IMPUTATION DE L'ENQUÊTE BUDGET DE FAMILLE 2000

Nicolas CHOPIN (*), Emmanuel MASSE (**)

(*) INSEE, Unité Méthodes Statistiques
(**) Ministère de l'Ecologie et Développement Durable

Introduction

L'enquête « Budget de Famille» est réalisée auprès de plus de dix mille ménages tout les cinq ans. Elle vise à reconstituer toute la comptabilité du ménage (dépenses, consommation et ressources). Les données collectées sont essentiellement monétaires. L'enquête «Budget de Famille » utilise deux instruments de collecte : un questionnaire, qui enregistre les revenus et les dépenses importantes ou régulières sur les derniers mois, et un carnet de dépenses qui enregistre les dépenses quotidiennes sur une période de quatorze jours.

Cette enquête est affectée par de la non-réponse totale et partielle. Le premier type de non-réponse est corrigé classiquement par des méthodes de repondération. Le second type de non-réponse impose l'utilisation de techniques d'imputation. De telles méthodes reviennent à remplacer les valeurs manquantes par des valeurs « admissibles » à la fois au niveau individuel et au niveau agrégé, c'est à dire qui respecte à la fois la structure globale de la variable imputée (moyenne, variance), et la cohérence des relations entre variables pour chaque individu.

Dans cet article, nous présentons les méthodes utilisées pour corriger la non-réponse partielle dans l'enquête Budget de Famille 2000. L'article est organisé comme suit. La première partie décrit les objectifs et la structure de l'enquête Budget de Famille 2000, ainsi que les caractéristiques de la non-réponse affectant cette enquête. La seconde partie rappelle quelques principes généraux sur les techniques d'imputation. Les troisième et quatrième parties décrivent les approches que nous avons suivies pour le traitement de la non-réponse, respectivement dans les questionnaires (techniques de modélisation économétrique) et dans les carnets (méthodes des donneurs). Nous évoquerons en guise de conclusion quelques pistes d'amélioration.

1. Présentation de l'enquête Budget de Famille

1.1. Objectifs de l'Enquête

L'enquête « Budget de Famille » est réalisée par l'INSEE tous les cinq ans. Son but est de reconstituer la comptabilité du ménage : enregistrement de la totalité des dépenses et des ressources.

L'objectif principal de l'Enquête est l'étude des dépenses. On enregistre, pour chaque ménage enquêté, le montant et la nature de ses dépenses, cette dernière étant codée par une nomenclature contenant plus de 900 postes. Notons que toutes les dépenses sont couvertes, y compris celles ne

relevant pas de la consommation de biens et services : impôts et taxes, remboursement de crédits, dons à d'autres ménages, etc.

Cette enquête permet de brosser un panorama relativement exhaustif sur la consommation et le niveau de vie des différentes catégories de ménage, tant sur le plan économique (tendance à long terme et facteurs explicatifs de la consommation) que sociale (études de la pauvreté, des inégalités, etc.)

1.2. Structure de l'Enquête

Le champ de l'enquête recouvre l'ensemble des ménages résidant en France métropolitaine[1]. La collecte est effectuée sur toute l'année (huit vagues de six semaines, ce qui exclue deux semaines début août et deux semaines fin décembre) afin de tenir compte de la saisonnalité du phénomène étudié. Deux instruments de collecte sont utilisés :

- Un questionnaire, qui enregistre les caractéristiques socio-démographiques du ménage, ses revenus et ses dépenses importantes ou irrégulières (biens d'équipement, logement, impôts, etc.).
- Un carnet de dépense, que doit remplir chaque membre du ménage de plus de quatorze ans, et qui enregistre les dépenses faibles ou fréquentes, sur une période de quatorze jours.

2. Principes généraux des méthodes d'imputation

2.1. La notion de non-réponse

On distingue généralement deux types de non-réponse :

- **la non-réponse totale**, qui correspond à l'absence complète d'information sur une unité interrogée.
- **la non-réponse partielle**, qui correspond à une absence d'information limitée à certaines variables de l'unité interrogée.

La non-réponse totale est généralement due à l'impossibilité d'interroger l'unité sélectionnée (dans le cas qui nous intéresse le ménage), soit parce que celle-ci est difficile à joindre, soit parce qu'elle refuse de coopérer, pour des raisons qui lui sont propres. La non-réponse partielle correspond au cas où l'unité accepte d'être enquêtée, mais décide au cours de l'entretien de ne pas répondre à certaines questions, ou de ne répondre que de façon imparfaite. Notons que cette seconde possibilité n'est pas toujours évoquée dans la définition de la non-réponse partielle, mais c'est pourtant un phénomène fréquent. Ainsi, dans l'enquête Budget de Famille, certains individus ont refusé de déclarer leurs revenus exacts, mais ont accepté de dire dans quelle tranche (selon une classification proposée par l'enquêteur) se situait ces même revenus.

La distinction entre non-réponses partielle et totale est moins claire qu'il ne peut sembler au premier abord. Dans certains cas, la non-réponse partielle porte sur un nombre si grand de variables que l'utilisation du peu d'information disponible sur l'unité étudiée porte à caution. On pourra alors préférer ignorer cette information parcellaire, et considérer l'unité correspondante comme affectée de non-réponse totale. Il arrive notamment que certaines personnes interrogées interrompent l'entretien avec l'enquêteur avant sa fin, par lassitude ou manque de temps. Il est courant alors de ne pas exploiter ces entretiens incomplets, et de traiter l'unité interrogée comme « totalement non-répondante ». Ce principe a été appliqué notamment dans l'Enquête Budget de Famille. Plus généralement, on retiendra que la distinction entre non-réponses partielle et totale dépendra en partie de certains partis pris des concepteurs et des analystes d'une enquête donnée.

[1] L'enquête métropolitaine est complétée par une enquête dans les départements d'Outre-Mer, portant sur environ 5000 ménages. Cette note ne traite que du traitement de la non-réponse de l'enquête métropolitaine. Le traitement de l'enquête DOM n'est pas encore effectuée à la date de rédaction de cet article.

2.2. Traitement de la non-réponse

La non-réponse totale est classiquement traitée par des méthodes de repondération. Cela revient à augmenter les poids des unités répondantes, de manière à corriger le biais introduit par la non-réponse totale, voir par exemple [1]. Ce point ne sera pas traité dans ce document.

Après repondération, nous obtenons une base de données couvrant l'ensemble des unités ayant répondu au moins partiellement à l'enquête. Cette base n'est pas directement exploitable, car les techniques d'analyse courante (analyse de données, régressions, etc.) nécessitent l'observation de toutes les variables, pour toutes les unités étudiées. On peut décider de restreindre l'analyse aux unités n'étant pas affectées de non-réponse partielle, mais cela pose deux problèmes :

- La proportion des unités telles qu'au moins une variable n'est pas renseignée est souvent élevée. La perte d'information est donc importante.
- La décision de ne pas répondre à une question donnée peut être corrélée avec certaines caractéristiques socio-économiques de l'unité étudiée. Ne pas tenir compte des unités affectées de non-réponse partielle introduit alors un biais éventuellement élevé.

Il est donc préférable dans la plupart des cas de conserver toutes les unités répondantes dans la base étudiée, quitte à « imputer » les données manquantes. L'imputation consiste à remplacer toutes les variables non renseignées par des valeurs « aussi proches que possibles » des valeurs réelles. Pour clarifier cette notion, nous proposons maintenant différents critères d'évaluation d'une méthode d'imputation donnée. La présentation de ces critères est volontairement non technique. Pour simplifier, nous supposerons qu'une seule variable, notée Y, est affectée de non-réponse, les autres variables, notées de façon générique X, étant renseignées pour tous les individus. Nous appellerons R la variable qui vaut 1 si Y est correctement renseigné, 0 sinon. Cette notation sera reprise dans tout l'article.

- Critères « descriptifs » : les valeurs imputées doivent obéir à la « structure » apparente des données observées. Au niveau global, les valeurs imputées doivent respecter la répartition de la variable Y : l'histogramme de Y obtenu après imputation doit être aussi proche que celui que l'on obtiendrait si toutes les valeurs étaient observées[2]. Au niveau individuel, la valeur imputée doit être cohérente avec les variables observées X : les corrélations observées entre X et Y doivent être prises en compte.
- Critères « inférentiels » : nous rangeons dans cette catégorie tous les critères, plus techniques, relevant de l'impact de l'imputation sur la propriété des méthodes d'estimation utilisées a posteriori. On peut par exemple s'interroger sur le maintien du caractère « sans biais » d'un estimateur donné (tel que l'estimateur des moindres carrés ordinaires) lorsqu'il est appliqué à des données imputées. De plus, l'imputation va généralement se traduire par une augmentation de la variance de cet estimateur. Il faut alors vérifier si cet accroissement reste faible, et s'il peut être évalué facilement.
- Critères de faisabilité : la mise en œuvre d'une méthode donnée dépendra aussi dans la pratique du temps (humain et machine) que l'on est prêt à investir, en égard aux résultats attendus. Ainsi, on n'hésitera pas à utiliser des méthodes « presse-bouton » et peu robustes lorsque la non-réponse n'affecte que très peu d'unités, car l'impact sur les résultats sera sans doute très faible.

Nous décrivons dans la partie suivante deux techniques d'imputation très simples (imputation de la moyenne, méthode du « hot-deck ») et les évaluons à l'aune des critères proposés, afin d'illustrer leur pertinence. Dans les autres parties de cet article, nous reviendrons à ces critères à chaque fois qu'un nouveau principe d'imputation sera présenté.

[2] En revanche, il n'a pas à ressembler à l'histogramme de Y calculée à partir des données uniquement observée car, comme nous l'avons déjà indiqué, ce dernier peut être biaisé, si la décision de ne pas répondre n'est pas indépendante de la valeur prise par Y. L'imputation permet alors éventuellement de redresser ce biais.

2.3. Deux techniques d'imputation simples

L'imputation par la moyenne consiste à remplacer toutes les valeurs manquantes par la moyenne observée de la variable *Y*. Cette méthode est très simple et très rapide à mettre en œuvre (troisième type de critère). Les valeurs imputées ne sont pas aberrantes au niveau global. En revanche, l'histogramme de *Y* sera sans doute trop concentré autour de la moyenne, et l'utilisateur sous-estimera donc la variabilité de *Y* (premier type de critère). D'un point de vue inférentiel, considérons la moyenne empirique calculée sur données complétées. Cet estimateur de l'espérance de *Y* reste sans biais si appliqué sur données imputées, sous l'hypothèse (très forte) que la non-réponse *R* est indépendante de *Y*. Sous la même hypothèse, sa variance est égale à celle de la moyenne calculée sur les données observées, ce qui reste raisonnable. En revanche, on voit bien que l'estimation de la variance de *Y* est forcément biaisée vers le bas. De plus, on peut démontrer que, même sous l'hypothèse mentionnée plus haut (indépendance de *R* et *Y*), l'estimateur des moindres carrés ordinaires de la régression de *Y* sur *X* est systématiquement biaisé. En bref, l'imputation par la moyenne a que peu de propriétés intéressantes, et ne doit être utilisée que lorsqu'un nombre très faible de valeurs doivent être imputées.

L'imputation par hot-deck consiste à remplacer chaque valeur manquante par une valeur tirée aléatoirement[3] parmi toutes les valeurs observées de *Y*. Cette méthode est un peu plus complexe que la précédente (le temps machine sera beaucoup plus élevé que pour l'imputation par la moyenne). La structure globale de *Y* (moyenne, variance) est respectée, sous l'hypothèse d'indépendance de *Y* et de *R*. Si cette hypothèse n'est pas vérifiée, l'histogramme de *Y* devient biaisé. De même, d'un point de vue inférentiel, les estimateurs de type « moyenne des *Y* » ou « estimateur des moindres carrés ordinaires de la régression de *Y* sur *X* » resteront sans biais, sous l'hypothèse déjà mentionnée, mais leurs variances seront fortement augmentées de part la nature même de la méthode utilisée (caractère aléatoire des valeurs imputées).

En résumé, l'imputation par hot-deck apparaît comme plus efficace que l'imputation par la moyenne. Son plus grand défaut cependant est de pas tenir compte des relations entre *X* et *Y*. Les méthodes plus évoluées que nous présenterons dans les parties suivantes permettent de corriger ce défaut.

3. Imputation des questionnaires

3.1. Imputation des revenus

L'enquête Budget de Famille mesure l'ensemble des ressources annuelles perçues au cours des douze mois précédant l'enquête. Les variables dites de revenus enregistrent les montants exacts de ces ressources, pour chaque catégorie (salaire, retraite, épargne, prestations familiales, etc.). La plupart des variables de revenus sont affectées de non-réponse partielle. La proportion de ménages n'ayant pas déclaré au moins un montant exact est de 3,6%. Notons qu'à défaut d'un montant exact, certains ménages ont accepté d'indiquer dans quelle tranche se situait ce montant.

La méthode d'imputation retenue pour ces variables de revenus se base sur une modélisation économétrique. Nous en décrivons le principe dans la partie suivante, et présentons leur mise en œuvre effective pour l'enquête Budget de Famille en 3.1.2.

[3] La probabilité de tirage d'une unité observée est proportionnelle à sa pondération après calage. Si l'échantillon n'est pas pondéré, le tirage est uniforme.

3.1.1. Imputation économétrique

Les techniques d'imputation économétrique reviennent à supposer que la variable affectée de non-réponse *Y* est liée aux variables complètement observées *X* par une relation économétrique simple, de type régression linéaire :

$$Y = Xb + sU \qquad (1)$$

où Xb représente le produit scalaire entre les vecteurs X et b, U est une variable gaussienne centrée réduite, et s est l'écart-type des résidus. Cette modélisation n'est valide que lorsque *Y* est une variable continue. Lorsque *Y* est discrète, on adoptera un modèle linéaire généralisé (logit ou probit par exemple).

Ce modèle étant posé, l'imputation se fait en deux étapes. On estime tout d'abord les paramètres b et s (estimateur par moindres carrés ordinaires dans le cas simple). On remplace ensuite la valeur manquante Y_i du ménage i :

- soit par la moyenne conditionnelle (estimée) $X_i\hat{b}$. On parle alors d'imputation par la moyenne conditionnelle.
- soit par la valeur simulée $X_i\hat{b} + \hat{s}U_i$, où U_i est un tirage aléatoire dans une loi normale centrée réduite. Cette seconde méthode est appelée méthodes des « résidus simulés ».

Le grand intérêt des méthodes économétriques est de pouvoir s'affranchir de l'hypothèse irréaliste d'indépendance entre *R* et *Y* (voir 2.3). La relation économétrique donnée en (1) revient en effet à supposer que *R* et *Y* sont indépendants *conditionnellement à X*, ce qui est une hypothèse beaucoup plus faible[4]. En d'autres termes, on suppose que le biais introduit par la non-réponse non uniformément répartie sur *Y* est entièrement corrigé par la prise en compte de l'information apportée par *X*.

Les méthodes économétriques présentent deux inconvénients. Tout d'abord, elles sont complexes et fastidieuses à mettre en œuvre : l'analyste doit sélectionner les variables explicatives à introduire dans son modèle, voir en construire de nouvelles (croisement d'indicatrices, transformations non linéaires de variables). Il peut aussi appliquer différentes transformations à la variable *Y* (polynomiale, logarithmique, etc.). Le but est en quelque sorte de maximiser la capacité prédictive du modèle (ou de façon équivalente de minimiser la variance du bruit estimé), de manière à tirer le plus d'information possible des variables *X*. A ce titre, on ne cherchera pas à limiter le nombre de régresseurs (principe de parcimonie) comme dans les cas courants d'analyse par régression, car on ne s'intéresse pas à la pertinence du modèle, mais plutôt encore une fois à son pouvoir prédictif.

Le second inconvénient des méthodes économétriques est le caractère « normatif » du modèle paramétrique considéré, qui revient à supposer une certaine régularité des données non forcément vérifiée en pratique. A cet égard, il est souhaitable de s'assurer que les résultats de la régression restent stables lorsque celle-ci est appliquée à des sous-populations données (obtenues par exemple en partitionnant selon une variable socio-démographique, telle que sexe, classe d'âge, etc.). En cas de trop forte instabilité, on préfèrera construire un modèle de régression distinct pour chaque sous-population. Malgré ces précautions, l'hypothèse d'une relation économétrique entre variables reste simplificatrice, et devra au moins être prise en compte dans les analyses ultérieures de la base complétée.

Notons enfin que la méthode des résidus simulés est en général préférée à la méthode de la moyenne conditionnelle, car elle permet une meilleur estimation de la variance de *Y*. Mais comme nous allons le voir dans l'exemple suivant, l'imputation de la moyenne conditionnelle peut s'avérer plus robuste dans les cas où le pouvoir explicatif du modèle reste faible.

[4] L'hypothèse d'indépendance de *Y* et *R* est communément appelée hypothèse MCAR (*missing completely at random*). L'hypothèse d'indépendance conditionnelle est appellée MAR (*missing at random*). La première hypothèse est généralement reconnue comme irréaliste, la seconde est adoptée dans la plupart des travaux traitant de la non-réponse, voir par exemple [2].

3.1.2. Imputation économétrique dans l'enquête Budget de Famille

La méthode précédente a été appliquée à chacune des variables de revenus, de manière indépendante (modèles de régression indépendants pour chaque variable de revenu, avec pour variables explicatives l'ensemble des variables entièrement enregistrées). Il a été décidé de modéliser le logarithme des montants : cette transformation de variable améliore l'ajustement du modèle (R^2 plus élevés) et donne des résultats plus robustes.

Une difficulté de l'imputation des variables de revenus est la possibilité, pour les répondants refusant de déclarer un montant exact, de donner la tranche dans laquelle se trouve le montant. Ceci complique légèrement l'estimation des paramètres de la régression (1). Notons E l'ensemble des individus i déclarant un montant exact y_i, et T l'ensemble des individus i déclarant un numéro de tranche n_i, entier compris entre 1 et K, le nombre de tranches, correspondant à des intervalles $[t_{k=1}, t_k]$, pour k=1,…,K. La vraisemblance du modèle s'écrit alors :

$$L(b,s) = \prod_{i \in E} (2pis)^{-1/2} \exp\left\{-(y_i - x_i b)^2 / 2s^2\right\} \prod_{i \in T} \left[\Phi\left\{(t_{n_i} - x_i b)/s\right\} - \Phi\left\{(t_{n_i - 1} - x_i b)/s\right\}\right]$$

où Φ est la fonction de répartition d'une loi normale.

Dans la pratique, l'estimateur du maximum de vraisemblance correspondant à ce modèle peut être obtenu par la procédure LIFEREG du logiciel SAS. Notons de plus que les montants imputés doivent appartenir à la tranche déclarée par l'individu, lorsque celui-ci décide effectivement de donner cette information. La simulation du résidu doit tenir compte de cette contrainte[5].

Pour la plupart des variables des revenus, nous avons retenu la méthode des résidus simulés. Celle-ci redonne, comme prévu, une meilleure approximation de la variance de la variable imputée. Notons cependant que cette méthode donne dans certains cas des points aberrants aux extrêmes des revenus. Ceci est dû en partie à la modélisation logarithmique : une forte valeur (positive ou négative) de résidu simulé à un effet multiplicatif sur la valeur imputée. Dans le premier cas (résidu très positif), on obtient éventuellement un montant plus important que le montant le plus élevé parmi les déclarés. Dans le second cas (résidu négatif), le montant imputé risque, après arrondi, de prendre une valeur nulle. Ceci remet en cause la robustesse du modèle utilisé.

Le nombre de ces montants imputés « extrêmes » est heureusement très faible (une dizaine sur l'ensemble des revenus), et nous avons pu les corriger manuellement. Nous avons décidé dans la plupart des cas de remplacer la valeur imputée par la moyenne conditionnelle (cela revient à prendre un résidu nul), ce qui donnait des valeurs largement plus satisfaisantes. Il s'est présenté aussi le cas d'une variable de revenu telle qu'un seul individu (parmi les déclarations exactes comme parmi les déclarations en tranche) se situe dans la tranche la plus élevée. Nous avons alors imputé le montant de cet individu par la borne inférieure de cette tranche, car la moyenne conditionnelle était inférieure à celle-ci. D'une manière plus générale, la qualité de l'imputation des revenus élevés doit être vérifiée avec un soin particulier, car, d'une part, ceux-ci vont contribuer fortement à l'estimation de la moyenne des revenus, et d'autre part, le pouvoir prédictif du modèle pour les revenus élevés est sujet à caution, de part le faible nombre de montants élevés parfaitement observés.

[5] Mathématiquement, cela revient à tirer aléatoirement le résidu dans une loi normale tronquée, c'est à dire restreinte à l'intervalle correspondant à la tranche déclarée. Le plus simple est de tirer plusieurs variables gaussiennes centrées réduites, et de retenir la première qui respecte la contrainte voulue, mais il existe des méthodes plus rapides, basés sur le principe d'acceptation/rejet, voir [4], p. 49.

3.2. Imputation des dépenses

Plus de 130 variables de montants de dépenses sont affectées par de la non-réponse, mais toujours dans des proportions très faibles (moins de 1% en moyenne). Devant l'ampleur de la tâche, une approche systématique et volontairement simple a été adoptée. Deux modes d'imputation ont été retenus :

- L'imputation économétrique, dès que le nombre de montants observés est suffisant (pour assurer une certaine qualité de l'estimation) et que le pouvoir explicatif du modèle est satisfaisant (critère retenu : R^2 supérieur à 0.2).

- L'imputation par hot-deck stratifié, dans tous les autres cas.

Nous décrivons dans la partie suivante le principe du hot-deck stratifié, et son application dans l'enquête Budget de Famille.

3.2.1. Imputation par hot-deck stratifié

Le hot-deck stratifié revient à découper la population étudiée en différentes strates, en fonctions de variables socio-économiques déterminées a priori, puis d'appliquer dans chacune des strates une imputation par hot-deck (voir 2.3). L'avantage de cette stratification est d'améliorer, par rapport au hot-deck simple, la cohérence pour une unité donnée entre le montant imputé et les variables socio-économiques renseignées.

Plusieurs études sur la consommation montrent que le déterminant principal de la consommation est le revenu disponible, voir par exemple [3] . Il a donc été décidé de construire dix strates, correspondant aux déciles du revenu total disponible (revenu de l'épargne non inclus, car celui-ci semble principalement réinvesti).

4. Imputation des carnets

4.1. Description de la non-réponse et objectifs de l'imputation

4.1.1. Des taux de non-réponse faible

Le niveau global de non-réponse des carnets est relativement faible : sur les 1 110 284 lignes que comportent l'ensemble des carnets agrégés, seules 6,3 % ne sont pas complètement renseignées lors du codage de la dépense (1,9 % des libellés sont codés sur moins de 4 positions).
Plus précisément, on peut distinguer plusieurs niveaux de non-réponse. Le codage des produits s'effectue à partir d'une nomenclature sur 6 positions. Le logiciel de codage automatique SICORE et les experts tentent à partir de l'information disponible, soit dans le libellé du produit, soit dans le type de magasin, de coder le plus précisément possible les produits achetés. Le tableau 1 montre quelques exemples de codifications incomplètes.

Libellé Produit	Code Produit	Libellé magasin	Code magasin	Montant
Fruits et légumes	011***	Intermarché	1112	18,10
Course	******	Auchan	1111	1226,76
EDF GDF (facture)	045***	EDF	6711	1960,00
Fruits et légumes	011***	Champion	1112	14,60
Fruits et légumes	011***	Champion	1112	13,90
Fruits et légumes	011***	Champion	1112	8,20
Livre revue	095***	Tabac	2224	6,00

Tableau 1 : Illustration de codes produits et codes magasins

Suivant le degré de précision des libellés, les codes produits et magasins sont plus ou moins renseignés. Le tableau 2 donne la répartition du nombre de codes à étoiles. Les positions renseignées apportent une information essentielle qui doit être prise en compte dans les procédures d'imputation.

Libellés	Nombre de libellés à imputer
* * * * * *	5 729
x * * * * *	2 428
x x * * * *	807
x x x * * *	12 355
x x x x * *	30 661
x x x x x *	18 068

Tableau 2 : Répartition des 70 048 libellés à imputer

4.1.2. Une très grande diversité des données manquantes

On observe dans l'échantillon plusieurs types de données manquantes : dans certains cas l'enregistrement non ou partiellement renseigné semble correspondre à un unique bien acheté, comme pour le libellé « Livre revue » avec le montant 6,00 qui a été codé 2224. En revanche, dans le cas du libellé «Course », associé à une dépense de 1226,76 dans un magasin d'alimentation, il est très vraisemblable que ce montant corresponde à l'achat de plusieurs produits (« ticket de course »). Il existe aussi certains cas intermédiaires où il est délicat de savoir si l'enregistrement correspond à un ou plusieurs dépenses, par exemple, le cas du libellé «Fruits et légumes » avec comme montant 18,10.
Le code du magasin ne suffit pas en général pour déterminer la nature de l'ensemble des dépenses. Les achats réalisés dans un magasin d'alimentation peuvent en effet être assez variés et ne pas se limiter simplement à des denrées alimentaires. Les procédures d'imputation doivent nécessairement tenir compte de cette diversité dans les codes à imputer.

4.1.3. Une imputation cohérente aux niveaux micro et macro-économique

Plusieurs méthodes sont envisageables pour l'imputation des carnets. Une approche élémentaire consiste à ne pas tenir compte de l'information partielle disponible (dans les codes produits et magasins) et à repartir pour chaque donnée manquante le montant associé dans l'ensemble des postes de la nomenclature au prorata des dépenses des répondants. Cette méthode présente l'avantage de conserver le montant total des dépenses et la structure par poste de la consommation. Cette approche présente cependant deux inconvénients majeurs :

- les données disponibles ne sont que partiellement exploitées (en particulier les codes produits et codes magasins) ;
- A chaque dépense à imputer, on associe un très grand nombre d'enregistrements avec des montants très faibles, ce qui biaise la structure microéconomique de consommation du ménage. Les montants des codes produits à imputer sont en effet dispersés sur de très nombreux postes, ce qui conduit à observer par exemple des montants très faibles (moins de 1 franc) d'achat de viande.

Cette méthode d'imputation conduit à avoir des données relativement cohérentes au niveau macroéconomique mais difficilement exploitables au niveau microéconomique.

Plusieurs voies sont envisageables pour améliorer la méthode exposée ci-dessus. On peut par exemple tenter de prendre en compte l'information contenue dans les libellés produits partiellement renseignés. Ainsi, dans le cas d'un libellé alimentation, l'allocation se fait en accord avec la structure de consommation en alimentation de l'ensemble des répondants. Même si elle exploite mieux une partie

de l'information disponible, les imputations ainsi réalisées conservent l'inconvénient majeur de ne pas permettre d'exploitation microéconomique fine.

L'objectif est donc de trouver une méthode d'imputation réaliste qui permette d'obtenir à la fois des données cohérentes au niveau macroéconomique (conservation de la structure de consommation des ménages) et exploitable pour les études microéconomiques.

4.2. La méthode des donneurs

L'application de méthodes économétriques est peu adaptée dans le cas où les données à imputer sont des variables polytomiques non ordonnées avec de nombreuses modalités. Le choix s'est donc porté sur la mise en œuvre d'une méthode d'imputation par le plus proche donneur.

4.2.1. Le principe théorique

La méthode des donneurs consiste à remplacer les valeurs manquantes d'une unité affectée de non-réponse (le receveur) par celles correspondant à une autre unité (le donneur), dont les caractéristiques sont proches du receveur. La définition « normative » d'une fonction de distance entre deux enregistrements permet de comparer l'ensemble des donneurs potentiels au receveur. On fait donc l'hypothèse sous-jacente que les receveurs ont en terme de consommation des caractéristiques comparables aux donneurs. Le choix de la fonction distance joue un rôle crucial dans cette approche.

4.2.2. Des traitements semi-automatiques

Dans un premier temps, on applique une série de traitements semi-automatiques. L'expert SICORE est chargé d'isoler les libellés dont l'information est fortement dégradée par codage automatique. Prenons l'exemple des fruits et légumes : ceux-ci sont sur un certain nombre d'enregistrements déclarés simultanément, le carnet ne contenant pas l'information détaillée permettant de savoir si l'achat portait sur des fruits, des légumes ou les deux à la fois. Le code attribué par SICORE est alors 011***. Les méthodes automatiques par le plus proche donneur que nous détaillerons par la suite pourrait conduire à imputer un code commençant par 011***, mais par nécessairement fruits 0116** ou légumes 0117**.
Une première étape d'imputation consiste donc à attribuer aléatoirement dans les proportions de la consommation des répondants soit le code « fruits » soit le code « légumes ».
Ces traitements semi-automatiques ont été réalisés à partir de fichiers fournis par l'expert SICORE. Ils concernent les catégories de dépenses suivantes : épicerie, jardinage, fruits et légumes, lingettes, textile, droguerie pharmacie.

4.2.3. Une mise en application délicate

Après traitements semi-automatiques, il reste 14 871 libellés à traiter pour obtenir une imputation sur quatre positions.
Comme évoqué précédemment, deux catégories de libellés sont distinguées :

1. D'une part les libellés isolés (c'est-à-dire tels que le ménage n'a pas réalisé d'autres dépenses dans le même magasin le même jour). Pour ces enregistrements, on ignore s'ils correspondent à une ou plusieurs dépenses. On peut par exemple avoir le cas d'un ticket qui n'est pas détaillé (la personne interrogée n'ayant fourni que le montant totale de la dépense). Il est également possible d'avoir une unique dépense dont le libellé n'était pas suffisamment précis pour être codé sur au moins 4 positions.

2. D'autre part les libellés qui correspondent à une unique dépense effectuées parmi d'autres par le ménage dans un magasin un jour donné (on vérifie dans le carnet que d'autres dépenses ont été effectuées par le ménage le même jour dans le même magasin). On peut légitiment supposer que chaque enregistrement correspond à une seule dépense.

Il faut alors distinguer plusieurs étapes dans la mise en place de la méthodes d'imputation :

1. Création des tables des receveurs. La première table de receveurs est constituée des enregistrements tels que le libellé n'a pas pu être codé sur quatre positions, et tels que le ménage n'ait aucune autre dépense le même jour dans le même magasin (il s'agit a priori soit de tickets soit de dépenses isolées). La seconde table de receveurs correspond aux dépenses (non codées sur quatre positions) effectuées parmi d'autres le même jour dans le même magasin, par le ménage.

2. De façon similaire, on constitue deux tables de donneurs qui correspondent cette fois aux libellés renseignés sur six positions. La première table est constituée de l'agrégation de l'ensemble des dépenses effectuées par un ménage dans un magasin un jour donné. Les montants de dépenses sont cumulés et on conserve le code produit qui est majoritaire en montant. L'idée est de constituer une table de donneur qui correspondent à des tickets (à un même donneur est associé plusieurs enregistrements). La deuxième table est constituée de l'ensemble des dépenses renseignée sur 6 positions et qui ne sont pas isolées.

3. Une fois constituée les tables des receveurs et des donneurs, il suffit de définir les deux fonctions distance.

Pour les libellés de second type, c'est-à-dire ceux correspondant à une dépense unique parmi d'autres (il ne peut a priori donc par s'agir de ticket), le donneur est choisi dans l'ensemble des unités dont les caractéristiques suivantes sont identiques à celles du receveur :

- catégorie de commune (communes de plus ou de moins de 200 000 habitants);
- tranche d'âge de la personne de référence (moins de 35 ans, de 35 à 60 ans, plus de 60 ans);
- code produit de la partie renseignée ;
- code magasin de la partie renseignée ;

Parmi ces unités, on choisit pour donneur celui qui minimise l'écart entre les montants déclarés (avec tirage éventuel en cas d'ex æquo).

Dans le cas des libellés du premier type (ceux correspondant éventuellement à des tickets), l'échantillon de donneurs potentiels est défini par correspondance avec les caractéristiques suivantes du receveur :

- catégorie de commune (voir plus haut);
- tranche d'âge de la personne de référence (voir plus haut) ;
- code produit majoritaire en montant pour la partie renseignée ;
- code magasin pour la partie renseignée ;

Parmi ces unités, on choisit pour donneur celle qui minimise l'écart entre les montants des tickets déclarés (avec tirage éventuel en cas d'ex æquo).

Dans le cas des tickets (receveurs et donneurs de type deux), une fois le donneur sélectionné, on remplace l'enregistrement à imputer par l'ensemble des dépenses associées au ticket du donneur. Les montants des enregistrements du ticket donneur sont modifiés de sorte à les ajuster au montant de l'enregistrement du receveur.

	Vagues 1 et 2	Vagues 3 et 4	Vagues 5 et 6	Vagues 6 et 7	Total
Enregistrements (fichier initial)	267 151	293 537	293 976	295 950	1 150 614
Enregistrements (parmi les répondants)	258 022	283 978	283 549	284 735	1 110 284
Enregistrements à imputer	4 750	4 990	5 130	5 466	20 336
Enregistrements à imputer (après traitement semi-automatique)	3 479	3 665	3 722	4 005	14 871
dont enregistrements isolés à imputer (tickets) : type 2	1 313	1 514	1346	1 454	5 627
dont enregistrements au sein d'un ticket à imputer : type 1	2 166	2 151	2376	2 551	9 244
Nombre final d'enregistrements	263 528	289 778	288468	291 634	1 133 408

Tableau récapitulatif des imputations

4.2.4. Des difficultés informatiques

L'ensemble de la procédure d'imputation a été écrite sous SAS à l'aide de la proc SQL. L'idée principale est de réaliser un produit cartésien (fusion sans contrainte) des tables donneurs et receveurs. Il est alors aisé de calculer l'ensemble des valeurs pour la fonction distance et de sélectionner la plus faible par une sous-réquête.
Le principal problème résulte de la taille de la table obtenue par le produit cartésien. En effet, pour les enregistrements à imputer au sein d'un ticket (type 1), si l'on effectue l'opération sur l'ensemble de la table, on obtient une table d'environ 10 000 000 000 de lignes, ce qui est excessif pour les capacités d'un ordinateur de bureau. Pour diminuer le temps nécessaire, les opérations ont été découpées en quatre étapes (par bloc de deux vagues). Ce découpage présente également l'avantage de fournir des donneurs à des périodes comparables de l'année . Les stratifications permettent également de diminuer la taille de la table produit.
En définitif, sur un ordinateur de bureau, l'imputation de l'ensemble des carnets prend environ 10 heures.

4.2.5. Le codage à 6 positions

La méthode explicitée précédemment permet d'imputer l'ensemble des dépenses sur 4 positions. Le passage à 6 positions s'effectue simplement en attribuant aléatoirement un code à 6 positions à l'ensemble de la dépense et cela au prorata du montant des dépenses parmi les répondants.

4.3. Des résultats satisfaisants

4.3.1. Au niveau macroéconomique

Le nombre de valeur à imputer étant relativement faible (par rapport au nombre d'enregistrements des carnets), l'impact au niveau macroéconomique est peu important. La méthode utilisée est très proche d'un hot-deck stratifié. Théoriquement, si les non répondant ont par strate la même structure de dépense que les répondants, l'imputation donne des estimateurs non biaisés de la moyenne et de la variance[6]. La prise en compte du donneur le plus proche en terme de montant permet de conserver une certaine cohérence des prix des produits.

[6] L'utilisation d'une fonction distance ne permet pas d'affirmer que le biais du à l'imputation est nul, néanmoins, dans la mesure où le donneur est toujours sélectionné dans la même strate que le receveur, on peut supposer que le biais est faible.

4.3.2. Au niveau microéconomique

La méthode utilisée est particulièrement efficace au niveau microéconomique, elle permet de conserver des types de dépenses et des montants crédibles. La distinction de deux types d'imputation permet de convenablement imputer les différentes sortes de non-réponse (tickets, dépenses isolées, libellés ambigus). Dans le cas de l'imputation de tickets, par exemple, le choix d'un donneur lui-même sélectionné parmi les tickets (ensemble des dépenses agrégées un jour, dans un magasin, par un ménage) complètement remplis permet d'imputer convenablement la structure des dépenses du ménage.

5. Conclusion

L'hétérogénéité des sources (carnets et questionnaires) et des variables à imputer (revenus, montants des dépenses) a imposé le recours à plusieurs méthodes d'imputation : résidus simulés, hot-deck stratifié, et méthode des donneurs. L'objectif était de respecter une certaine cohérence des données, tant au niveau global (moyenne, variance) qu'au niveau individuel (liens entres variables). Une amélioration envisageable est l'automatisation des procédures utilisées (écriture de macros SAS), de manière à simplifier l'imputation de prochaines enquêtes.

Bibliographie

[1] Deville J.C., et Särndall C.E., «Calibration estimators in survey sampling", *Journal of the American Statistical Association*, vol 87, n° 11, pp 376-382, 1992.

[2] Rubin, D.B., « Multiple imputation for nonresponse in surveys », *Wiley*, 1987.

[3] Herpin, N. et Verger, D., « La consommation des Français », La Découverte, 2000.

[4] Rober, C.P., « Monte Carlo Statistical Methods », Springer-Verlag, 1999.