet de choisir, dans la grande majorité des cas, un donneur très proche du receveur. La distance entre le receveur et le donneur sélectionné, normalisée, est en moyenne de 0,11. Elle n'est jamais supérieure à 0,42. En moyenne, moins de deux des neuf variables auxiliaires prennent des modalités différentes pour le receveur et le donneur sélectionné. Le nombre moyen de distances partielles non nulles pour les dix autres variables est légèrement supérieur à un. Au total, le receveur et le donneur ont en moyenne moins de trois caractéristiques différentes. Pour environ 95 % des receveurs, le receveur et le donneur diffèrent pour moins de cinq caractéristiques. Le nombre de caractéristiques différentes n'est jamais supérieur à neuf.

4.3. Conséquences de la correction de la non-réponse sur les résultats globaux de l'étude

Tableau 1 : Répartion (en %) de la variable catégorie socio-professionnelle du conjoint.

Catégorie socio-professionnelle du conjoint	***Répartition avant l'imputation***		Après hot-deck séquentiel		Après hot-deck aléatoire (14 cas)		Après hot-deck métrique		Après hot-deck métrique et calage sur marges	
	Proportion	***Intervalle de confiance***	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance
Agriculteurs exploitants	***4,8***	***[4,7;4,9]***	4,8	[4,7;4,9]	4,8	[4,7;4,9]	4,9	[4,8;5,0]	4,9	[4,8;5,0]
Artisans, commerçants, chefs d'entreprise	***7,4***	***[7,3;7,5]***	7,4	[7,3;7,5]	7,4	[7,3;7,5]	7,3	[7,2;7,4]	7,3	[7,2;7,4]
Cadres et professions intellectuelles supérieures	***8,7***	***[8,6;8,8]***	8,6	[8,5;8,7]	**8,5**	[8,4;8,6]	**8,5**	[8,4;8,6]	**8,4**	[8,3;8,5]
Professions intermédiaires	***18,8***	***[18,7;18,9]***	**18,6**	[18,5;18,7]	18,7	[18,6;18,8]	**18,5**	[18,4;18,6]	**18,5**	[18,4;18,6]
Employés	***29,7***	***[29,5;29,9]***	29,9	[29,7;30,1]	29,9	[29,7;30,1]	29,8	[29,6;30,0]	29,8	[29,6;30,0]
Ouvriers	***26,1***	***[25,9;26,3]***	26,0	[25,8;26,2]	26,2	[26,0;26,4]	26,3	[26,1;26,5]	26,3	[26,1;26,5]
Sans activité professionnelle	***4,5***	***[4,4;4,6]***	**4,7**	[4,6;4,8]	4,6	[4,5;4,7]	**4,7**	[4,6;4,8]	**4,9**	[4,8;5,0]
Ensemble	***100,0***		100,0		100,0		100,0		100,0	
Champ	**Hors non répondants pour les variables concernées, pondération initiale (avant calage)**		**Complet, pondération initiale (avant calage)**						**Complet, pondération finale (après calage)**	
Effectif pondéré (en milliers)	***29 580***		**32 517**						**35 765**	

Note : Les intervalles de confiance sont calculés sous l'hypothèse d'un tirage aléatoire simple[5].

La distribution des variables est très peu modifiée par l'imputation et le calage. Par exemple, la proportion de personnes vivant avec un conjoint cadre ou exerçant une profession intellectuelle supérieure passe de 8,7 % avant l'imputation à 8,6 % après imputation par hot-deck séquentiel, à 8,5 % après imputation par hot-deck aléatoire ou par hot-deck métrique et à 8,4 % après imputation par hot-deck métrique et calage sur marges (cf. tableau 1). L'intervalle de confiance associé à la proportion de personnes vivant avec un conjoint cadre, calculée avant l'imputation (8,7 %), est de [8,6 ; 8,8]. Seul le résultat obtenu avec le hot-deck séquentiel reste inclus dans cet intervalle de confiance[6]. Les corrections effectuées ne sont donc pas tout à fait neutres sur la distribution des variables. Celle-ci tend à varier légèrement plus avec l'imputation par hot-deck aléatoire qu'avec l'imputation par hot-deck séquentiel, et elle est aussi un tout petit peu plus sensible à l'imputation par hot-deck métrique qu'à l'imputation par hot-deck aléatoire, mais les variations sont du même ordre.

Les conséquences de l'imputation et du calage sur les résultats globaux de l'étude sont limitées mais ne sont pas non plus négligeables (cf. tableau 2).

La proportion de couples constitués de deux personnes du même âge à plus ou moins un an, calculée sur les seuls répondants, est de 29,0 %. Après imputation, elle passe à 28,9 %, quelle que soit la méthode utilisée. Après imputation par hot-deck métrique et calage sur marges, elle est à nouveau de 29,0 %. Les résultats obtenus après correction de la non réponse restent inclus dans l'intervalle de confiance de la proportion calculée avant l'imputation ([28,8 ;29,2]). La plupart des résultats reposant sur une seule variable imputée et calculés sur l'ensemble de la population ne sont pas sensiblement modifiés par les corrections effectuées. Mais les corrections ont un impact un peu plus net sur les résultats relatifs à des sous-ensembles de population, du fait de leur plus petite taille. Ainsi, la proportion de couples constitués de deux personnes du même groupe social varie très peu selon la méthode employée, et le résultat obtenu après correction de la non-réponse reste toujours inclus dans l'intervalle de confiance associé à la proportion calculée avant l'imputation. En revanche, la proportion d'ouvriers vivant avec un conjoint ouvrier passe de 35,2 % avant l'imputation à 34,8 % après imputation par hot-deck métrique et calage sur marges, et sort de l'intervalle de confiance ([35,0 ;35,4]).

[5] Les intervalle de confiance ne sont donc pas tout à fait exacts, et ont ici une valeur indicative, mais la prise en compte du vrai plan de sondage ne les modifierait probablement pas, notamment du fait de la grande taille de l'échantillon (de l'ordre de 300 000 individus), qui explique d'ailleurs la faible étendue des intervalles de confiance.

[6] La relative stabilité des résultats ne doit pas être interprétée comme un indicateur de « qualité » de l'imputation. Il en est de même du fait qu'une proportion calculée après l'imputation reste incluse dans l'intervalle associé à la proportion calculée avant l'imputation.

Lorsque les résultats portent sur plusieurs variables imputées, ils sont beaucoup plus sensibles aux corrections effectuées et à la méthode utilisée, et ce d'autant plus qu'ils portent sur de petites sous-populations. Par exemple, la proportion de personnes françaises vivant avec un conjoint français varie très peu (entre 98,1 et 98,2 %) selon la méthode utilisée. En revanche, la proportion de personnes étrangères vivant avec un conjoint de même nationalité qu'elles, qui est calculée sur un effectif nettement plus faible, est de 66,3 % avant l'imputation, passe à 63,3 % après imputation par hot-deck séquentiel et à 63,0 % après imputation par hot-deck aléatoire, et sort donc de l'intervalle de confiance ([66,1 ;66,5]). En revanche, après imputation par hot-deck métrique, cette proportion est de 66,6 % et le calage la fait passer à 66,1 %. Ces différences s'expliquent par le fait que le hot-deck métrique tient compte du lien entre les variables auxiliaires et les variables imputées, entre les différentes variables imputées pour une observation donnée, mais aussi du lien entre les variables imputées et les variables renseignées (parmi les dix) pour cette observation.

Tableau 2 : Quelques résultats de l'étude.

Résultats globaux sur l'homogamie	**Répartition avant l'imputation**		Après hot-deck séquentiel		Après hot-deck aléatoire (14 cas)		Après hot-deck métrique		Après calage et hot-deck métrique	
	Proportion	**Intervalle de confiance**	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance	Proportion	Intervalle de confiance
Résultats obtenus à partir de variables dont une au plus a pu être imputée										
% couples 2 pers. même groupe social	**30,0**	**[29,8;30,2]**	30,1	[29,9;30,3]	30,0	[29,8;30,2]	29,9	[29,7;30,1]	29,9	[29,7;30,1]
% d'ouvriers avec un conjoint ouvrier	**35,2**	**[35,0;35,4]**	**34,9**	[34,7;35,1]	**34,9**	[34,7;35,1]	35,0	[34,8;35,2]	**34,8**	[34,6;35,0]
% couples 2 pers. nées une même année civile	**10,5**	**[10,4;10,6]**	10,4	[10,3;10,5]	10,5	[10,4;10,6]	10,5	[10,4;10,6]	10,5	[10,4;10,6]
% couples 2 pers. même âge +/- un an	**29,0**	**[28,8;29,2]**	28,9	[28,7;29,1]	28,9	[28,7;29,1]	28,9	[28,7;29,1]	29,0	[28,8;29,2]
Résultats obtenus à partir de variables dont plusieurs ont pu être imputées										
% couples 2 pers. nées en France	**81,4**	**[81,3;81,5]**	**80,8**	[80,7;80,9]	**81,0**	[80,9;81,1]	**81,2**	[81,1;81,3]	81,4	[81,3;81,5]
% couples 2 pers. nées à l'étranger	**7,3**	**[7,2;7,4]**	**7,1**	[7,0;7,2]	**6,9**	[6,8;7,0]	7,3	[7,2;7,4]	7,2	[7,1;7,3]
% couples une pers. née en France et l'autre à l'étranger	**11,4**	**[11,3;11,5]**	**12,1**	[12,0;12,2]	**12,1**	[12,0;12,2]	11,4	[11,3;11,5]	11,4	[11,3;11,5]
% couples 2 pers. nées dans le même département parmi ceux constitués de 2 pers. nées en France	**49,9**	**[49,7;50,1]**	**49,1**	[48,9;49,3]	**46,9**	[46,7;47,1]	**50,2**	[50,0;50,4]	50,1	[49,9;50,3]
% couples 2 pers. nées dans le même pays parmi ceux constitués de 2 pers. nées à l'étranger	**83,7**	**[83,6;83,8]**	**80,9**	[80,8;81,0]	**82,6**	[82,5;82,7]	83,6	[83,5;83,7]	**83,4**	[83,3;83,5]
% pers. françaises avec un conj. français	**98,2**	**[98,2;98,2]**	98,2	[98,2;98,2]	**98,1**	[98,1;98,1]	98,2	[98,2;98,2]	98,2	[98,2;98,2]
% pers. étrangères avec un conj. de même nationalité	**66,3**	**[66,1;66,5]**	**63,3**	[63,1;63,5]	**63,0**	[62,8;63,2]	**66,6**	[66,4;66,8]	66,1	[65,9;66,3]
% pers. étrangères avec un conj. français	**29,9**	**[29,7;30,1]**	**30,6**	[30,4;30,8]	**33,1**	[32,9;33,3]	29,7	[29,5;29,9]	**30,2**	[30,0;30,4]
% pers. même niveau d'études	**56,0**	**[55,8;56,2]**	**55,7**	[55,5;55,9]	**55,0**	[54,8;55,2]	56,2	[56,0;56,4]	56,2	[56,0;56,4]
Champ	**Hors non répondants pour les variables concernées, pondération initiale (avant calage)**		**Complet, pondération initiale (avant calage)**						**Complet, pondération finale (après calage)**	
Effectif pondéré (en milliers)			**32 517**						**35 765**	

Note : Les intervalles de confiance sont calculés sous l'hypothèse d'un tirage aléatoire simple.

La non prise en compte de toutes les corrélations entre variables a des conséquences sensibles sur les résultats de l'imputation. Ainsi, le calcul de la proportion de personnes de nationalité portugaise parmi celles nées au Portugal repose sur deux variables susceptibles d'avoir été imputées : le lieu de naissance et la nationalité de la personne. Parmi les personnes pour lesquelles aucune de ces deux variables n'a été imputée, cette proportion est de 76 %, tandis que parmi les personnes pour lesquelles au moins l'une de ces deux variables a été imputée, elle est de 29 % avec le hot-deck séquentiel, de 11 % avec le hot-deck aléatoire et de 84 % avec le hot-deck métrique.

Conclusion

Finalement, les différentes méthodes d'imputation utilisées ici ont un impact relativement faible, mais non négligeable, sur les résultats de l'étude, notamment s'ils portent sur plusieurs variables imputées et/ou sur de petites populations. Il faut de plus garder à l'esprit que les conséquences des corrections effectuées pourraient être encore beaucoup plus marquées si l'échantillon était plus petit, les taux de non-réponse plus élevés, les variables à imputer plus nombreuses ou les corrélations entre variables plus fortes.

Parce qu'elle permet notamment de tenir compte de l'ensemble des liens entre variables, l'imputation par hot-deck métrique est une méthode particulièrement adaptée au cas où plusieurs variables potentiellement corrélées sont à imputer, et ce quelle que soit la nature (qualitatives ou quantitatives) des variables à imputer et des variables auxiliaires utilisées pour réaliser l'imputation. Mais cette méthode nécessite un temps de calcul important, ce qui la rend difficile à employer dans le cas de gros volumes de données.

Il faut surtout retenir de cette expérience que l'imputation de plusieurs variables qui peuvent être corrélées présente des risques auxquels il faut prêter une attention particulière.

Bibliographie

[1] Caron N., « Les principales techniques de correction de la non-réponse et les modèles associés »,
document de travail série méthodologie statistique Insee, n°9604, 1996 (version actualisée disponible courant 2005).

[2] Dupont F., « Imputation procedures for quantitative and qualitative variables », *document de travail de la direction des statistiques démographiques et sociales Insee*, n°F9406, 1994.

[3] Murthy M. N., Chacko E., Hossain. M., « Condensing and weighting in multivariate nearest neighbour method of imputation ».

[4] Sautory O., « La macro CALMAR – redressement d'un échantillon par calage sur marges », *document de travail de la direction des statistiques démographiques et sociales Insee*, n°F9310, 1993.

[5] Afsa Essafi C., « Les modèles polytomiques non ordonnées : théorie et applications », *document de travail série méthodologie statistique Insee*, n°0301, 2003.