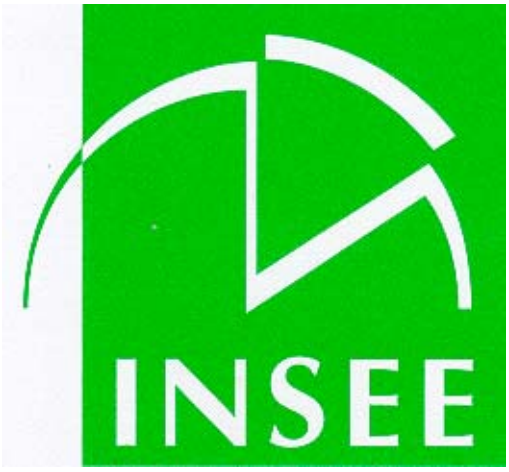
Direction des Statistiques Démographiques et Sociales

N° F0901

La non-réponse partielle aux variables financières de l'enquête Logement 2006 : mise en oeuvre de nouvelles procédures de redressement et comparaison de méthodes d'imputation

Sophie O'PREY

Document de travail



Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail
de la

DIRECTION DES STATISTIQUES DEMOGRAPHIQUES ET SOCIALES

Division Logement

N°F0901

La non-réponse partielle aux variables financières de l'enquête Logement 2006 : mise en œuvre de nouvelles procédures de redressement et comparaison de méthodes d'imputation

Sophie O'PREY
Division Logement

Résumé

Les informations à caractère financier de l'enquête Logement font l'objet d'un certain nombre de non-réponses (« Refus » ou « Ne Sait Pas ») et de valeurs aberrantes. Il est nécessaire de les corriger en aval de la collecte avant toute exploitation des variables concernées car la non-réponse partielle pose des problèmes en termes de biais et de précision des estimateurs. Les redressements exposés dans ce rapport concernent des variables qui présentent des particularités différentes. D'une part, on a cherché à redresser la mauvaise déclaration dont les placements financiers et les charges locatives avaient fait l'objet. D'autre part, pour les revenus - qui pouvaient être renseignés en clair ou en tranches - on a employé des méthodes d'imputation permettant de tenir compte de l'information en tranches disponible.

En prolongement de ce travail, on a comparé certaines méthodes d'imputation entre elles. Du point de vue de la distribution des variables redressées, on constate peu de différences lorsque les domaines étudiés sont suffisamment grands car les taux de non-réponse sont dans l'ensemble faibles et l'échantillon de taille importante. Par contre, les estimations sont de moins bonne qualité lorsque l'on travaille sur de petites sous-populations, en particulier pour des ratios susceptibles de comporter des variables imputées au numérateur et au dénominateur comme les taux d'effort. Il est alors préférable d'utiliser des méthodes d'estimation sur petits domaines ou du moins de tenir compte des intervalles de confiance des estimateurs, notamment si l'on souhaite effectuer des comparaisons temporelles ou entre des sous-populations.

Mots-clefs : non-réponse partielle, méthodes d'imputation, imputation par hot-deck, imputation par la méthode des résidus simulés, information en tranches.

Abstract

The Insee Housing survey contains financial information in which appear a number of missing values - "Refusal" or "Do not know" - and outliers. It is necessary to correct these variables before processing them because non-response will produce a bias in estimation and reduce the accuracy of estimators.
The procedures described in this document concern two types of variables possessing certain particularities. First of all, figures for financial investments and rent had been under-estimated and so had to be corrected. Secondly, with the question about income it proved necessary to resort to imputation methods taking interval-censored data into account, since it was possible to answer in two ways, by giving an exact figure or else by indicating the interval in which it was comprised.

Further to this study, some of the imputation procedures were compared. As regards the distribution of the corrected variables, there proved to be little difference in cases where areas of interest were large enough because of the generally low non-response rates and the sufficiently large sample size. However estimations for small sub-populations were of poorer quality, in particular as regards ratios which featured imputed variables both in the numerator and denominator, such as housing budget share. In this case it is preferable to use small area estimation or at any rate to take into account the confidence intervals of estimators, especially if one wishes to make comparisons in time or between sub-populations.

Keywords: partial non-response, imputation procedures, hot-deck procedures, interval-censored data.

Sommaire

Introduction

Le questionnaire de l'enquête Logement comporte un aspect financier et budgétaire important. Parmi ses principaux objectifs figurent en effet l'évaluation des dépenses consacrées au logement et de leur poids dans le budget des ménages (taux d'effort), ainsi que l'étude des conditions de logement selon le niveau de vie. Or, si les enquêtés sont généralement enclins à parler de leur cadre de vie, les informations à caractère financier sont plus difficiles à obtenir : elles peuvent faire l'objet de non-réponses (« Refus » ou « Ne sait pas ») ou de valeurs aberrantes.

Quelle que soit l'origine de la non-réponse partielle, il est nécessaire de la corriger en aval de la collecte avant toute exploitation des variables concernées car elle pose des problèmes en termes de biais et de précision des estimateurs. Ces derniers peuvent être biaisés si l'on se restreint aux seuls répondants, dans la mesure où la décision de ne pas répondre à une question est souvent corrélée avec des caractéristiques socio-économiques du ménage. De plus, les non-réponses diminuent la fiabilité des résultats puisqu'on perd de l'information : la taille du sous-échantillon exploitable pour la variable d'intérêt est plus petite que s'il n'y avait pas eu de non-réponse partielle, donc la variance des estimateurs est plus grande.

A l'occasion de ce travail, on a procédé à un approfondissement méthodologique qui a consisté à chercher les meilleures méthodes d'imputation possibles pour certaines variables financières de l'enquête Logement 2006, compte-tenu de l'information disponible. Les redressements exposés dans ce document concernent des variables présentant des particularités différentes : d'une part, les placements financiers et les charges locatives, qui ont fait l'objet d'une mauvaise déclaration ; d'autre part, les salaires et les allocations RMI, qui pouvaient être renseignés en clair ou en tranches. Par rapport aux enquêtes précédentes, on a été amené à mettre en œuvre de nouvelles méthodes, soit parce que les questions ont été modifiées en 2006, soit parce que l'on souhaite prendre en compte davantage d'information.

On a ainsi cherché à redresser la sous-déclaration donc les questions sur la *détention de placements financiers* et le *paiement de charges locatives* ont fait l'objet. En effet, comme les « Ne sait pas » et « Refus » n'étaient pas autorisés à ces questions, certaines réponses négatives correspondent en réalité à des refus de répondre ou à une méconnaissance de la part du ménage. L'objectif des redressements est donc non seulement d'imputer les montants manquants mais aussi de corriger cette mauvaise déclaration, ce qui nécessite de procéder en deux temps : il faut d'abord corriger la sous-déclaration, puis imputer les montants non renseignés. Pour les placements financiers, on doit effectuer une imputation simultanée des types d'actifs détenus et de la tranche de montant associée pour les ménages auxquels on affecte la détention de placements.

En ce qui concerne les *revenus*, il faut utiliser de nouvelles méthodes de redressement en 2006 car des questions en tranches ont été ajoutées en cas de non-réponse aux questions en clair : on doit donc avoir recours à des méthodes de correction de la non-réponse qui permettent de prendre en compte l'information éventuellement apportée par les tranches. Dans le cas des revenus individuels, on est confronté à une difficulté supplémentaire qui est liée à la manière dont les questions sont posées : alors que le montant demandé en clair concerne le revenu de l'individu, la question posée en tranches porte sur le montant perçu par l'ensemble des membres du ménage, donc on doit traiter de l'information qui est disponible à deux niveaux.

En prolongement de cet approfondissement méthodologique, on compare certaines méthodes d'imputation entre elles du point de vue de la distribution des variables redressées. On cherche à voir si elles ont des effets sur des variables agrégées, c'est-à-dire au niveau macroéconomique, ainsi que sur de petites sous-populations. On s'intéresse aussi à leur impact sur des ratios susceptibles de comporter des variables imputées au numérateur et au dénominateur : les taux d'effort, qui visent à mesurer le poids des dépenses consacrées au logement dans le budget des ménages.

Le document est organisé selon le plan suivant :

- on présente dans une première partie l'enquête Logement 2006 et les variables concernées par la non-réponse partielle.

- la deuxième partie concerne la méthodologie adoptée pour corriger la non-réponse partielle : elle décrit les différentes étapes des redressements et présente les méthodes d'imputation utilisées pour l'enquête Logement 2006.

- l'application de ces méthodes fait l'objet de la troisième partie, dans laquelle sont exposés les redressements des placements financiers et des charges locatives d'une part, et les redressements des salaires et des allocations RMI d'autre part.

- enfin, on compare certaines des méthodes d'imputation utilisées (méthode des résidus simulés et hot-deck aléatoire par classes) du point de vue de leur impact sur la distribution de variables agrégées et sur la distribution des salaires pour des sous-populations. On étudie aussi leur effet sur les taux d'effort.

****

Ce travail a fait l'objet d'un mémoire de FCDA réalisé entre mars 2007 et janvier 2008. Les tables recodifiées et pondérées de l'enquête Logement 2006 n'ont été disponibles qu'à partir de la mi-février 2008. Pour les travaux présentés dans ce document, on a donc utilisé une pondération provisoire (seuls certains résultats ont été calculés avec les poids définitifs ; ils sont alors signalés).

Je tiens à remercier Nathalie Caron (DEPP) pour ses conseils.

Première partie : Présentation de l'enquête Logement 2006 et des variables concernées par les redressements

I. PRÉSENTATION DE L'ENQUÊTE LOGEMENT 2006

L'enquête Logement est l'une des principales enquêtes ménages de l'Insee par la taille de son échantillon, son ancienneté (depuis 1955) et la portée de ses résultats : au niveau national, elle est une source d'information très riche pour décrire le parc de logements et les conditions d'occupation par les ménages de leur résidence principale, car elle est plus complète et précise que les recensements sur ces thèmes. Enfin, l'un des points forts des enquêtes Logement est leur régularité (elles ont lieu tous les quatre ou cinq ans) et leur forme similaire, ce qui permet des comparaisons temporelles.

La collecte de la dernière enquête Logement s'est déroulée en six vagues du 27 février au 23 décembre 2006 (la précédente enquête avait eu lieu en 2002). L'enquête Logement 2006 a été réalisée par sondage auprès de 65 000 logements dont 57 000 en France métropolitaine. Le recueil des informations s'effectuait sur une seule visite, avec une durée moyenne d'entretien évaluée à un peu moins d'une heure. Seuls les ménages qui étaient dans leur résidence principale ont été enquêtés. Au sein du ménage, les enquêteurs devaient interroger une personne du groupe de référence afin que le répondant soit en mesure de répondre à l'ensemble des questions et en particulier, celles relatives à des montants (loyers, remboursements d'emprunts, revenus…). Parmi les ménages appartenant au champ de l'enquête (c'est-à-dire ceux qui ont été enquêtés dans leur résidence principale), près de huit sur dix ont accepté de répondre au questionnaire, soit un peu plus des deux tiers de l'ensemble des ménages échantillonnés[1]. Au total, un peu plus de 39 000 ménages ont répondu à l'enquête en France métropolitaine.

II. L'ENJEU DU RECUEIL DES INFORMATIONS FINANCIÈRES DANS L'ENQUÊTE LOGEMENT

L'enquête Logement cherche à répondre à plusieurs grandes questions :
- décrire les conditions de logements objectives (confort, peuplement…) et subjectives des ménages (appréciation sur le logement, l'immeuble, le quartier, souhaits de mobilité…)
- analyser les comportements (accession à la propriété…)
- aider à l'évaluation et à l'orientation de la politique du logement (développement de logements sociaux, attribution d'aides au logement, efficacité des prêts à taux zéro…)
- évaluer les dépenses en logement et leur poids dans le budget des ménages (taux d'effort)

Pour y apporter des éléments de réponse, le questionnaire aborde les thèmes suivants :
- caractéristiques physiques des logements
- qualité de l'habitat
- modalités juridiques d'occupation du logement
- difficultés d'accès au logement, solvabilité des ménages, fonctionnement des rapports locatifs
- dépenses associées au logement et aides dont bénéficient les occupants
- ressources perçues par les différents membres du ménage
- patrimoine en logements des ménages, utilisation de logements autres que la résidence principale
- mobilité résidentielle des ménages
- opinion des ménages à l'égard de leur logement et désir éventuel d'en changer

De nombreuses données financières sont donc recueillies. L'enquête permet de connaître précisément les dépenses effectuées par le ménage pour se loger, c'est-à-dire les montants des loyers, des charges locatives ou de copropriété, le prix et le financement des logements achetés récemment, les remboursements d'emprunt des accédants à la propriété ainsi que les consommations d'eau et d'énergie. L'allègement des dépenses que procurent à leurs bénéficiaires les aides publiques au logement (AL, APL) est également mesuré. Le ménage est par ailleurs interrogé sur les ressources qu'il

[1] Taux de collecte = nombre de répondants / nombre de fiches-adresses = 66,2 % (65,3 % en France métropolitaine).
Taux de réussite = nombre de répondants / nombre de résidences principales = 77,8 % (76,7 % en France métropolitaine).

perçoit (revenus d'activité, de placements, prestations sociales…), ce qui permet d'évaluer son revenu global.

A la lumière des études menées sur les précédentes enquêtes Logement, on peut isoler trois grands objectifs liés à la collecte de ces informations financières.

2.1. Mesurer les dépenses liées au logement

L'enquête permet de connaître les dépenses liées au logement dans l'habitat individuel et collectif, parmi les locataires et les propriétaires (charges, loyers, remboursements d'emprunts…). Elle vise aussi à mesurer les dépenses d'énergies suivant le type de combustible utilisé. Une attention particulière est accordée aux achats récents de résidences principales et à leurs modalités de financement (description des prêts, de l'apport personnel…) : on peut ainsi mettre en rapport le coût des logements avec leur qualité et savoir comment a évolué le marché de l'immobilier depuis la précédente enquête Logement.

2.2. Étudier les conditions de logement selon les revenus

Un autre objectif de l'enquête est d'évaluer le revenu global annuel de chaque ménage et la dispersion réelle des sommes perçues, ce qui permet d'étudier les conditions de logement de certaines catégories de population définies à partir des tranches de revenus (ménages pauvres par exemple). Il est donc important d'avoir à l'issue des imputations une distribution des revenus aussi proche que possible de la réalité. Le revenu total peut également être rapporté au nombre d'unités de consommation du ménage afin de comparer les variables liées au logement en référence à des tranches de revenus homogènes (niveau de vie).

2.3. Évaluer le poids des dépenses de logement dans le budget des ménages : le « taux d'effort »

Enfin, en rapportant le montant des dépenses que les ménages effectuent pour se loger à celui de leurs revenus, l'enquête permet de mesurer le taux d'effort, c'est-à-dire la part des dépenses consacrées au logement. Grâce aux taux ainsi calculés, il est possible de comparer l'effort financier consenti par différentes catégories de ménages pour leur habitation (accédants à la propriété par rapport aux locataires, locataires en HLM par rapport aux locataires en secteur libre…). Cela permet aussi de voir quels ménages ont besoin d'aides personnelles et de quel montant.

Pour pouvoir répondre à ces objectifs, il est donc important de disposer de variables financières bien renseignées. Or elles font l'objet de non-réponses partielles qu'il faut corriger avant de pouvoir les exploiter.

III. LA NON-RÉPONSE PARTIELLE AUX VARIABLES FINANCIÈRES DE L'ENQUÊTE LOGEMENT

3.1. La notion de non-réponse

Comme toute enquête ou recensement, l'enquête Logement 2006 est affectée par de la non-réponse, qui se manifeste sous deux formes différentes : non-réponse totale et partielle.

La non-réponse totale correspond à une absence complète d'information pour un ménage échantillonné appartenant au champ de l'enquête. Il s'agit d'une non-réponse chronique à l'ensemble des questions : aucune variable n'est renseignée. On la rencontre dans les cas où l'enquêteur n'a pu interroger le ménage sélectionné parce que ce dernier était injoignable, a refusé de répondre au questionnaire ou n'a pu le faire (en raison de problèmes de santé, linguistiques…), ou encore a abandonné au début de l'entretien. Dans l'enquête Logement, la non-réponse totale a été corrigée par repondération : on a augmenté les poids des unités répondantes de manière à corriger le biais qu'elle introduit.

La non-réponse partielle est une non-réponse ponctuelle à une ou quelques questions pour un ménage échantillonné. L'absence d'information est limitée à certaines variables du questionnaire : l'enquêté a accepté de répondre à l'enquête mais n'a pas su ou voulu répondre à une partie des questions

(montants des revenus par exemple)[2]. Pour corriger la non-réponse partielle, on recourt à des techniques d'imputation, qui consistent à créer des valeurs artificielles aussi proches que possible des valeurs réelles.

Le travail de redressement de la non-réponse partielle dont il est question dans ce document est effectué en aval de la collecte, après la constitution des fichiers contenant les données collectées. Cependant, avant d'aborder les différentes méthodes de correction de la non-réponse partielle, il est utile de s'interroger sur ses origines possibles et sur les moyens de l'éviter, car la meilleure façon de traiter les non-réponses est encore de ne pas en avoir…

3.2. L'origine des non-réponses partielles à l'enquête Logement 2006

Les enquêteurs ont signalé lors des tests que les enquêtés parlent en général volontiers de leur cadre de vie, notamment lorsqu'ils sont propriétaires. Les « Ne sait pas » et « Refus » sont cependant autorisés à la plupart des questions relatives à des montants. En effet, certains ménages refusent de répondre à des questions qui leur semblent indiscrètes ou sans rapport avec le thème de l'enquête, comme celles portant sur le montant de leurs placements financiers. Il arrive aussi que les enquêtés ne connaissent pas la réponse ou ne disposent pas des documents qui leur permettraient de répondre (relevés annuels de charges par exemple). Enfin, certaines questions posent des problèmes de compréhension car elles sont plus complexes ou techniques ; en particulier, la description des prêts nécessite de savoir lire un tableau d'amortissement.

En pratique, on assimile aussi à des non-réponses partielles les réponses invraisemblables (montants anormalement élevés ou faibles, ou incohérents au regard des autres informations recueillies) sauf s'il est possible de les corriger (erreur de zéro par exemple).

Sont également considérées comme des non-réponses partielles les valeurs « imparfaites » : elles concernent les ménages qui n'ont pas donné le montant exact de leurs revenus mais ont accepté de les situer dans une tranche[3]. Ces réponses imprécises apportent toutefois une information supplémentaire importante : on utilisera des méthodes d'imputation adaptées afin d'en tenir compte lors de la correction de la non-réponse partielle à la question en clair.

Il arrive que la non-réponse partielle porte sur un trop grand nombre de variables pour que le questionnaire soit exploitable : il est alors préférable de ne pas le conserver puisqu'il ne contient pas suffisamment d'informations, ce qui revient à le traiter comme s'il avait fait l'objet d'une non-réponse totale. Dans l'enquête Logement, ont été assimilés à des non-réponses totales l'ensemble des questionnaires interrompus avant la fin de l'entretien (pour des raisons de lassitude ou de manque de temps du ménage, ou encore parce certaines questions lui semblent trop indiscrètes), car on a estimé qu'ils étaient insuffisamment remplis.

3.3. La prévention de la non-réponse partielle

Pour réduire les sources de valeurs manquantes et aberrantes, donc limiter le recours à des méthodes de redressement en aval de la collecte, une attention particulière a été portée à la prévention de la non-réponse partielle lors de la préparation de l'enquête (cf. annexe 1). Le questionnaire a été testé plusieurs fois sur le terrain afin entre autres de s'assurer que les questions posées étaient suffisamment claires et précises. Pour éviter les erreurs de saisie ou de compréhension des questions, de nombreux contrôles de cohérence programmés sous CAPI se déclenchent lors de l'entretien en cas de situation atypique : l'enquêteur ainsi alerté peut alors corriger la réponse s'il s'agit d'une erreur, ou éventuellement rédiger un commentaire explicatif s'il s'agit d'un cas particulier. Toutefois, ces moyens de prévention ne dispensent pas de l'utilisation de méthodes d'imputation car on rencontre à l'issue de la collecte un nombre non négligeable de non-réponses partielles.

[2] Ne sont bien-sûr pas prises en compte les non-réponses à des questions filtrées non posées.
[3] En cas de non-réponse aux questions en clair sur le montant des revenus, des tranches étaient proposées.

3.4. Les variables de l'enquête Logement concernées par les redressements

Au total, une vingtaine de variables de l'enquête Logement doivent être redressées, relatives à quatre grands thèmes :

- les revenus (perçus au niveau individuel et du ménage)
- les ménages locataires et logés gratuitement
- les propriétaires
- les dépenses en eau et en énergies

La proportion de non-réponse est plus ou moins importante selon les variables : le taux de non-réponse varie de 0,5 % pour les loyers à 24 % pour les revenus non salariaux (cf. annexe 2). Les variables concernant les dépenses en logement sont dans l'ensemble mieux renseignées que celles portant sur les revenus. Les taux de non-réponses observés sont pour la plupart très proches de ceux relevés lors de l'enquête Logement de 2002.

Les redressements dont il est question dans ce document concernent des variables qui présentent des caractéristiques particulières et pour lesquelles on a mis en œuvre de nouvelles procédures par rapport aux enquêtes précédentes :

- les *placements financiers* et les *charges locatives* ont fait l'objet d'une mauvaise déclaration que l'on corrige en effectuant des redressements en deux temps. Pour les placements financiers, on procède également à l'imputation simultanée de deux variables liées entre elles : les types de placements détenus et le montant correspondant.
- à l'enquête 2006, des questions en tranches ont été ajoutées en cas de non-réponse au montant en clair des revenus : c'est notamment le cas des *salaires* et des *allocations RMI*. Pour corriger la non-réponse aux montants en clair, on doit utiliser des méthodes d'imputation permettant de mobiliser l'information apportée par les tranches.

Deuxième partie : Méthodologie des redressements

Le travail de redressement consiste en premier lieu à établir l'ordre dans lequel effectuer les imputations. En effet, une fois une variable redressée, elle participe éventuellement à l'imputation des autres variables en temps que variable auxiliaire : on utilise de l'information auxiliaire en partie artificielle (les valeurs imputées sont alors supposées vraies). Les redressements des différentes variables se déroulent en plusieurs étapes :

- vérification des données
- imputation des valeurs manquantes
- vérification de la cohérence du fichier imputé

I. PREMIÈRE ÉTAPE DES REDRESSEMENTS : LA VÉRIFICATION DES DONNÉES

Avant de procéder aux imputations, une étape préalable d'analyse et de validation des données collectées est nécessaire pour s'assurer que les informations recueillies sont exploitables. Il faut donc veiller à ce que les données soient homogènes : par exemple, on doit ramener les charges à leur équivalent annuel, ou encore réévaluer les prix des logements à leur valeur en 2006 à l'aide des indices notaires. On peut ensuite procéder à la vérification de la cohérence des données recueillies.

1.1. Comment détecter les montants douteux ?

Pour repérer les valeurs suspectes, on procède pour chaque variable financière à des analyses exploratoires univariées et multivariées. Pour les salaires par exemple, on a édité la distribution par catégorie sociale, sexe et temps de travail afin de comparer les valeurs extrêmes aux autres valeurs très élevées ou faibles : si la valeur la plus élevée est de très loin supérieure à la deuxième valeur la plus élevée, il y a des chances pour que ce soit une erreur. On peut en effet avoir affaire à des montants qui ne sont pas extrêmes en tant que tels mais sont très élevés au vu de la catégorie sociale par exemple. Avant de supprimer sans autre vérification ces valeurs suspectes pour les remplacer par des montants imputés, on effectue un contrôle de cohérence interne des questionnaires concernés (micro-contrôles) : on regarde si ces valeurs peuvent s'expliquer par d'autres caractéristiques de l'individu concerné (âge, CDD ou CDI…) ou si l'enquêteur a émis une remarque pour les justifier. Ce travail peut être long car les valeurs extrêmes aberrantes ont un effet masquant : en les mettant de côté, on en découvre d'autres. On peut aussi comparer les valeurs extrêmes à des seuils minimum et maximum issus de sources externes, en se référant à des barèmes (prestations familiales, allocations logement…) ou à d'autres enquêtes ménages : par exemple, pour avoir un ordre de grandeur des seuils maximum de salaires par catégorie sociale, on s'est référé à l'enquête Revenus Fiscaux 2005[4].

Des erreurs peuvent cependant subsister au terme de ces contrôles « manuels ». Pour les principales variables d'intérêt qui sont modélisables (salaires par exemple), on a ajouté une étape de vérification supplémentaire : on effectue une régression robuste, qui permet de détecter les anomalies influentes sur la régression[5], et on confronte ses résultats à ceux de l'analyse exploratoire réalisée précédemment. Elle permet de repérer des valeurs suspectes supplémentaires ou de confirmer le diagnostic de certains montants douteux.

Au terme de ces « contrôles qualité » se pose la question du traitement des anomalies détectées. Selon que les valeurs suspectes correspondent à une erreur (valeur aberrante) ou à un cas très particulier (valeur atypique), on procède différemment.

1.2. Les valeurs aberrantes sont corrigées ou supprimées

Les valeurs aberrantes sont des montants invalides qui correspondent à une erreur de mesure : elles sont impossibles en tant que telles ou incohérentes au vu des autres réponses données par l'enquêté ou des réponses des autres ménages. Le plus souvent, elles proviennent d'une erreur de saisie (zéro en trop ou en moins, erreur de décimale) : en effet, il n'y a pas de contrôles sous CAPI pour toutes les

[4] On peut également confronter les effectifs : par exemple, on a comparé le nombre de ménages ayant déclaré disposer de placements financiers avec celui issu de l'enquête Patrimoine 2004.

[5] On utilise ici la régression robuste comme méthode de détection et non de modélisation, c'est-à-dire pour sa capacité à repérer les observations atypiques et aberrantes (« outliers »). Cf. annexe 3.

questions et, même en présence d'un message d'avertissement, il arrive que l'enquêteur lise trop rapidement le message et confirme sa réponse à tort (si le salaire donné est déjà élevé par exemple). Les valeurs aberrantes peuvent également être dues à une mauvaise compréhension de la question par l'enquêté, à une confusion mensuel/annuel ou Francs/Euros, ou tout simplement à une erreur (s'il répond de mémoire ou lit mal un document). Les montants très faibles peuvent aussi correspondre à des refus déguisés (sous-déclaration volontaire du salaire par exemple).

Lorsque c'est possible, on corrige les valeurs aberrantes manuellement (erreur flagrante de zéro par exemple). Sinon on les efface[6] pour les imputer au même titre que les valeurs non renseignées car elles ont un effet distordant donc biaiseraient les estimations : on génère de la non-réponse partielle postérieurement à la collecte[7].

1.3. Les valeurs atypiques sont conservées mais ne participent pas à l'imputation

Les valeurs atypiques sont des montants corrects mais rares. On les rencontre par exemple dans le cas d'un millionnaire ou, plus couramment, dans celui d'une personne qui vient de commencer à travailler et dont le montant des salaires perçus au cours de l'année écoulée est par conséquent très faible. Il faut donc les conserver mais on ne les prend pas en compte lors de l'imputation pour ne pas donner trop de poids à ces cas très particuliers (dans le cas d'une imputation par donneur) ou pour faire des régressions propres (pour ne pas biaiser l'estimation des coefficients dans le cas d'une imputation économétrique) : ils ne participeront pas au modèle ou seront retirés des donneurs. Il faut néanmoins veiller à ne pas classer trop de valeurs comme atypiques pour ne pas fausser la distribution : lorsqu'on effectue une régression robuste, on n'exclut généralement des imputations qu'un sous-ensemble des valeurs classées en atypiques (les régressions robustes conduisent généralement à considérer 3 à 4 % des observations comme des « outliers », ce qui est trop important).

Il est important de bien distinguer les valeurs atypiques des valeurs aberrantes pour éviter de modifier des choses vraies et d'éliminer à tort les queues de distribution, sans pour autant conserver des valeurs erronées. En cas de doute, on conserve la réponse du ménage et on la considère comme atypique pour ne pas l'utiliser lors des imputations.

Une fois les valeurs atypiques et aberrantes repérées, on peut procéder à la correction des non-réponses partielles.

II. SECONDE ÉTAPE DES REDRESSEMENTS : LA CORRECTION DES NON-RÉPONSES PARTIELLES PAR IMPUTATION

L'objectif d'une enquête est de permettre l'estimation de paramètres d'intérêt sur l'ensemble de la population ou sur des domaines (totaux, moyennes, ratios de deux variables d'intérêt, médianes, variances, coefficients de régression et de corrélation...). Pour mettre en œuvre les méthodes d'analyse courantes, il faut que les variables soient correctement renseignées pour tous les ménages enquêtés.

2.1. Pourquoi corriger la non-réponse partielle ?

On pourrait envisager de ne « rien » faire, c'est-à-dire de ne travailler que sur le sous-échantillon des unités pour lequel la non-réponse n'affecte aucune variable d'intérêt ou auxiliaire (les logiciels comme SAS fonctionnent d'ailleurs sur ce principe). Mais la présence de non-réponse partielle a des effets sur les estimateurs : pour chaque variable d'intérêt, l'échantillon des répondants résulte d'une sélection qui est « inconnue » et il est impossible d'exploiter le fichier sans comprendre ce processus de sélection. Plus précisément, la restriction de l'analyse aux questionnaires non affectés de non-réponse partielle pose des problèmes pour les estimateurs des paramètres de la population en termes de biais et de précision, deux des indicateurs fréquemment utilisés pour évaluer la qualité des estimateurs.

[6] Les données brutes sont naturellement conservées : on peut ainsi savoir quelles données ont été corrigées ou imputées.

[7] Si on effectue une procédure « automatique » de validation des données, il faut être prudent lors de l'interprétation des résultats (par exemple, si on fixe un seuil maximum pour un type de revenu et qu'on met à valeur manquante les montants qui le dépassent, il ne faut pas déduire à tort lors de l'exploitation que ce type de revenu est toujours inférieur au seuil fixé).

a) Les estimateurs calculés à partir des seuls répondants peuvent être biaisés

Les non-répondants à une question sont susceptibles d'avoir un comportement et des caractéristiques différents de ceux des répondants car la décision de ne pas répondre est souvent corrélée avec des caractéristiques socio-économiques du ménage. Par exemple, il est probable que les ménages ayant des revenus élevés ont tendance à moins répondre aux questions sur les revenus. Il existe donc un biais de sélection dans l'échantillon car il ne représente pas l'ensemble de la population mais seulement ceux qui ont accepté de répondre : les non-répondants sont certainement parmi les plus riches. La variable d'intérêt n'est donc observée que pour une partie non aléatoire de la population.

Si on calcule les estimateurs à partir des seuls répondants avec les formules habituelles, la présence de non-réponse partielle risque de mener à des estimateurs biaisés, le biais étant d'autant plus important que le taux de non-réponse est élevé[8]. Ainsi, l'estimateur du revenu moyen obtenu à partir des répondants est biaisé puisque les hauts montants sont sous-représentés : il sous-estime le vrai revenu moyen. De même, l'estimateur des MCO calculé à partir des seuls répondants est biaisé : pour avoir des estimations sans biais des coefficients de régression, on doit recourir à des méthodes d'économétrie particulières comme la méthode en deux étapes d'Heckman.

La réduction du biais de sélection par la méthode en deux étapes d'Heckman

La méthode d'Heckman (1979) vise à réduire le biais de sélection. Elle se déroule en deux étapes :
- la première étape repose sur l'estimation d'une équation de sélection. On estime un modèle probit portant sur la variable qualitative déterminant la sélection, à partir duquel on obtient un paramètre : l'inverse du ratio de Mills, qui correspond au hasard de non-sélection.
- lors de la deuxième étape, on intègre ce terme correcteur parmi les variables explicatives du modèle de régression, ce qui permet de réduire le biais des coefficients estimés.

En présence d'un biais de sélection, tout se passe en effet comme si une variable omise biaisait les coefficients estimés : le ratio de Mills inversé permet de donner une idée de ce facteur, c'est-à-dire de corriger le fait que l'estimation ne concerne pas l'ensemble des individus mais seulement une partie. On gagne ainsi de l'information et les coefficients estimés par les MCO sont moins biaisés.

Cependant, les exploitations envisagées ne se limitent pas à l'estimation de coefficients de régression donc on doit avoir recours à des méthodes d'imputation pour corriger la non-réponse partielle.

b) En présence de non-réponse partielle, les estimateurs sont moins précis

Un autre problème posé par la présence de non-réponse partielle est la perte de précision des estimateurs. En effet, la taille du sous-échantillon exploitable pour la variable d'intérêt est plus petite que s'il n'y avait pas eu de non-réponse partielle (on peut voir la non-réponse partielle comme une phase supplémentaire d'échantillonnage non contrôlé) : on perd de l'information et la variance des estimateurs est plus grande. Par conséquent, les non-réponses diminuent la fiabilité des résultats, notamment si l'on travaille sur des sous-populations.

On ne peut donc pas se limiter aux non-répondants pour les analyses (sauf éventuellement pour estimer des coefficients de régression) : il faut conserver tous les ménages enquêtés et corriger la non-réponse partielle. C'est le biais des estimateurs ponctuels qui est le problème le plus important à résoudre.

2.2. La correction de la non-réponse partielle par imputation

Le principal objectif du traitement de la non-réponse partielle est de permettre la production d'estimations approximativement sans biais. On la corrige par des méthodes d'imputation, qui consistent à remplacer les réponses manquantes par des valeurs « plausibles » à la fois au niveau individuel et au niveau agrégé.

Les valeurs imputées pour les non-répondants sont issues ou estimées à partir de la distribution des répondants de façon déterministe ou aléatoire ; or ces derniers n'ont pas forcément les mêmes caractéristiques que les non-répondants (c'est justement l'une des raisons pour lesquelles on cherche à corriger la non-réponse). Pour prendre en compte les différences entre répondants et non-répondants lors de l'imputation, on utilise des informations connues à la fois sur les répondants et les non-

[8] Le but de la prévention de la non-réponse partielle est précisément de minimiser la quantité d'information manquante pour réduire autant que possible le biais dû à la non-réponse.

répondants : les **variables auxiliaires**. Ce sont des variables de l'enquête, autres que la variable d'intérêt, dont on tient compte afin que le biais des estimateurs obtenus sur les données imputées soit plus faible que celui des estimateurs calculés à partir des seuls répondants (cela permet aussi de réduire la variance si la méthode est aléatoire). En principe, une variable est dite auxiliaire si elle est disponible pour toutes les unités échantillonnées mais en pratique, une variable pour laquelle il y a très peu de non-réponses peut être utilisée comme variable auxiliaire même si elle n'est pas connue pour l'ensemble de l'échantillon.

La qualité des imputations dépend en grande partie de l'information auxiliaire disponible donc le choix des variables auxiliaires pertinentes est très important. Pour déterminer lesquelles retenir, on peut modéliser la variable d'intérêt afin de sélectionner un ensemble de variables auxiliaires final (le biais sera d'autant plus faible que les variables auxiliaires choisies sont bien corrélées non seulement avec la variable d'intérêt mais aussi avec la probabilité de réponse). Selon la méthode d'imputation utilisée, les variables auxiliaires peuvent être incorporées comme variables explicatives dans un modèle ou être utilisées pour former des classes d'imputation, c'est-à-dire des groupes d'unités homogènes à l'intérieur desquels on effectue les imputations.

a) Classifications des méthodes d'imputation

La majorité des méthodes d'imputation peuvent être représentées par le modèle suivant :

$$y_i = f(z_i) + \varepsilon_i,$$
$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = 0, i \neq j, \quad E(\varepsilon_i^2) = \sigma_i^2$$

où y est la variable d'intérêt que l'on veut imputer et z un vecteur de variables auxiliaires disponible pour toutes les unités de l'échantillon s (répondantes ou non).
Soit y_i^* la valeur imputée utilisée pour remplacer la valeur manquante y_i.

Les principales méthodes d'imputation peuvent être classées en deux groupes :

- méthodes **déterministes (ou prédictives)** : pour un échantillon donné, elles donnent une valeur fixe si le processus d'imputation est répété. Elles consistent à obtenir y_i^* en estimant la fonction f par $\hat{f}_r$ au moyen des unités répondantes $i \in s_r$: $y_i^* = \hat{f}_r(z_i)$.

- méthodes **stochastiques (ou aléatoires)** : leur but est de restituer une distribution proche de celle que l'on aurait obtenue en l'absence de non-réponse. Elles ont une composante aléatoire donc ne fournissent pas nécessairement la même valeur imputée étant donné l'échantillon si l'on répète le processus d'imputation. L'imputation stochastique peut être vue comme une imputation déterministe à laquelle on a ajouté un terme résiduel aléatoire e* : $y_i^* = \hat{f}_r(z_i) + e_i^*$

Une classification alternative peut être utilisée :
- méthodes d'imputation **par donneur** : on a recours aux valeurs d'individus répondants (donneurs) pour imputer les valeurs manquantes. Le donneur peut être choisi au hasard ou non.
- méthodes d'imputation **par valeur prédite**, c'est-à-dire artificielle (non observée) : elles utilisent diverses fonctions des valeurs des répondants pour obtenir les valeurs de remplacement

b) L'imputation offre plusieurs avantages

L'imputation facilite l'analyse des données car elle mène à la création d'un fichier complet à partir duquel on peut effectuer des analyses de données, des régressions... : elle assure donc la cohérence des résultats issus de différentes analyses. Par exemple, si on impute les y_i manquants par $\hat{y}_k$, on obtient un ensemble de données complétées :

$$y_k^* = \begin{cases} y_k, & k \in r \text{ (ensemble des répondants)} \\ \hat{y}_k, & k \in s\text{-}r \text{ (non-répondants)} \end{cases}$$

L'imputation évite également la perte d'informations car elle utilise toutes les données observées sur les répondants partiels. Si les méthodes d'imputation ont toutes pour objectif d'améliorer la qualité des données finales, elles ne résolvent cependant pas tous les problèmes liés à la présence de non-réponse partielle et peuvent même en amener de nouveaux sur les statistiques produites *in fine*.

2.3. Les limites des méthodes d'imputation

Les méthodes d'imputation offrent de nombreux avantages mais il ne faut pas oublier lors de l'exploitation que la présence de données imputées a des conséquences sur la qualité des inférences[9], surtout si beaucoup de réponses ont été imputées : aucune méthode d'imputation ne remplacera une vraie réponse.

a) Les estimateurs imputés peuvent être biaisés

Le biais des estimateurs imputés n'est rendu négligeable que sous certaines hypothèses concernant le mécanisme de réponse[10] et l'information auxiliaire.

On peut considérer que l'échantillon des répondants à une question est sélectionné parmi les questionnaires collectés selon un mécanisme de réponse, qui correspond à la distribution des probabilités de réponse à la variable d'intérêt. Il est inconnu puisque l'on ignore la loi des non-réponses : on doit donc établir une hypothèse sur sa nature, qui peut être de trois types :

- **mécanisme uniforme** (MCAR : Missing Completely At Random) : il suppose que les données sont manquantes de façon aléatoire. La probabilité de réponse pour la variable d'intérêt est la même pour toutes les unités de la population et ne dépend donc ni des variables auxiliaires, ni de la variable d'intérêt. Cette hypothèse correspond au cas idéal puisque les estimateurs sont non biaisés en présence de données imputées. Mais elle est peu réaliste car les non-répondants ont souvent des caractéristiques différentes des répondants.

- **mécanisme ignorable** (MAR : Missing At Random) : sous cette hypothèse, la non-réponse et la variable d'intérêt sont indépendantes conditionnellement à de l'information auxiliaire ; en d'autres termes, la probabilité de réponse n'est pas liée à la valeur de la variable d'intérêt une fois prise en compte l'information auxiliaire. C'est cette hypothèse, plus réaliste que celle du mécanisme uniforme, qui est généralement retenue dans les travaux d'imputation. Comme on suppose que les données sont manquantes de façon aléatoire conditionnellement à l'information auxiliaire, cela implique que le biais introduit par la non-réponse peut être entièrement corrigé par l'utilisation d'une bonne information auxiliaire.

- **mécanisme non-ignorable** (NMAR : Not Missing At Random) : sous cette hypothèse, la probabilité de répondre à la variable d'intérêt dépend de la variable d'intérêt elle-même (et éventuellement d'autres variables non observées). C'est la situation la plus problématique car la non-réponse n'est explicable que par la variable d'intérêt.

A moins de faire l'hypothèse forte d'un mécanisme uniforme, des biais liés à la non-réponse partielle subsistent après l'imputation :

- **si, dans le cas d'un mécanisme ignorable, on ne tient pas compte de certaines variables qui expliquent la sélection**. En effet, le biais ne peut être supprimé que si l'on utilise une bonne information auxiliaire : les estimateurs sont biaisés si on a imputé avec un mauvais modèle, si on a oublié une variable auxiliaire importante parmi les variables explicatives (dans le cas d'une imputation économétrique), ou encore si on a utilisé de mauvaises variables pour former les classes d'imputation.

- **si les variables auxiliaires disponibles ne suffisent pas pour appréhender la sélection, ce qui est le cas lorsque le mécanisme est non-ignorable**. Pour les revenus fonciers par exemple, comme la non-réponse ne dépend pas seulement de variables auxiliaires mais aussi du montant perçu, l'utilisation d'information auxiliaire ne permet pas de supprimer le biais. Au terme des imputations, il subsiste un biais de sélection dont il faut tenir compte lors de l'exploitation, en utilisant par exemple la méthode d'Heckman.

Les méthodes de correction de la non-réponse partielle reposent en effet sur l'hypothèse d'un mécanisme de non-réponse ignorable (MAR) : elles supposent que les non-répondants répondent

[9] Voir Haziza (2002) pour plus de détails.
[10] Gautier (2005).

comme les répondants conditionnellement à un ensemble de variables. En présence d'un mécanisme non-ignorable (NMAR), elles ne permettent pas de résoudre le problème de l'effet de sélection car on ne peut « inventer » de l'information dont on ne dispose pas : par définition, les méthodes par donneur ne changent jamais la valeur la plus élevée. Quant aux méthodes économétriques, elles ne sont pas non plus la solution car on effectue la modélisation à partir des ménages répondants (qui ont sans doute perçu en moyenne des revenus fonciers plus faibles) : le modèle estimé risque de ne pas être adéquat pour les valeurs très élevées des variables auxiliaires puisque peu de grandes valeurs sont observées.
Si l'on ne prend pas en compte ce biais de sélection lors de l'inférence, on sous-estime les revenus fonciers de l'ensemble de la population car les données ont été imputées à partir des répondants moins riches[11] : les estimations sont d'autant plus biaisées que le taux de non-réponse est important.

b) Les estimateurs imputés sont moins précis que les estimateurs spontanés

Une autre limite des méthodes d'imputation concerne la précision des estimateurs. Loin de l'améliorer, le fait d'imputer une partie des données ajoute de l'imprécision : la variance des estimateurs imputés est toujours plus grande que celle des estimateurs spontanés, c'est-à-dire obtenus sur données réelles. Par conséquent, si on traite les données imputées comme si elles avaient été observées, la variance totale de l'estimateur risque d'être sous-estimée (d'autant plus que le taux de non-réponse est élevé). Le fait de surestimer sa précision peut entraîner une sous-estimation de la longueur des intervalles de confiance et mener à des tests erronés.

c) Certaines méthodes d'imputation distordent la distribution des variables d'intérêt

Les méthodes déterministes introduisent une distorsion dans la fonction de répartition de la variable imputée. On évitera donc de les utiliser si l'on souhaite conserver la structure globale de la variable d'intérêt et on privilégiera des méthodes aléatoires, qui permettent d'obtenir une distribution redressée proche de celle de la variable d'origine. Plus précisément, les valeurs imputées doivent respecter la distribution de la variable d'intérêt telle qu'elle serait si toutes les valeurs avaient été observées (distribution « naturelle ») et non telle qu'elle apparaît à partir des seules données observées, puisque sa répartition initiale peut être biaisée si la décision de ne pas répondre n'est pas indépendante de la valeur de la variable d'intérêt (l'imputation vise justement à corriger ce biais).

d) L'imputation peut modifier les relations entre variables

L'imputation peut altérer les corrélations entre variables donc il faut être vigilant lors des inférences si l'on utilise des données imputées dans des analyses multivariées (notamment des régressions). En effet, si l'imputation est réalisée indépendamment pour chaque variable, on néglige les liens qui peuvent exister entre les différentes variables imputées. Si l'on souhaite préserver la cohérence des relations entre certaines variables pour chaque individu, il est préférable de s'orienter vers des méthodes par donneur qui permettent d'utiliser un donneur unique pour imputer plusieurs variables et donc de conserver les liens. Dans certains cas, un redressement par étapes peut également permettre de tenir compte des liens entre les variables à imputer : c'est le cas pour les variables relatives aux prêts immobiliers (les résultats sont présentés en annexe 7).

2.4. Sur quels critères comparer les méthodes d'imputation ?

La comparaison directe des propriétés des estimateurs obtenus sur les données après imputation est complexe. En ce qui concerne le biais des estimateurs imputés, il est difficile de l'estimer puisque par définition on ne connaît pas la réponse des non-répondants.

Quant à la précision des estimateurs, elle est compliquée à calculer en présence de données imputées[12]. En effet, l'imputation se traduit par l'ajout d'un terme de variance supplémentaire : pour estimer la variance totale des estimateurs, il faut donc tenir compte non seulement de la variance d'échantillonnage et de la variance induite par le mécanisme de non-réponse totale, mais aussi de la variance liée à l'imputation (elle peut représenter près de 15 % de la variance totale[13]). Chaque technique d'imputation conduit à une formule de variance et à une estimation de variance

[11] D'autant plus que certains répondants ont pu volontairement sous-estimer le montant de leurs revenus.
[12] Il est important de pouvoir identifier dans les tables d'exploitation les valeurs imputées pour calculer la variance liée à l'imputation.
[13] Voir Caron (1999) pour plus de détails.

particulières[14]. Pour calculer la précision des estimateurs imputés, on peut utiliser une approche en trois phases ou une approche modèle[15].

Notons que le calcul de précision ne permet de comparer entre elles que les méthodes d'une même famille (aléatoire ou déterministe) ; du point de vue de la précision, les méthodes aléatoires sont en effet toujours moins bonnes que les méthodes déterministes puisque l'ajout d'un terme aléatoire introduit un terme de variance supplémentaire.

Comme il est impossible d'évaluer le biais et compliqué de calculer la variance des estimateurs, c'est plutôt à partir de leurs effets sur la distribution de la variable imputée que l'on va comparer les différentes méthodes d'imputation.

2.5. Avantages et inconvénients des méthodes d'imputation selon leur nature

On présente ici les avantages et inconvénients des grands types de méthodes d'imputation (déterministes ou aléatoires, par donneur ou par valeur prédite). Les méthodes choisies pour l'enquête Logement seront détaillées dans la partie suivante.

a) Méthodes déterministes

Du point de vue de la distribution de la variable d'intérêt, les méthodes déterministes sont peu intéressantes car elles introduisent une distorsion artificielle : l'imputation par la moyenne par classes peut par exemple créer des pics. On sous-estime donc la variabilité naturelle, c'est-à-dire celle que l'on aurait observée en l'absence de non-réponse. Les estimateurs des caractéristiques de la distribution (quantiles, variance empirique…) sont donc biaisés, de même que les estimateurs des MCO.
En revanche, comme la précision des estimateurs obtenus à l'issue de l'utilisation de méthodes déterministes est toujours meilleure qu'avec des méthodes aléatoires, on peut les mettre en œuvre si l'objectif est d'estimer des totaux car elles fournissent des estimateurs plus précis.

b) Méthodes aléatoires (stochastiques)

Par rapport aux méthodes déterministes, les méthodes aléatoires augmentent la variabilité des estimateurs mais offrent davantage de propriétés intéressantes. Elles produisent des estimateurs sans biais d'un maximum de grandeurs sur la loi (moyenne, quantiles, variance empirique, estimateur des MCO…). Elles préservent la distribution et la variabilité de la variable d'intérêt : la variance empirique calculée sur l'ensemble des données est proche de celle calculée sur l'ensemble des répondants. On s'oriente donc vers des méthodes aléatoires lorsque l'objectif est multiple (analyses multivariées, calcul de médianes…).

c) Méthodes par donneur ou méthodes par valeur prédite

Par rapport aux méthodes par valeur prédite, les méthodes par donneur (déterministes ou stochastiques) présentent plusieurs avantages par rapport aux méthodes par valeur prédite. Elles permettent d'imputer des données qui existent puisqu'elles ont été observées sur des répondants. Elles peuvent donc être utilisées pour imputer des variables qualitatives ou quantitatives alors que les méthodes par valeur prédite s'appliquent pour des variables quantitatives. Enfin, il est possible de préserver les liaisons entre variables si l'on utilise un donneur unique. Par contre, les méthodes par donneur ne permettent pas d'utiliser facilement beaucoup d'information auxiliaire, contrairement aux méthodes économétriques.

III. PRÉSENTATION DES MÉTHODES D'IMPUTATION RETENUES POUR L'ENQUÊTE LOGEMENT 2006

Compte-tenu des avantages et inconvénients listés ci-dessus, on essaie de choisir pour chaque variable de l'enquête la méthode d'imputation la plus pertinente, en fonction :
- des hypothèses qu'elle requiert (variable d'intérêt modélisable…)
- de la disponibilité d'information auxiliaire de bonne qualité
- de l'usage qui sera fait de la variable lors des exploitations ultérieures : est-ce que la distribution est importante, est-ce qu'on s'intéresse à l'estimation de totaux, de moyennes ?…

[14] Caron (2005).
[15] Haziza et Rancourt (2004).

- du temps disponible, compte-tenu du nombre de variables à redresser et de l'importance de la variable dans l'enquête

Lorsque les valeurs manquantes peuvent être corrigées directement à l'aide des autres variables disponibles, on utilise une méthode déductive. Par exemple, pour les prêts à taux fixe (non renégociés ni remboursés par anticipation) dont les remboursements sont constants, le montant des mensualités peut être déduit à partir de la durée totale du prêt et du montant emprunté (cf. annexe 7). Pour les autres variables, on ne dispose pas de suffisamment d'informations pour recourir à ce type de méthode donc on s'oriente vers d'autres techniques.

3.1. Pour les principales variables financières, on privilégie des méthodes aléatoires

Pour les variables essentielles de l'enquête (revenus, loyers, prix du logement…), on cherche des méthodes permettant de tenir compte du maximum d'information auxiliaire disponible et préservant la variabilité de la variable d'intérêt. On privilégie donc des méthodes d'imputation aléatoires : méthode des résidus simulés et hot-deck aléatoire ou séquentiel.

a) Présentation de l'imputation par la méthode des résidus simulés

Cette méthode d'imputation économétrique (par valeur prédite) peut être appliquée lorsqu'une variable Y continue est manquante pour la population des non-répondants mais que l'on dispose d'une ou de plusieurs variables renseignées (X_1,…X_n) sur cette population.

On suppose que la variable à imputer Y est liée aux variables auxiliaires X par une relation économétrique simple de type régression linéaire :

$Y = X\beta + \varepsilon$.

L'imputation se déroule en deux phases :
- dans un premier temps, on calcule une prédiction des montants non déclarés par régression
- dans un deuxième temps, on simule ou on tire aléatoirement des résidus que l'on ajoute à l'estimateur calculé précédemment : l'ajout de cet aléa permet de ne pas toujours imputer la moyenne conditionnelle

Première étape : recherche d'un modèle ayant un bon pouvoir prédictif puis calcul des prédictions $X_i\hat{\beta}$ pour les non-répondants

i) On commence par estimer le modèle sur la population des répondants

Les valeurs atypiques repérées lors de l'étape de vérification des données sont exclues. Le modèle peut être estimé par les MCO.
Lorsque la variable à imputer peut être renseignée en clair ou en tranches, on utilise une PROC LIFEREG[16] (cf. annexe 3). Elle permet de traiter des modèles censurés, c'est-à-dire dans lesquels la variable expliquée est soumise à une troncature (censurée à gauche et/ou à droite) : elle peut notamment être utilisée pour ajuster des modèles « mixtes », dans lesquels la variable d'intérêt est observée sous forme de tranches pour une partie de l'échantillon et en clair pour certaines observations (c'est le cas des revenus[17]). La PROC LIFEREG étant sensible aux valeurs atypiques et aberrantes, il est important de bien les avoir repérées avant afin de les écarter de la régression.
On peut aussi estimer le modèle par une régression robuste si l'on souhaite limiter l'influence des valeurs atypiques. En effet, lorsque les données sont contaminées par des erreurs, l'estimation par maximisation de vraisemblance ou par les MCO donne des résultats biaisés.

On cherche à maximiser la capacité prédictive du modèle afin de réduire le biais des estimateurs imputés. Contrairement aux modélisations que l'on effectue habituellement, on ne cherche pas à limiter le nombre de régresseurs (principe de parcimonie) car l'objectif n'est pas de construire un modèle économétrique pour faire de la prévision. En vue des imputations, il est essentiel de ne pas oublier de variables auxiliaires lors de la modélisation donc on introduit un maximum de variables explicatives et

[16] Voir Lollivier (1997) pour plus de détails.
[17] En effet, si la variable d'intérêt peut être renseignée en tranches, on ne peut pas utiliser les MCO classiques ou la régression robuste puisque le revenu pourrait prendre des valeurs n'appartenant pas à la tranche déclarée. Par ailleurs, exclure de l'estimation les répondants en tranches pour se restreindre au groupe des répondants en clair pourrait engendrer des biais de sélection.

on les conserve même si elles ne sont pas très significatives, dès lors qu'elles ont un sens vis-à-vis de la variable à expliquer.

Pour améliorer l'ajustement du modèle et avoir des résultats plus robustes, il peut être préférable d'utiliser la transformation logarithmique lorsque la variable d'intérêt est un montant (qui suit approximativement une loi log-normale). En effet, la réalisation de la loi normale n'est pas forcément un nombre positif donc le passage au logarithme permet de bien spécifier le modèle.

Il est par ailleurs important de vérifier que le modèle est stable sur les sous-populations : si ce n'est pas le cas, on peut être amené à effectuer des modèles de régressions séparés pour chaque catégorie. Par exemple, on a modélisé les salaires des hommes et des femmes de façon distincte.

Enfin, pour pouvoir effectuer des régressions propres, on doit aussi s'assurer que les variables explicatives sont renseignées pour toutes les observations et que les effectifs des modalités sont suffisants (au moins 5 % des individus), en regroupant au besoin celles dont les effectifs sont trop faibles.

ii) <u>On calcule ensuite les prédictions</u> $X_i\hat{\beta}$ <u>pour les Y_i non renseignés</u> en appliquant aux non-répondants les coefficients $\hat{\beta}$ estimés sur les répondants

Deuxième étape : on simule des résidus que l'on ajoute aux prédictions calculées à l'étape précédente afin de calculer les imputations $\hat{Y} = X\hat{\beta} + \hat{u}$

i) <u>Les résidus aléatoires peuvent être générés de deux façons : simulation à partir d'une loi normale ou tirage dans la loi empirique des résidus</u>

Si les résidus des répondants suivent une loi proche d'une loi normale, on peut simuler les résidus aléatoires û à partir d'une distribution normale d'espérance nulle et de variance égale à la variance empirique des résidus des répondants. Si les résidus observés s'écartent beaucoup d'une loi normale, il est préférable de tirer les résidus au hasard parmi l'ensemble des résidus observés sur les répondants afin de respecter l'hétérogénéité des données.

L'inconvénient du tirage dans la loi empirique des résidus est qu'il peut mener à des estimations trop dispersées. Pour éviter d'avoir des valeurs trop faibles ou trop élevées, il peut donc être plus approprié de procéder à une simulation des résidus dans la loi normale même si les résidus observés sur les répondants ne suivent pas tout à fait une loi normale (c'est le cas pour les variables monétaires par exemple).

Analyse des résidus de l'équation (hypothèses de normalité et d'homoscédasticité)

Pour choisir la manière dont on va générer les résidus aléatoires û, on doit donc examiner les hypothèses de normalité et d'homoscédasticité des résidus observés pour les répondants. Quand il y a un grand nombre d'observations, l'estimateur des MCO est sans biais même si les résidus ne sont pas distribués selon une loi normale et même en présence d'hétéroscédasticité, dès lors que les résidus sont d'espérance nulle et les variables exogènes (et que le modèle est bon). Par contre, dans la perspective de simuler les résidus, il faut vérifier la normalité et l'hétéroscédasticité des résidus observés. En effet, pour obtenir au final une « bonne » variabilité de la variable imputée, les résidus imputés doivent avoir la même distribution que les résidus tirés de l'équation de la variable d'intérêt : si on ne tient pas compte de la vraie distribution des résidus, on peut sous-estimer ou surestimer la vraie variance des montants estimés.

- Vérification de l'hypothèse de normalité

Pour savoir quelle méthode de simulation des résidus utiliser, on vérifie la normalité des résidus observés (en éditant par exemple l'histogramme des résidus afin de voir s'ils s'écartent beaucoup d'une loi normale, ou en comparant leurs quantiles à ceux d'une loi normale).

- Vérification de l'hypothèse d'homoscédasticité

L'homoscédasticité est le fait que la variance des résidus est constante conditionnellement à la valeur des variables explicatives. Si l'on constate au contraire que la variance des termes d'erreur a tendance à adopter un comportement systématique conditionnellement à l'une au moins des variables explicatives, on parle d'hétéroscédasticité. Le test de White permet de repérer la présence d'hétéroscédasticité et les variables les plus en cause peuvent être identifiées grâce au test de Breusch-Pagan (cf. annexe 4).
On n'est pas obligé de corriger[18] l'hétéroscédasticité lors de la modélisation puisqu'elle ne biaise pas l'estimation des coefficients. En revanche, ces derniers ne sont plus de variance minimum en présence d'hétéroscédasticité[19] : l'estimateur n'est plus optimal car les MCO ne tiennent pas compte du fait que les résidus connaissent un niveau de variance très différent selon les valeurs prises par la ou les variables explicatives en cause dans l'hétéroscédasticité. Il faut donc tenir compte de l'hétéroscédasticité lors du tirage ou de la simulation des résidus :
- si on simule les résidus, il faut « stratifier » les résidus par la variable en cause, en calculant les écarts-types des résidus observés pour chaque strate.
- si on tire les résidus parmi les résidus observés, il faut créer des classes parmi les observations afin d'affecter à un receveur le résidu d'un donneur proche, c'est-à-dire partageant certaines particularités.

ii) <u>Les résidus aléatoires ainsi générés sont ensuite ajoutés aux prédictions calculées à l'étape précédente</u> : chaque valeur manquante Y_i est remplacée par la valeur imputée $\hat{Y}_i = X_i\hat{\beta} + \hat{u}_i$. Si de l'information en tranches est disponible, on impose au résultat de se trouver dans la tranche déclarée.

L'ensemble de cette deuxième étape (c'est-à-dire la simulation des résidus et le calcul des imputations) peut être réalisé avec la macro %simul (cf. annexe 5).

Avantages et inconvénients de l'imputation par les résidus simulés

A l'enquête Logement 2006, on a utilisé la méthode des résidus simulés pour imputer des variables importantes comme les salaires, les loyers, les charges ou encore le prix des logements. Cette méthode offre en effet plusieurs avantages : elle préserve la distribution de la variable d'intérêt et permet d'utiliser un grand nombre de variables auxiliaires. Les valeurs imputées sont sans biais sauf si l'on a utilisé une transformation logarithmique de la variable d'intérêt (car ce sont alors les estimateurs sur le logarithme qui sont sans biais ; on peut cependant considérer que ce biais est négligeable)[20].

La méthode des résidus simulés présente cependant plusieurs inconvénients. Elle nécessite de disposer d'une information auxiliaire de bonne qualité et d'effectuer un travail de modélisation important (diagnostics de qualité, tests de l'homoscédasticité et de la normalité des résidus, transformation éventuelle de la variable d'intérêt…) car la validité de l'inférence en présence de données imputées dépend de la validité du modèle choisi. Par ailleurs, les relations entre variables sont détruites par la méthode des résidus simulés puisqu'elle ne permet d'imputer qu'une variable à la fois. De plus, comme toute méthode aléatoire, elle augmente la variance des estimateurs.

Un autre inconvénient de la méthode des résidus simulés est le risque de générer des résidus trop forts. En effet, comme on observe généralement peu de montants très élevés ou très faibles (revenus par exemple), le pouvoir prédictif du modèle peut être moins bon pour ces types de montants. De plus, si une partie des valeurs atypiques n'a pas été repérée, la variance des résidus est plus élevée. Or, si on simule des résidus trop importants, on risque d'imputer des montants trop grands, ce qui peut conduire à fortement surestimer la moyenne. L'effet peut être très fort si l'on a procédé à une transformation logarithmique de la variable d'intérêt pour la modélisation car les valeurs élevées sont aggravées par le passage final à l'exponentielle. A l'issue de l'imputation, il est donc important de regarder si des montants sont beaucoup plus élevés que le montant maximum chez les déclarés. Si

[18] On peut corriger l'hétéroscédasticité par l'approche des MCG : il s'agit de diviser, pour chaque individu, tous les termes du modèle par la variance des résidus, puis d'estimer le modèle obtenu à l'aide des MCO.
[19] D'après le théorème de Gauss-Markov, l'estimateur MCO est BLUE (Best Linear Unbiased Estimator), c'est-à-dire sans biais, de variance minimale et convergent si : 1) $E(u_i) = 0 \ \forall\ i = 1, \ldots n$; 2) $Var(u_i) = \sigma^2 \ \forall\ i = 1, \ldots n$: hypothèse d'homoscédasticité des résidus ; 3) $Cov(u_i,u_j) = 0$ pour tout $i \neq j$: hypothèse d'absence d'autocorrélation des résidus ; 4) $Cov(X_j,u_j) = 0$: hypothèse d'exogénéité des variables explicatives (indépendance avec le terme d'erreur).
En présence d'hétéroscédasticité, l'estimateur des MCO est sans biais mais non optimal.
[20] Voir Caron (1999).

peu de cas sont relevés, on peut se contenter de modifier les quelques observations concernées mais s'ils sont nombreux, il faut revoir la modélisation.

b) Présentation de l'imputation par hot-deck

L'imputation par hot-deck est une méthode par donneur, qui consiste à remplacer la valeur manquante d'un individu non-répondant (receveur) par la valeur observée d'un répondant (donneur). Il existe plusieurs types de hot-deck : pour les redressements de l'enquête Logement, on a eu recours au hot-deck aléatoire et au hot-deck séquentiel.

i) Hot-deck aléatoire

Le hot-deck aléatoire consiste à remplacer chaque valeur manquante par la valeur observée pour un répondant choisi au hasard. Elle peut être vue comme une imputation par la moyenne à laquelle on a ajouté un résidu aléatoire existant. Le fait de tirer le répondant au hasard ajoute une variabilité dans les valeurs imputées, qui doit être prise en compte si on veut estimer la variance.

Le recours à cette méthode suppose que les répondants aient un comportement homogène afin que les réponses dupliquées ne soient pas trop éloignées de la vraie valeur. Lorsque la population est trop hétérogène, il est donc préférable de partitionner l'échantillon de manière à regrouper les individus qui se ressemblent, c'est-à-dire de constituer des classes d'imputation. On applique ensuite le hot-deck aléatoire à l'intérieur de ces sous-populations plus homogènes : on tire un donneur au hasard dans la classe à laquelle appartient le receveur et non plus dans l'ensemble de la population. Ainsi, pour chaque valeur manquante, on impute la valeur d'un répondant « proche ».

La formation de classes d'imputation homogènes permet de réduire le biais dû à la non-réponse

En pratique, pour certaines méthodes d'imputation comme le hot-deck, on forme souvent des classes avant d'imputer pour assurer une certaine robustesse aux imputations. L'objectif des classes d'imputation est de réduire le biais dû à la non-réponse : pour les constituer, on doit donc former des groupes d'unités homogènes par rapport à la variable d'intérêt. Cela signifie qu'à l'intérieur de chaque classe, les unités doivent avoir approximativement la même valeur pour la variable d'intérêt (la variance intra-classe doit être petite). Comme par définition on ignore la valeur de la variable d'intérêt pour les non-répondants, on utilise de l'information auxiliaire pour constituer les classes.
Plusieurs méthodes peuvent être utilisées pour les former (pour le hot-deck, les classes doivent exister dans l'échantillon des receveurs et dans celui des donneurs). Pour déterminer les variables qui permettent d'obtenir des classes homogènes, on peut modéliser la variable d'intérêt afin de sélectionner les variables qui sont les plus corrélées avec elle. Pour les imputations de l'enquête Logement, on a constitué les classes en croisant plusieurs variables auxiliaires catégorielles (âge, sexe, catégorie sociale...).
Pour diminuer la variance des estimateurs, il faut éviter d'utiliser un trop grand nombre de classes, donc de croiser trop de variables, surtout si elles comportent un grand nombre de modalités : on risque en effet de constituer des classes vides de donneur ou de taille insuffisante, c'est-à-dire comportant plus de receveurs que de donneurs, ce qui conduirait à dupliquer certaines valeurs. On peut s'assurer que le nombre de répondants à l'intérieur d'une classe soit supérieur à certain seuil afin que le donneur soit tiré aléatoirement et, si cette contrainte n'est pas respectée, regrouper des classes en éliminant des variables auxiliaires ou en regroupant des modalités.
Il ne faut cependant pas chercher à avoir un nombre de donneurs très élevé dans chaque classe si cela nécessite d'éliminer beaucoup de variables de classe, car si on a trop d'hétérogénéité dans certaines classes (c'est-à-dire si une variable auxiliaire fortement significative n'est pas utilisée), on risque d'aboutir à des estimations biaisées.

Le tirage des donneurs est de préférence sans remise, surtout si on impute plusieurs variables, car l'imputation par hot-deck aléatoire sans remise mène à une plus petite variance que l'imputation avec remise. Mais la sélection des donneurs est plus simple avec remise car, dans le cas sans remise, il faut qu'il y ait plus de donneurs que de receveurs : la macro %hotdeck, qui permet de faire du hot-deck aléatoire par classes tout en tenant compte d'information en tranches, effectue d'ailleurs des tirages avec remise (cf. annexe 6). Cependant, si le ratio *nombre de receveurs / nombre de donneurs* est faible (c'est-à-dire si la proportion de non-répondants est petite), il n'y a pas beaucoup de différences entre les deux méthodes car la probabilité de sélectionner plusieurs fois un même donneur est alors faible.

Avantages et inconvénients de l'imputation par hot-deck aléatoire par classes

Le hot-deck aléatoire permet d'obtenir des estimateurs sans biais et préserve la distribution de la variable d'intérêt : la variance de la distribution après imputation est proche de celle de la distribution initiale. Il fournit des données réelles et offre la possibilité de conserver une partie des liens entre

variables puisqu'on peut imputer plusieurs variables simultanément en utilisant un donneur unique. Cet aspect a été utilisé à l'enquête Logement pour imputer simultanément des types de placements financiers et de la tranche de montant associée.

En revanche, la mise en œuvre du hot-deck aléatoire implique de définir ce qu'est la ressemblance : un travail de modélisation peut être nécessaire pour déterminer les variables de classes. L'imputation par hot-deck aléatoire ne permet pas d'utiliser beaucoup d'information auxiliaire car il faut limiter le nombre de variables et de modalités pour avoir suffisamment de donneurs dans chaque classe. De plus, elle peut donner trop de poids à des valeurs aberrantes ou atypiques donc il est important de bien les repérer avant imputation. Enfin, comme c'est une méthode d'imputation aléatoire, elle augmente la variance des estimateurs.

On l'a utilisée entre autres pour imputer les revenus non salariaux, les revenus fonciers ou encore les revenus « autres ».

Le recours à l'échantillonnage équilibré pour tirer les donneurs permet de réduire la variance d'imputation

Pour réduire la variance supplémentaire générée par le hot-deck aléatoire, une amélioration possible de la méthode (non mise en œuvre ici) est d'introduire des contraintes pour sélectionner les donneurs : plus précisément, il s'agit de recourir à l'échantillonnage équilibré[21].
L'idée est de considérer les donneurs comme un échantillon de répondants auprès desquels on va recueillir les valeurs à imputer. Soit un échantillon de taille n dans lequel on suppose que la réponse manque au hasard avec la même probabilité pour chaque unité. La mise en œuvre du hot-deck aléatoire revient à tirer un échantillon de n-m donneurs par sondage aléatoire simple parmi les m répondants. Si on tire au hasard cet échantillon, on augmente la variance des estimateurs car chaque répondant va compter aléatoirement pour 1 (s'il ne sert pas de donneur) ou 2 (s'il sert de donneur) : la variance d'une moyenne peut ainsi augmenter de plus de 10 %.
Par contre, si on tire l'échantillon des donneurs de manière à ce qu'il soit équilibré sur la moyenne de la variable d'intérêt observée sur les répondants, l'échantillon des receveurs fournit la même moyenne que celle calculée sur les répondants : elle n'est donc pas affectée par la variance d'imputation[22]. Les échantillons équilibrés peuvent être sélectionnés de manière aléatoire grâce à la macro %CUBE.
Le recours à l'échantillonnage équilibré permet ainsi de combiner les avantages des méthodes déterministes et stochastiques : on obtient des estimateurs plus précis (car la variance due au hot-deck aléatoire est éliminée) tout en préservant la distribution de la variable d'intérêt puisque l'imputation est aléatoire.

ii) Hot-deck séquentiel

Le hot-deck séquentiel consiste à classer l'échantillon dans un certain ordre puis, pour chaque valeur manquante, à imputer la valeur du répondant qui la précède[23]. Pour choisir la ou les variables de tri, il faut déterminer, parmi les variables renseignées pour les répondants et les non-répondants, celles qui expliquent le mieux la variable à imputer (à partir des observations correspondant aux seuls répondants).

Comme l'estimateur obtenu est fonction de l'ordre dans lequel les données échantillonnées apparaissent dans le fichier, il faut éviter de choisir une variable de tri corrélée avec la probabilité de non-réponse afin de ne pas dupliquer plusieurs fois la même valeur. En effet, la duplication des mêmes valeurs distord la distribution et peut aussi augmenter le biais dans le cas où le donneur est atypique.

Le hot-deck séquentiel présente plusieurs avantages : très simple à programmer, il fournit des estimateurs sans biais et permet d'imputer des données réelles. En revanche, il peut donner trop de poids à des valeurs aberrantes ou atypiques, d'où l'importance de les repérer en amont, et le nombre de variables auxiliaires utilisables est très limité. A l'enquête Logement, on a eu recours au hot-deck séquentiel pour imputer les allocations RMI perçues par une personne du ménage autre que la personne de référence ou son conjoint, car on dispose de peu de variables auxiliaires.

c) Imputation par la méthode des résidus simulés ou par hot-deck ?

Il est difficile de trancher entre l'imputation par les résidus simulés (méthode par valeur prédite) et hot-deck (méthode par donneur) car utiliser l'un ou l'autre ne change pas grand-chose pour la distribution

[21] Un échantillon est dit équilibré sur une ou plusieurs variables disponibles dans la base de sondage lorsque, pour chacune d'entre elles, l'estimateur de Horvitz-Thompson coïncide exactement avec le vrai total issu de la base de sondage. Pour plus de détails, voir Deville (2005), Favre, Matei et Tillé (2005) et Rousseau et Tardieu (2004).
[22] On peut équilibrer l'échantillon sur d'autres caractéristiques estimées de la variable d'intérêt comme l'écart-type ou la médiane.
[23] Il faut donc que la première valeur soit renseignée. Si elle est manquante, on peut l'initialiser en traitant le fichier comme une liste circulaire.

(comme on le verra dans la quatrième partie) ni pour la variance[24]. La méthode des résidus simulés nécessite un travail de modélisation important mais, dans le cas d'une imputation par hot-deck, on peut aussi être amené à modéliser la variable d'intérêt afin de sélectionner les variables auxiliaires les plus corrélées, pour constituer les classes d'imputation dans le cas du hot-deck aléatoire ou pour choisir les variables de tri dans le cas du hot-deck séquentiel.

Pour les redressements de l'enquête Logement, on utilise la méthode des résidus simulés quand on trouve un modèle ayant un « bon » pouvoir prédictif (on a estimé qu'un $R^2 > 0,25$ était satisfaisant car on travaille sur des données individuelles) : elle permet d'utiliser un maximum d'information auxiliaire. Par contre, on a recours au hot-deck si on n'arrive pas à bien spécifier le modèle, lorsqu'il y a peu de variables explicatives de la variable d'intérêt par exemple (c'est le cas pour les revenus non salariaux).

Notons que pour les prestations familiales, scolaires, les allocations RMI ou les aides au logement, on a préféré utiliser ces méthodes plutôt que de procéder à une imputation par barème. En effet, le système des prestations est complexe et on ne dispose pas toujours de l'information nécessaire (revenus des années N-1, N-2…) pour reconstituer les assiettes ressources : il aurait donc fallu faire des hypothèses simplificatrices (par exemple, neutralité du décalage entre l'année de perception des ressources et le moment où elle est prise en compte pour le calcul des prestations).

3.2. Pour les variables secondaires, on a recours à des méthodes déterministes

Un inconvénient des méthodes précédentes est qu'elles sont longues à mettre en œuvre : pour les résidus simulés, il faut procéder à un travail de modélisation et la sélection des variables de classe pour le hot-deck ne va pas forcément de soi. Compte-tenu du nombre de variables à imputer, on s'est donc contenté de méthodes plus simples et plus rapides à mettre en œuvre pour les variables de moindre importance ou ayant fait l'objet de très peu de non-réponses : on a eu recours à la prédiction par la moyenne par classes ou à l'imputation par le plus proche voisin. Ces méthodes ont notamment été utilisées pour imputer les taux d'intérêt des prêts immobiliers (cf. annexe 7).

a) Présentation de l'imputation par la moyenne

L'imputation par la moyenne est une méthode par valeur prédite qui consiste à remplacer les valeurs manquantes par la moyenne observée sur les répondants (c'est un cas particulier de l'imputation par la régression). Elle peut être appliquée à l'intérieur de classes d'imputation, ce qui permet de réduire le biais dû à la non-réponse puisqu'on affecte alors la moyenne des unités répondantes partageant certaines caractéristiques.

L'imputation par la moyenne est très facile à mettre en œuvre mais ses propriétés sont limitées. Elle distord la distribution de la variable d'intérêt car les valeurs sont concentrées autour de la moyenne (ou des moyennes par classe) : certaines caractéristiques de la distribution (quartiles…) sont biaisées et la variabilité de la variable d'intérêt est sous-estimée. Elle peut par contre être utile pour estimer des totaux et des moyennes si on ne s'intéresse pas à la distribution de la variable d'intérêt, ou s'il y a peu de non-réponses et que l'on veut redresser rapidement la variable.

A l'enquête Logement, on a utilisé cette méthode pour imputer des variables d'importance moindre comme le montant de la location d'un garage, box ou parking, ou encore les frais d'agence et/ou de notaire.

b) Présentation de l'imputation par le plus proche voisin (PPV) (hot-deck métrique)

L'imputation par le plus proche voisin est une méthode par donneur qui consiste à remplacer la donnée manquante du non-répondant par celle du répondant le plus proche au sens d'une distance, calculée à partir des informations auxiliaires disponibles sur les répondants et les non-répondants. A l'enquête Logement, elle a été mise en œuvre pour imputer la part du surloyer car il y avait très peu de non-réponses à corriger.

Cette méthode présente l'avantage de fournir des données réelles puisque c'est une méthode par donneur. En revanche, elle peut être difficile à mettre en œuvre lorsque l'on souhaite prendre en compte un certain nombre de variables auxiliaires, surtout si elles sont de nature différente (qualitatives et quantitatives), car elle implique de définir ce qu'est la ressemblance. Il faut choisir une distance qui

[24] Voir Caron (1999) pour plus de détails.

permette de donner plus d'importance aux variables auxiliaires « expliquant » le mieux la variable d'intérêt : elle peut être calculée comme la somme pondérée et normalisée de distances partielles définies au niveau de chaque variable auxiliaire[25].

IV. DERNIÈRE ÉTAPE DES REDRESSEMENTS : LA VERIFICATION DE LA COHÉRENCE DU FICHIER IMPUTÉ

A l'issue des redressements, il est important de vérifier la cohérence du fichier imputé. A défaut de pouvoir évaluer le biais des estimateurs en présence de données imputées, on procède si possible à des comparaisons avec des données externes (autres enquêtes auprès des ménages, données de la CNAF…), tout en sachant que chaque enquête a des intervalles de confiance différents.

Les vérifications peuvent être effectuées au niveau de chaque variable mais aussi au niveau macroéconomique, sur des agrégats comme le revenu total disponible par exemple.

[25] La distance partielle est égale à 0 si le receveur et le donneur potentiel prennent la même modalité pour la variable auxiliaire (elle vaut 1 sinon). Les poids peuvent être calculés comme une mesure de la corrélation entre la variable d'intérêt et la variable auxiliaire ; lorsque les variables sont qualitatives, il n'est pas possible d'utiliser le coefficient de corrélation linéaire donc il faut avoir recours à des indicateurs de corrélation tels que le V de Cramer. Voir Vanderschelden (2005) pour plus de détails.

Troisième partie : Les nouvelles procédures de redressement mises en œuvre à l'enquête Logement 2006

A – LA CORRECTION DES MAUVAISES DÉCLARATIONS : REDRESSEMENTS EN DEUX TEMPS DES PLACEMENTS FINANCIERS ET DES CHARGES LOCATIVES

Les questions qualitatives de l'enquête Logement ne font en principe pas l'objet de redressements car les non-réponses (« Ne Sait Pas » et « Refus ») ne sont autorisées que pour les variables quantitatives. Cependant, à l'enquête 2006, on a dû corriger une partie des réponses négatives à la détention de placements financiers et au paiement de charges locatives. En effet, au vu des remarques des enquêteurs et de sources externes, la détention de placements financiers et le paiement de charges locatives sont sous-déclarés : certaines réponses négatives correspondent en réalité à des refus de répondre ou à une méconnaissance de la part du ménage. On doit donc mettre en œuvre des redressements en deux temps :
- dans un premier temps, correction de la sous-déclaration : affectation de détention de placements financiers à des ménages non-détenteurs, et du paiement de charges à des ménages ayant déclaré ne pas en payer (la réponse initiale du ménage est naturellement conservée parmi les données brutes). Pour les placements financiers, on procède à une imputation stochastique de la détention tandis que pour les charges, on se contente d'une imputation déterministe.
- dans un deuxième temps, imputation de l'ensemble des montants manquants, y compris pour les ménages « récupérés » à l'étape précédente.

I. LE REDRESSEMENT DES PLACEMENTS FINANCIERS

Le questionnement sur les placements financiers est organisé en deux temps : on recense les types de placements détenus par le ménage, puis on lui demande d'indiquer dans quelle tranche se situe le montant total de ces placements.

Les « Ne Sait Pas » et « Refus » n'étaient autorisés qu'à la question sur le montant. Cependant, un certain nombre de questionnaires indiquant que le ménage ne détient aucun placement financier comportent une remarque de l'enquêteur signalant qu'il s'agissait en réalité d'un refus de répondre de la part du ménage. En effet, comme ni le « Ne Sait Pas » ni le « Refus » n'étaient autorisés à la question sur la détention, l'enquêteur était obligé de choisir la modalité « aucun placement financier » lorsque le ménage refusait d'indiquer les types de placements qu'il possédait. Pour évaluer l'ampleur de cette sous-déclaration, on a confronté les résultats de l'enquête Logement 2006 avec une autre source : l'enquête Patrimoine 2004, qui fait davantage référence en ce qui concerne les taux de détention d'actifs financiers. Les résultats de cette comparaison figurent dans le tableau suivant :

	Enquête Patrimoine 2004	Enquête Logement 2006
Nombre total de ménages	24 737 732	25 001 715
Nombre de ménages détenteurs d'actifs financiers	22 566 904	18 688 436
% de ménages détenteurs	91,2	74,7

Selon l'enquête Patrimoine, neuf ménages sur dix disposent d'un patrimoine financier, sous forme d'épargne liquide (livrets d'épargne) ou d'épargne plus longue (épargne-logement, assurance-vie, épargne-retraite, valeurs mobilières)[26]. Or, d'après l'enquête Logement, seuls les trois quarts des ménages sont détenteurs d'un placement ; la sous-estimation du nombre de détenteurs est de l'ordre de quatre millions de ménages.

Il y a donc deux types de non-réponses partielles à corriger :
- la non-réponse à la tranche dans laquelle se situe le montant détenu, parmi les ménages ayant déclaré détenir des placements financiers (11 % d'entre eux n'ont pas indiqué la tranche dans laquelle il se situait).
- la non-réponse « déguisée » à la question sur les types de placements détenus

[26] Rougerie et Cordier (2004).

On met en place un redressement en deux étapes :

1) Redressement de la sous-déclaration de placements financiers : imputation de détention
Les remarques des enquêteurs ne suffisent pas pour redresser la variable. En effet, seuls certains d'entre eux ont émis un commentaire et, même en présence d'une telle remarque, on ne peut pas savoir si le ménage est ou non détenteur. Pour attribuer la détention de placements financiers à certains ménages ayant déclaré ne pas en avoir, on a donc recours à une source externe : l'enquête Patrimoine 2004.

2) Imputation de la tranche dans laquelle se situe le montant total détenu, pour les ménages :
- qui ont indiqué les placements qu'ils détenaient mais n'en ont pas donné le montant
- auxquels on a affecté de la détention lors de la première étape : imputation simultanée de types de placement et d'une tranche, en excluant des donneurs potentiels les donneurs et receveurs de l'étape précédente.

1.1. Première étape : correction de la non-réponse « déguisée » à la question sur la détention

L'objectif de la première étape est d'obtenir un nombre de ménages détenteurs cohérent avec les données issues de l'enquête Patrimoine. On modélise par une régression logistique la probabilité de détenir au moins un placement financier dans l'enquête Patrimoine, puis on importe le modèle dans l'enquête Logement afin d'attribuer par un tirage à probabilités inégales la détention de placements à certains des ménages qui n'en ont pas déclarés.

a) Estimation d'une équation de détention sur les données de l'enquête Patrimoine

Pour corriger la sous-déclaration de la détention de placements financiers, on cherche à sélectionner, parmi les ménages ayant déclaré ne pas détenir de placements, ceux pour lesquels cette réponse négative a des chances d'être une non-réponse déguisée. Pour cela, on recourt à une imputation économétrique sur la base d'une équation estimée sur une source exogène, l'enquête Patrimoine, car ses données semblent plus fiables en matière de placements financiers.

L'idée est d'estimer dans l'enquête Patrimoine une équation de la probabilité de détenir un placement puis d'appliquer ce modèle à l'enquête Logement. On modélise la probabilité de détention en effectuant une régression logistique, le but étant de déterminer la meilleure équation possible en se restreignant aux variables communes aux deux enquêtes (il y a donc des contraintes sur les variables explicatives[27]).

Le modèle s'écrit : $P(D=1) = F(\alpha_0 + \sum_{i=1}^{n} \alpha_i \mathrm{X}_i)$

où les X_i sont les variables explicatives prises en compte dans le modèle et les α_i les paramètres à estimer. F est la fonction de répartition d'une loi logistique : $F(t) = e^t/(1+e^t) = 1/(1+e^{-t})$.

Le pourcentage de paires concordantes, c'est-à-dire le taux de bien classés, est de 80,7 % pour le modèle retenu.

b) Calcul de la probabilité de détention pour les ménages de l'enquête Logement

On importe ensuite le modèle dans l'enquête Logement afin de calculer, à partir des coefficients estimés par la régression logistique, la probabilité de détenir un placement financier.

On calcule pour chaque ménage la probabilité de détenir un placement financier par la formule suivante :

$$p(D=1) = \frac{1}{1+\exp(-\hat{\alpha}_0 - \sum_{i=1}^{n} \hat{\alpha}_i \mathrm{X}_i)} = \frac{\exp(\hat{\alpha}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \mathrm{X}_i)}{1+\exp(\hat{\alpha}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \mathrm{X}_i)}$$

où $\hat{\alpha}_0$ est la constante obtenue par l'estimation du modèle logit sur les données de l'enquête Patrimoine, et $\hat{\alpha}_i$ le coefficient estimé relatif à la variable explicative X_i.

[27] Les variables retenues sont les suivantes : âge et diplôme le plus élevé de la personne de référence, statut d'occupation du logement, catégorie sociale de la personne de référence (ou statut si elle est au chômage ou à la retraite), détention de patrimoine immobilier, niveau de vie, taille de l'unité urbaine, composition du ménage.

Notons qu'on trouve bien des probabilités élevées pour les ménages ayant indiqué détenir des placements.

c) Tirage à probabilités inégales d'un échantillon de détenteurs parmi les ménages non-détenteurs de l'enquête Logement

La troisième phase de la première étape consiste à déterminer, parmi les ménages qui se sont déclarés non-détenteurs à l'enquête Logement, ceux auxquels on va affecter la détention d'un placement financier. Ces nouveaux détenteurs sont sélectionnés aléatoirement selon un tirage à probabilités inégales[28], les probabilités étant issues du modèle logit. Comme on doit imputer la détention de placements à environ quatre millions de ménages, on tire un échantillon de 6 000 ménages « faux non-détenteurs », soit 3 975 661 ménages en effectif pondéré. Parmi l'échantillon de nouveaux détenteurs identifiés, on retrouve bien des ménages au sujet desquels les enquêteurs avaient émis des doutes.

1.2. Deuxième étape : imputation des types de placements et des tranches de montants manquants par hot-deck aléatoire

Pour les ménages ayant indiqué les placements qu'ils détenaient mais n'ayant pas voulu ou pu donner la tranche du montant détenu, on procède par hot-deck aléatoire (cf. annexe 8). La répartition en tranches est peu modifiée par l'imputation.
Pour les ménages auxquels on a attribué la détention de placements financiers, on impute simultanément les types de placements et la tranche de montant associée. On procède par hot-deck aléatoire afin d'utiliser un même donneur pour imputer les deux variables[29], en constituant les classes à partir de la catégorie sociale, de l'âge et du revenu de la personne de référence. Les donneurs et receveurs du hot-deck précédent sont exclus des donneurs potentiels.

1.3. Comparaison des résultats avec l'enquête Patrimoine 2004

Le tableau suivant présente les proportions de ménages détenant les différents types de placements financiers dans les enquêtes Logement et Patrimoine :

	Enquête Logement 2006						**Enquête Patrimoine 2004**
	Avant imputation		**Après imputation…** **des tranches manquantes** uniquement pour les ménages ayant déclaré la détention de placements financiers		**Après imputation…** **de la détention**, des types de placement et des tranches manquantes		
	Effectif	%	Effectif	%	Effectif	%	%
Ensemble des ménages	**25 001 715**	**100,0**	**25 001 715**	**100,0**	**25 001 715**	**100,0**	**100,0**
Ménages non détenteurs	6 313 279	25,3	6 313 279	25,3	2 337 619	9,3	8,8
Ménages détenant au moins un placement financier	18 688 436	74,7	18 688 436	74,7	22 664 096	90,7	91,2
Placements détenus (parmi l'ensemble des ménages) :							
- livrets d'épargne exonérés	16 206 606	64,8	16 206 606	64,8	19 638 294	78,5	82,6
- livrets soumis à l'impôt	618 447	2,5	618 447	2,5	745 441	3,0	5,9
- épargne logement	6 989 592	28,0	6 989 592	28,0	8 360 032	33,4	41,3
- valeurs mobilières	4 252 909	17,0	4 252 909	17,0	5 018 651	20,1	24,2
- assurance-vie, épargne retraite	7 556 741	30,2	7 556 741	30,2	8 998 429	36,0	43,7
- bons d'épargne, bons anonymes, bons du trésor, de capitalisation…	142 672	0,6	142 672	0,6	164 842	0,7	1,0
- autre placement financier	875 438	3,5	875 438	3,5	1 075 053	4,3	16,7

L'estimation du nombre de ménages détenteurs est plus proche de celui de l'enquête Patrimoine après imputation qu'avant. Le redressement permet aussi de se rapprocher de l'enquête Patrimoine en ce qui concerne les proportions des ménages détenant les divers placements financiers : avant l'imputation de détention, seuls 65 % des ménages de l'enquête Logement possèdent un livret d'épargne exonéré ; après l'imputation, ils sont 79 % à détenir un tel placement, contre 83 % dans l'enquête Patrimoine.

[28] La mise en œuvre sous SAS s'effectue au moyen d'une PROC SURVEYSELECT avec l'option METHOD = PPS SYS (PPS Systematic Sampling). Elle sélectionne des unités en effectuant un tirage systématique aléatoire. Une méthode alternative mais plus déterministe aurait consisté à classer les ménages par probabilité décroissante et à sélectionner les ménages au-dessus d'un certain seuil de probabilité prédite. Mais le fait de tirer un échantillon à probabilités inégales permet de respecter le caractère « aléatoire » de la méthode mise en œuvre.

[29] Si on n'avait eu que les types de placement à imputer, on aurait pu utiliser un modèle polytomique non ordonné. Mais on souhaite ici imputer simultanément les types de placement et la tranche associée, d'où l'intérêt d'une méthode par donneur.

Par ailleurs, l'enquête Logement aboutit à des montants détenus un peu inférieurs à ceux de l'enquête Patrimoine. Le tableau ci-dessous permet de comparer la répartition des montants des placements financiers parmi les ménages détenteurs :

Enquête Logement 2006		Enquête Patrimoine 2004	
Moins de 3 000 €	33,1%	Moins de 3 000 €	25,7%
3 000 à moins de 7 000 €	16,2%	3 000 à moins de 7 500 €	17,4%
7 000 à moins de 15 000 €	17,3%	7 500 à moins de 15 000 €	15,5%
15 000 à moins de 30 000 €	12,0%	15 000 à moins de 30 000 €	14,8%
30 000 à moins de 50 000 €	7,7%	30 000 à moins de 45 000 €	8,0%
50 000 à moins de 100 000 €	7,1%	45 000 à moins de 75 000 €	6,9%
		75 000 à moins de 105 000 €	3,8%
100 000 à moins de 150 000 €	3,2%	105 000 à moins de 150 000 €	3,3%
150 000 à moins de 200 000 €	1,5%	150 000 à moins de 225 000 €	2,1%
Plus de 200 000 €	2,4%	225 000 à moins de 300 000 €	1,0%
		300 000 à moins de 450 000 €	0,7%
		Plus de 450 000 €	0,7%

II. LE REDRESSEMENT DES CHARGES LOCATIVES

L'objectif est de connaître le montant mensuel de charges payé par les ménages locataires ou logés gratuitement. La non-réponse n'était autorisée qu'au montant des charges mais il avait été remarqué lors des enquêtes Logement précédentes que le nombre de locataires ayant déclaré payer des charges était inférieur à celui auquel on aurait pu s'attendre. En effet, certains locataires ignorent payer des charges car ils versent un montant global sans disposer de la ventilation du montant payé entre loyers, taxes et charges locatives.

Pour corriger cette sous-déclaration, on procède en deux étapes :
- imputation de charges à des ménages locataires ayant déclaré ne pas en avoir mais qui doivent en payer au vu des caractéristiques de leur logement
- imputation des montants manquants pour les ménages ayant déclaré des charges mais n'en ayant pas donné le montant, ainsi que pour les ménages auxquels on a affecté des charges à l'étape précédente

La population à redresser comprend trois catégories de ménages :
- les ménages locataires ou logés gratuitement ayant déclaré payer des charges en dehors du loyer mais n'ayant pas donné leur montant
- les locataires en immeuble collectif ayant déclaré un loyer charges et/ou taxes comprises
- les locataires ayant déclaré ne pas payer de charges alors qu'ils doivent en payer au vu des caractéristiques de leur logement[30]

Bilan de la non-réponse :

	Effectif	Répondants	Non-répondants	% de non-réponses
Individuel	**1 582**	**1 537**	**45**	**2,8**
Locataires	1 522	1 483	39	2,6
Logés gratuit	60	54	6	10
Collectif	**14 031**	**11 120**	**2 911**	**20,7**
Locataires	11 037	10 914	123	1,1
Logés gratuit	224	206	18	8,0
Locataires ayant donné un loyer charges/taxes incluses	2 522	-	2 522	100,0
Locataires n'ayant pas déclaré payer des charges mais devant en payer au vu des caractéristiques de leur logement	248	-	248	100,0
Ensemble	**15 613**	**12 657**	**2 956**	**18,9**

2.1. Première étape : imputation du paiement de charges

A défaut de connaître le nombre de locataires payant des charges, on déduit la population des locataires susceptibles d'en payer à partir de l'existence d'éléments entraînant le paiement de charges. Sont considérés comme devant payer des charges les locataires qui résident dans un immeuble collectif[31] possédant l'un au moins des éléments d'équipement ou de service suivants : existence d'un

[30] A l'issue des imputations des charges, le montant imputé est déduit du loyer.
[31] Le paiement de charges est plus rare dans l'habitat individuel et parmi les ménages logés gratuitement.

ascenseur, présence d'un gardien, eau chaude collective, facture d'eau payée au propriétaire ou à un gérant, chauffage collectif.

2.2. Deuxième étape : imputation des montants de charges manquants par la méthode des résidus simulés

Une fois la sous-déclaration corrigée, il reste à imputer les montants manquants. On choisit de les imputer par la méthode des résidus simulés. Comme on dispose de davantage de variables explicatives pour les logements situés dans un immeuble collectif, on redresse les charges séparément selon que le logement relève de l'habitat individuel ou collectif.

a) Imputation du montant des charges locatives dans l'habitat collectif

Pour modéliser le montant des charges dans l'habitat collectif, on régresse le logarithme du montant mensuel des charges sur les variables explicatives suivantes : mode de paiement de l'eau, eau chaude collective, chauffage collectif, présence d'un ascenseur, d'un gardien, autres éléments de charges, surface du logement, taille d'unité urbaine, niveau de vie, année d'achèvement de la construction de l'immeuble, nombre d'étages de l'immeuble, date d'emménagement du ménage dans le logement, secteur locatif (HLM ou privé), nombre de pièces et nombre d'habitants du logement[32].

Pour les locataires en immeuble collectif qui ont indiqué un montant de loyer charges et/ou taxes comprises et pour ceux qui ont déclaré ne pas payer de charges, on fixe comme borne maximum du montant de charges un tiers du loyer (cela correspond au troisième quartile des ménages répondants aux charges et aux loyers). Comme le montant des charges doit être ensuite déduit du loyer, cela permet de garantir l'obtention d'un loyer hors charges et taxes positif et d'un montant suffisant.

Le coefficient de détermination est satisfaisant : $R^2 = 0,54$ (cf. annexe 9). On applique ensuite la macro %simul.

Comparaison de la distribution avant et après imputation :

	Avant	Après
Nombre de ménages	6 484 255	6 484 255
Effectif renseigné	5 176 865	6 484 255
Effectif non renseigné	1 307 390	-
% de non-réponses	20,2	-
Moyenne	89,8	82,3
Intervalle de confiance*	[86,2 ; 93,4]	[79,0 ; 85,6]
Écart-type	66,1	63,2
Minimum	1**	1
1er quartile	44	39
Médiane	77	68
3ème quartile	122	113
Maximum	1 152	1 152

* Les intervalles de confiance sont calculés sous l'hypothèse d'un sondage aléatoire simple[33]. Ils sont donnés à titre indicatif.
** Quelques ménages paient moins de 18 euros par an

L'imputation corrige à la baisse le montant des charges locatives dans l'habitat collectif (le montant moyen sort de l'intervalle de confiance calculé avant l'imputation). Cela s'explique par le fait que les montants de charges imputés aux locataires n'ayant pas déclaré en payer sont en moyenne inférieurs à ceux des autres locataires.

b) Imputation du montant des charges locatives dans l'habitat individuel

On régresse le logarithme du montant mensuel des charges sur les variables explicatives suivantes : mode de paiement de l'eau, autres charges, surface du logement, date d'emménagement du ménage dans le logement, taille d'unité urbaine, niveau de vie, année d'achèvement de la construction, secteur locatif, nombre de pièces et nombre d'habitants du logement. Le coefficient de détermination est également satisfaisant : $R^2 = 0,34$ (cf. annexe 9). On applique ensuite la macro %SIMUL.

[32] Cf. note n°119/F330 du 28 juin 2004.
[33] Cette hypothèse est d'autant plus acceptable qu'il avait été constaté à l'enquête Logement 1996 que la variance de l'enquête Logement était pratiquement égale à celle d'un sondage aléatoire simple. Pour la plupart des estimateurs, la variance estimée par POULPE pouvait être remplacée par une estimation approchée ne prenant pas en compte le plan de sondage, grâce notamment à la taille importante de l'échantillon. Voir Le Blanc (1998) pour plus de détails.

Comparaison de la distribution avant et après imputation :

	Avant	**Après**
Nombre de bénéficiaires	977 992	977 992
Effectif renseigné	956 213	977 992
Effectif non renseigné	21 779	-
% de non-réponses	2,2	-
Moyenne	30,0	29,7
Intervalle de confiance*	[29,0 ; 31,0]	[28,9 ; 30,5]
Écart-type	45,8	45,4
Minimum	1	1
1er quartile	10	10
Médiane	18	18
3ème quartile	30	30
Maximum	1 000	1 000

L'imputation ne modifie pas la distribution du montant des charges dans l'habitat individuel.

B - LE REDRESSEMENT DE VARIABLES RENSEIGNÉES EN CLAIR OU EN TRANCHES : LES REVENUS

L'objectif de la partie sur les revenus est de quantifier l'ensemble des ressources perçues par les différents individus du ménage au cours des douze derniers mois. Pour l'enquête Logement 2006, il a été nécessaire d'utiliser de nouvelles méthodes de redressement car des questions en tranches ont été ajoutées en cas de non-réponse aux montants en clair, la largeur des tranches proposées variant selon le type de revenu. On dispose donc d'une information supplémentaire importante en cas de non-réponse partielle au montant en clair. Les ménages pouvaient ne pas répondre à la question en tranches mais dans plus de la moitié des cas de non-réponse à la question en clair, les enquêtés ont accepté d'indiquer dans quelle tranche se situe le montant total perçu par l'ensemble du ménage, comme le montre le tableau suivant :

	Variables	% de non-réponses à la question en clair	% de non-réponses à la question en tranches (parmi les ménages n'ayant pas répondu en clair)
Revenus perçus au niveau individuel	salaires et primes	8,4	27,0
	indemnités de chômage	7,6	25,9
	retraites et pensions	11,7	41,5
	revenus non salariaux	23,8	39,3
Revenus perçus par le ménage	prestations familiales et liées au handicap aides à la scolarité d'un enfant	9,6 7,0	36,9
	revenus fonciers	23,9	46,8
	allocations RMI perçues par la personne de référence et/ou son conjoint allocations RMI perçues par une autre personne du ménage	4,6 9,5	36,0
	revenus de type « autre »	5,3	42,6

Le but des redressements des revenus est de reconstituer des montants en continu à partir de l'ensemble des déclarations (en clair ou en tranches) afin de disposer ensuite de caractéristiques de moyenne, de dispersion et de concentration. On pourrait envisager d'imputer par les centres des tranches : mais en utilisant cette méthode déterministe, on sous-estimerait la variabilité de la variable ainsi imputée puisqu'on ne restituerait pas la variabilité interne aux tranches (cela modifierait la distribution de la variable et créerait des « grumeaux ») ; en outre, on rencontrerait un problème pour la dernière tranche qui n'a pas de borne supérieure. On privilégie donc les deux méthodes aléatoires suivantes, qui permettent de prendre en compte l'appartenance à des tranches et possèdent de plus des propriétés intéressantes pour la distribution des variables : la méthode des résidus simulés et le hot-deck aléatoire. On doit redresser d'une part les revenus individuels, d'autre part les revenus de niveau ménage.

I. LE REDRESSEMENT DES REVENUS INDIVIDUELS : EXEMPLE DES SALAIRES

La difficulté posée par le redressement des revenus individuels est de prendre en compte de l'information disponible à deux niveaux, celui de l'individu et celui du ménage. Les montants en clair sont en effet demandés au niveau individuel alors que la question en tranches porte sur le montant perçu par l'ensemble du ménage. De plus, elle n'est posée que dans le cas où la personne de référence (et/ou son conjoint éventuel) a déclaré avoir perçu des salaires mais refuse ou ne peut donner le montant en clair correspondant. Par conséquent, on ne peut établir une correspondance entre la question en clair et celle en tranches que pour les ménages dans lesquels seul(e) la personne de référence ou son conjoint perçoit des salaires.

L'utilisation des tranches se révèle donc complexe pour les revenus individuels. Comme le redressement des salaires doit être fait au niveau individuel (puisque ce sont les caractéristiques de l'individu et non du ménage qui expliquent le montant perçu), on mobilise à deux moments l'information fournie par les tranches afin de faciliter le passage du niveau ménage au niveau individu :
- lors des imputations, on n'utilise l'information en tranches que si elle est à un niveau individuel, c'est-à-dire si seule la personne de référence ou son conjoint éventuel a perçu des salaires.
- un deuxième niveau d'utilisation des tranches intervient une fois que l'ensemble des salaires ont été imputés : pour les ménages dont on connaît la tranche, on vérifie que la somme des salaires s'y situe bien.

On tient donc bien compte de toutes les tranches renseignées[34].

On a par ailleurs profité de l'utilisation de méthodes permettant de prendre en compte les tranches pour corriger la non-réponse au montant des primes, qui peuvent être d'un montant non négligeable (aux enquêtes précédentes, seules les non-réponses au salaire étaient traitées). Pour les individus qui ont perçu des primes mais n'ont indiqué qu'un montant de salaires hors primes, on utilise en effet ce montant comme borne inférieure dans la procédure d'imputation, et le double de ce montant comme borne supérieure afin de ne pas imputer des montants de primes trop élevés[35].

Sont donc à redresser non seulement les non-répondants au salaire en clair, mais aussi les individus qui ont perçu des primes en plus des salaires et n'ont indiqué que le montant des salaires hors primes.

	Effectif	%
Individus ayant perçu des salaires	**38 217**	**100,0**
- montants en clair (salaires et éventuelles primes)	**34 995**	**91,6**
- non-réponses (aux salaires et/ou primes)	**3 222**	**8,4**
dont non-réponses pour lesquelles la tranche est connue*	709	1,9
*individus appartenant à un ménage pour lequel la tranche est connue***	*2 046*	*5,4*
ménages auxquels la question en tranches a été posée	**1 832**	**100,0**
- tranche renseignée	1 338	73,0
- Ne Sait Pas	191	10,4
- refus	303	16,5

* Il s'agit des ménages dans lesquels seule la personne de référence ou son conjoint a perçu des salaires. La question en tranches correspond alors au montant perçu par l'individu, et donc est utilisable au niveau individu (elle se situe au même niveau que la question en clair).
** Certains individus de ces ménages (autres que la personne de référence et son conjoint) peuvent être répondants puisque la question en tranches est posée dès lors que la personne de référence ou son conjoint est non-répondant.

Toutes choses égales par ailleurs, les cadres donnent moins souvent le montant de leur salaire

Pour déterminer quelques facteurs de la non-réponse aux salaires, on a modélisé la probabilité de donner le montant en clair au moyen d'une régression logistique (cf. annexe 10). Toutes choses égales par ailleurs, les cadres sont plus susceptibles de ne pas indiquer le montant de leur salaire. Le type d'emploi est également significatif : les salariés en CDD répondent moins que ceux qui ont un CDI. A autres caractéristiques comparables, les individus qui vivent seuls ont plus de chances de répondre. S'agissant de l'âge, les extrêmes ont tendance à moins répondre : être âgé de moins de 25 ans ou de 45 ans et plus va de pair avec une probabilité de réponse plus faible. Enfin, dans une moindre mesure, la taille de l'unité urbaine dans laquelle résident les salariés est également liée à la non-réponse aux salaires : ceux qui vivent en zone rurale ou en agglomération parisienne répondent davantage.

1.1. Préparation des données en vue des redressements

a) Distinction de trois catégories de salariés

On distingue trois catégories de salariés et anciens salariés[36] afin de mobiliser le maximum d'information auxiliaire disponible et d'utiliser pour chaque catégorie la méthode d'imputation la plus adaptée, tout en évitant de mélanger les différents profils de salariés. En effet, certaines questions sur l'activité professionnelle[37] n'ont été posées qu'à certains salariés alors qu'elles sont importantes pour redresser le montant des salaires ; il serait dommage de ne pas les utiliser lorsqu'elles sont disponibles.

Les trois catégories sont créées par croisement de la situation actuelle principale vis-à-vis du travail, du statut et de la catégorie sociale[38] :
- la première catégorie est la catégorie principale puisqu'elle comporte 95 % des salariés. Elle exclut les indépendants, les chefs d'entreprise salariés et les cas particuliers.
- la deuxième catégorie regroupe les indépendants et chefs d'entreprise salariés[39], soit 2 % des salariés.

[34] Certaines tranches n'ont pu être utilisées. En effet, il arrive qu'un membre du ménage autre que la personne de référence ou son conjoint ait donné son salaire alors que la personne de référence et/ou son conjoint n'a pas indiqué le(s) sien(leurs), mais que la tranche indiquée pour l'ensemble du ménage soit inférieure au salaire déclaré par l'individu. En l'absence de contrôles entre la tranche dans laquelle le ménage a déclaré se situer d'une part, et les montants individuels éventuellement renseignés d'autre part, on fait l'hypothèse que les salaires déclarés en clair sont plus fiables que la tranche qui porte sur l'ensemble du ménage (le répondant a pu croire à tort que la tranche ne portait que sur son salaire).
[35] Ainsi, les primes imputées ne peuvent pas être supérieures au montant du salaire donné par l'individu.
[36] Certains des individus ayant perçu des salaires au cours de l'année écoulée ne sont plus salariés au moment de l'enquête.
[37] Le type d'emploi et l'activité de l'établissement qui emploie l'individu ou que ce dernier dirige ne sont renseignés que pour les individus qui avaient un emploi salarié au moment de l'enquête.
[38] La catégorie sociale et le statut nous renseignent sur la situation de l'individu au moment de l'enquête (ou la plus récente s'il ne travaille pas actuellement) qui n'est pas forcément la même que lorsqu'il percevait des salaires (s'il vient de changer de travail par exemple).

- la troisième catégorie permet d'isoler les cas particuliers, soit 3 % des individus ayant perçu des salaires au cours de l'année écoulée. Leurs salaires semblent correspondre à un travail d'appoint ou à de « petits boulots » donc leurs montants risqueraient de parasiter l'imputation des autres salaires. Il s'agit essentiellement d'étudiants, d'hommes et femmes au foyer, de personnes handicapées et de personnes sans activité professionnelle de 60 ans et plus non retraitées.

b) Imputation préalable de l'activité de l'établissement par hot-deck aléatoire

L'activité de l'établissement qui emploie l'individu ou que ce dernier dirige apporte une information importante sur le montant des salaires mais n'est pas toujours renseignée. On peut néanmoins l'utiliser comme variable auxiliaire pour imputer les salaires des individus des catégories 1 et 2 car il y a seulement 6,7 % de non-réponses[40]. On impute les codes d'activité manquants en utilisant des méthodes de hot-deck aléatoires. On constitue des classes d'imputation à partir de la catégorie sociale et du statut ; comme la catégorie sociale détaillée n'est connue que pour les non-retraités, on procède à des hot-deck séparés pour les non-retraités et pour les retraités.

La macro %hotdeck fait du tirage avec remise donc on peut regarder combien de fois un donneur a été utilisé en moyenne ainsi que la part des donneurs utilisés plusieurs fois (cf. annexe 11) : la quasi-totalité des donneurs n'ont été utilisés qu'une seule fois. L'imputation modifie peu la répartition de l'activité de l'établissement :

	Avant		Après	
Activité	**Effectif**	**%**	**Effectif**	**%**
Effectif concerné	23 186 264	100,0	23 186 264	100,0
Effectif renseigné	21 595 722	93,3	23 186 264	100,0
Non-réponses	1 590 542	6,7	-	-
Répartition dans l'effectif renseigné				
Agriculture	463 185	2,1	500 520	2,2
Construction	1 904 272	8,6	2 035 855	8,8
Industrie	3 724 517	16,8	3 887 110	16,8
Tertiaire non marchand	7 338 558	33,2	7 603 587	32,8
Tertiaire marchand	8 708 316	39,3	9 159 191	39,5

1.2. Imputation des principaux montants de salaires par la méthode des résidus simulés

Pour la première catégorie, qui comprend 95 % des salariés, on trouve un modèle ayant un pouvoir explicatif satisfaisant (R^2 = 0,48) donc on choisit d'imputer les valeurs en clair manquantes par la méthode des résidus simulés. Elle permet de prendre en compte les tranches tout en utilisant un maximum de variables explicatives et évite d'introduire une distorsion dans la distribution des salaires. On travaille sur des sous-modèles séparés hommes – femmes car ce ne sont pas forcément les mêmes facteurs qui jouent dans le montant de leurs salaires : il vaut donc mieux estimer séparément les équations[41].

Présentation de l'imputation des salaires des hommes

On commence par construire une prédiction des salaires non déclarés en modélisant le logarithme du salaire dans le modèle linéaire gaussien : $\ln(Y) = X\beta + \sigma U$, où X est le vecteur des variables auxiliaires, σ l'écart-type des résidus et U suit une loi normale centrée réduite (σU est le vecteur des résidus).

Les variables explicatives[42] retenues

Les durées de perception des revenus durant les douze derniers mois n'étaient pas demandées car c'est le montant des revenus effectivement perçus au cours de l'année précédant l'enquête qui nous intéresse. On ne peut donc pas prendre en compte la durée de perception des salaires dans les imputations, ce qui pose problème pour les individus qui ont commencé à travailler récemment et n'ont perçu de salaire que sur une partie de l'année, mais aussi pour ceux qui ont perçu plusieurs types de

[39] Il s'agit plus précisément des indépendants, agriculteurs exploitants, artisans, commerçants et chefs d'entreprise (y compris ceux qui ne se sont pas classés comme indépendants), chefs d'entreprise salarié (individus qui sont chefs de leur entreprise et se versent un salaire), PDG, gérants minoritaires et associés.
[40] On ne l'utilise pas pour la catégorie 3 car elle manque pour 71 % des individus de cette catégorie.
[41] On aurait également pu effectuer des régressions indépendantes sur les populations des salariés des secteurs public et privé.
[42] Les variables habituellement prises dans les équations de salaires sont des descripteurs sociodémographiques des individus et des ménages (âge, sexe, être né à l'étranger ou en France, diplôme ou nombre d'années d'études, situation familiale, région de résidence ou taille de l'unité urbaine) et des caractéristiques de l'emploi exercé (catégorie sociale, secteur public ou privé, travail à temps plein ou partiel, activité, ancienneté dans l'entreprise, type de contrat). Voir Kramarz (2003), Koubi (2003), Meurs et Ponthieux (2006).

revenus[43] (salaires puis retraites par exemple). On inclut donc parmi les variables auxiliaires une indicatrice de la perception de revenus autres que les salaires[44]. En effet, procéder à des imputations séparées selon les types de revenus perçus en plus des salaires serait peu judicieux puisqu'on ne sait pas pendant combien de temps ces revenus ont été versés à l'individu.

On utilise donc comme variables explicatives la nationalité de l'individu, son statut matrimonial, son niveau de diplôme, la perception d'autres revenus en plus du salaire, le type d'emploi, le statut, la catégorie sociale, le secteur d'activité de l'établissement qui emploie l'individu ou que l'individu dirige, la taille de l'unité urbaine dans laquelle il réside, et le logarithme de l'âge de l'individu (l'effet de l'âge est marginalement décroissant donc il faut utiliser une fonction concave[45]).

Modélisation des salaires et calcul des prédictions

A défaut de disposer d'une méthode de régression robuste permettant de modéliser une variable d'intérêt partiellement renseignée en tranches, on choisit de combiner plusieurs types de modélisation.

Afin de détecter d'éventuels problèmes dans l'estimation de l'équation de salaire, on régresse le logarithme du salaire sur les variables explicatives par les MCO (cf. annexe 12). Au vu du coefficient de détermination R^2, la qualité de la régression est relativement satisfaisante étant donnée la nature des données (soumises à une variabilité individuelle) : 48 % de la variance totale du logarithme du salaire est expliquée par le modèle. Les résidus suivent une loi proche d'une loi normale et sont d'espérance nulle.

On effectue ensuite une régression robuste des montants en clair ; elle classe 7,2 % des salaires en atypiques, ce qui peut s'expliquer par l'hétérogénéité des données. Certains montants correspondent à des salaires perçus sur une année, d'autres sur quelques mois seulement : comme on ne peut pas prendre en compte la durée comme variable explicative, il y a davantage de « bruit ». La quasi-totalité des montants repérés lors de l'analyse exploratoire figurent parmi elles ; on sélectionne au final un sous-ensemble de 95 valeurs atypiques, soit 0,6 % des répondants, que l'on écarte des imputations.

Enfin, pour calculer les prédictions, on utilise une PROC LIFEREG afin de pouvoir régresser non seulement les salaires en clair non atypiques mais aussi ceux donnés en tranches.

Comparaison des PROC LIFEREG et ROBUSTREG

Pour mesurer l'apport de l'information en tranches pour les salaires, on a comparé les résultats des régressions du logarithme du salaire par les MCO et par la PROC ROBUSTREG à ceux obtenus avec la PROC LIFEREG (cf. annexe 12). Les coefficients estimés sont très proches quelle que soit la modélisation utilisée, sans doute parce que peu de tranches sont renseignées par rapport au nombre de réponses en clair : parmi les salariés ayant indiqué un montant (en clair ou en tranches), seuls 6,3 % des hommes et 4,8 % des femmes ont donné une réponse en tranches. Elles apportent donc peu d'information dans la PROC LIFEREG.

Simulation des résidus et calcul des salaires imputés

Une fois les prédictions de salaires calculées, on simule les résidus à partir d'une distribution normale car la loi des résidus observés sur les répondants est proche d'une loi normale. Par contre, d'après le test de White, il faut rejeter l'hypothèse d'homoscédasticité (cf. annexe 11). Selon les tests de Breusch-Pagan, c'est la variable relative au type d'emploi (CDD, CDI à temps complet, CDI à temps partiel ou autre) qui est le plus en cause, ce que confirme l'examen des écart-types des résidus selon le contrat de travail. On pourrait envisager de faire des estimations différentes selon le type d'emploi mais cela alourdirait les redressements tandis que le gain en terme d'adéquation du modèle, et donc de prédiction, risquerait d'être faible. On a essayé de grouper le type d'emploi en deux postes mais le problème d'hétéroscédasticité subsiste quelle que soit la manière dont on regroupe les modalités. Pour en tenir compte lors de la simulation des résidus, on stratifie ces derniers par la variable relative au type d'emploi dans la table en entrée de la macro %simul[46].

[43] 11 % des salariés ou anciens salariés ont déclaré avoir bénéficié d'autres types de revenus au cours de l'année précédant l'enquête. Il s'agit essentiellement de changements de situation (perte d'emploi, reprise d'activité, départ en retraite…).

[44] On ne prend pas en compte le montant des autres revenus perçus mais seulement le fait d'avoir perçu ou non d'autres revenus ; en effet, les salaires sont redressés en premier donc, à ce stade, le montant des autres revenus n'est pas connu pour tous les individus puisqu'il a pu faire l'objet de non-réponse partielle (par contre, pour le redressement des autres revenus, on peut utiliser le montant du salaire comme variable explicative).

[45] On aurait aussi pu prendre une fonction quadratique.

[46] On les stratifie par l'intermédiaire de la variable SIGMA : on calcule les écart-types des résidus observés pour chaque strate, c'est-à-dire chacun des quatre types d'emploi (SIGMA prend donc quatre valeurs au lieu d'une seule).

On tente enfin de reconstituer la variable latente : pour chaque enregistrement à imputer, on simule un résidu dans la loi normale d'espérance nulle et de variance égale à la variance empirique des résidus des répondants appartenant à la même strate de type d'emploi, puis on l'ajoute à la prédiction. Ce résidu doit être tel que le logarithme du salaire total imputé soit dans la tranche éventuellement déclarée : si le logarithme ainsi imputé se situe bien dans la tranche, on l'impute ; sinon on tire un nouveau résidu jusqu'à obtenir une valeur appartenant à la tranche. Enfin, on calcule le salaire comme l'exponentielle du montant imputé : $\hat{Y}_i = \exp(X_i \hat{\beta} + \hat{\sigma} U_i)$.

Au final, l'imputation ne modifie pas la distribution des salaires des hommes :

	Avant	**Après**
Nombre d'individus	11 497 939	11 497 939
Effectif renseigné	10 438 032	11 497 939
Effectif non renseigné	1 059 907	-
% de non-réponses	9,2	-
Moyenne	21 534,4	21 679,5
Intervalle de confiance*	[21 298 ; 21 771]	[21 453 ; 21 906]
Écart-type	15 392,6	15 461,4
Minimum	15	15
1er quartile	14 000	13 940
Médiane	18 200	18 294
3ème quartile	25 005	25 527
Maximum	370 000	370 000

* Calculé sous l'hypothèse d'un sondage aléatoire simple

Le fait d'avoir simulé les résidus dans la loi normale semble avoir permis d'éviter d'imputer des valeurs extrêmes. La valeur la plus élevée après imputation n'est pas aberrante (elle n'est pas modifiée par rapport à la valeur avant imputation) et les différences que l'on observe au niveau de certains petits domaines sont faibles : par exemple, le salaire maximum des hommes occupant une profession intermédiaire, qui s'élève à 200 000 € parmi les répondants, est de 202 129 € après imputation.

On procède de la même façon pour les femmes.

1.3. Imputation des salaires restants par hot-deck aléatoire par classes

On redresse ensuite les salaires des deux autres catégories de salariés. Les indépendants et chefs d'entreprise salariés sont traités isolément car on dispose de moins d'information sur eux que sur les individus de la première catégorie. De plus, le montant de leurs salaires est difficile à modéliser donc on procède à une imputation par hot-deck aléatoire par classes. Les variables qui expliquent le mieux le montant de leurs salaires sont l'âge, la perception de revenus en plus du salaire et le statut (agriculteur, indépendant, chef d'entreprise salarié) : on les utilise donc pour constituer les classes d'imputation.

Quant aux cas particuliers, on a peu d'information sur l'activité qui leur a rapporté des salaires car moins de questions leur ont été posées et les variables explicatives sont moins bien renseignées : on choisit donc également une imputation par hot-deck aléatoire par classes. Les variables utilisées pour constituer les classes sont l'âge, le type d'emploi et le fait d'être ou non étudiant.

1.4. Vérification de la cohérence des salaires individuels avec les tranches de niveau ménage

Enfin, à l'issue des imputations, on regarde dans quelle tranche se situe la somme des salaires perçus par l'ensemble des individus du ménage et on la compare à la tranche éventuellement déclarée : il s'agit du deuxième niveau d'utilisation des tranches. Jusqu'à présent, pour les ménages dans lesquels plusieurs personnes travaillent, aucune contrainte n'a été imposée pour que la somme des salaires individuels se situe dans la tranche éventuellement déclarée par le ménage. Dans le cas où on aurait simulé des résidus trop grands pour certains des individus de ces ménages, l'information apportée par la tranche permettrait de le repérer.

Il y a une différence pour 495 des 1 338 ménages qui avaient indiqué une tranche[47] ; toutefois, pour la moitié d'entre eux, les tranches sont immédiatement voisines (seuls 261 ménages ont un écart de plus d'une tranche). Pour la majorité des écarts constatés, les montants imputés sont supérieurs à la tranche mais ils ne sont pas beaucoup plus élevés. Pour qu'au final les salaires perçus par l'ensemble des membres du ménage se situent dans la tranche déclarée, on répartit les montants imputés de manière à ce que le nouveau total des salaires perçus par l'ensemble du ménage soit égal à la borne la

[47] Les ménages pour lesquels on observe des incohérences sont naturellement des ménages qui comportent plusieurs salariés. Lorsqu'un seul individu du ménage était salarié, l'information en tranches disponible était utilisable au niveau individuel donc prise en compte dans les macros %simul ou %hotdeck.

plus proche du total obtenu « spontanément » à l'issue des imputations. Ainsi, on ne modifie pas les valeurs données en clair.

A l'issue de cette étape, tous les salaires individuels sont cohérents avec la tranche éventuellement déclarée par le ménage. Les redressements élèvent un peu la moyenne des salaires mais celle-ci reste dans l'intervalle de confiance calculé avant l'imputation. La distribution[48] est peu modifiée, à l'exception des premiers déciles qui diminuent légèrement :

	Avant imputation	Après imputation
Moyenne	18 558	18 661
Intervalle de confiance*	[18 401 ; 18 715]	[18 510 ; 18 812]
Minimum	15	15
1er décile	5 245	5 208
2e décile	9 900	9 600
3e décile	12 500	12 300
4e décile	14 400	14 400
5e décile	16 500	16 500
6e décile	18 283	18 300
7e décile	21 000	21 085
8e décile	24 700	25 000
9e décile	31 500	32 000
Maximum	483 500	483 500

* Calculé sous l'hypothèse d'un sondage aléatoire simple.

II. LE REDRESSEMENT DES REVENUS DE NIVEAU MÉNAGE : EXEMPLE DES ALLOCATIONS RMI

Les questions portant sur les allocations RMI perçues par le ménage sont organisées de la manière suivante. On demande le montant des allocations dont a bénéficié la personne de référence et/ou son éventuel conjoint, ainsi que le montant perçu par les autres personnes du ménage. En cas de « Ne Sait Pas » ou de « Refus » à la question sur le montant perçu par la personne de référence et/ou son conjoint, on demande dans quelle tranche se situe le montant total perçu par l'ensemble des membres du ménage (y compris les autres personnes du ménage[49]).

Bilan de la non-réponse :

	Effectif	**%**
ménages dont la personne de référence ou le conjoint perçoit le RMI	1 877	100,0
- montants en clair	1 791	95,4
- non-réponses	86	4,6
ménages percevant le RMI au titre d'une autre personne que la personne de référence ou le conjoint	273	100,0
- montants en clair	247	90,5
- non-réponses	26	9,5
question en tranches posée	86	100,0
- tranche connue pour l'ensemble du ménage	55	64,0
- non-réponses	31	36,0

Un essai d'imputation par barème a été réalisé mais cette méthode n'a pas été retenue car on ne sait pas sur combien de temps portent les ressources déclarées, ni de quand date la demande. Or la situation du ménage a pu changer depuis : en essayant d'estimer les ressources à partir des revenus déclarés pour les douze derniers mois, on s'aperçoit en effet qu'un certain nombre de ménages non-répondants ont des ressources supérieures au seuil et n'auraient donc pas dû percevoir le RMI. Comme il est par ailleurs difficile de trouver un modèle explicatif du montant du RMI, on s'oriente vers des imputations par hot-deck. On impute séparément les montants perçus par la personne de référence et/ou son conjoint, pour lesquels on procède à un hot-deck aléatoire par classes, et ceux perçus par d'autres personnes du ménage, pour lesquels on met en œuvre un hot-deck séquentiel.

[48] Ces résultats sont calculés avec les poids définitifs. De nombreux salaires ont manifestement été arrondis (12 000 €, 18 000 €…), ce qui forme des grumeaux dans la distribution.

[49] La question en tranches est donc mal placée dans le questionnaire : elle porte sur le montant perçu par l'ensemble du ménage mais est seulement posée en cas de non-réponse de la personne de référence et/ou de son conjoint (elle devrait se situer après la question sur le montant perçu par les autres personnes du ménage, ce qui aurait permis d'avoir des tranches en cas de non-réponse à cette question).

2.1. Imputation par hot-deck aléatoire des allocations RMI perçues par la personne de référence (et/ou son conjoint éventuel)

Pour imputer les allocations perçues par la personne de référence et/ou son conjoint, on a cherché à utiliser des variables de classes proches des critères de versement du RMI[50] : ressources du ménage au cours de l'année écoulée, type de famille, déduction d'un « forfait logement » (ménage propriétaire ou logé gratuitement) ou perception d'une aide au logement. On utilise ensuite la macro %hotdeck afin de tenir compte de l'information en tranches.

Comparaison de la distribution avant et après imputation :

	Avant	Après
Nombre d'allocataires	581 440	581 440
Effectif renseigné	541 705	581 440
Effectif non renseigné	39 735	-
% de non-réponses	6,8	-
Moyenne	3 435,8	3 379,2
Intervalle de confiance*	[3 330 ; 3 541]	[3 276 ; 3 482]
Écart-type	2 235,4	2 227,1
Minimum	40	40
1er quartile	1 275	1 275
Médiane	3 659	3 540
3ème quartile	4 836	4 812
Maximum	13 032	13 032

* Calculé sous l'hypothèse d'un sondage aléatoire simple

La distribution est peu modifiée : l'imputation diminue légèrement la moyenne mais celle-ci reste comprise dans l'intervalle de confiance calculé avant imputation.

2.2. Imputation par hot-deck séquentiel des allocations RMI perçues par les autres personnes du ménage

Peu de montants sont à imputer et il est difficile de déterminer des variables auxiliaires. Plutôt que de choisir une méthode totalement aléatoire comme lors de l'enquête précédente (hot-deck aléatoire sans variables de classe), on choisit une méthode d'imputation par hot-deck séquentiel. Elle permet de prendre en compte le peu d'information auxiliaire disponible tout en présentant l'avantage d'être très simple à programmer. Elle repose sur le tri du fichier selon des variables auxiliaires : ici, on choisit le type de famille et le niveau de revenus.

Au final, l'imputation ne modifie pas la distribution :

	Avant	Après
Nombre d'allocataires	124 755	124 755
Effectif renseigné	115 348	124 755
Effectif non renseigné	9 407	-
% de non-réponses	7,5	-
Moyenne	3 531,5	3 520,7
Intervalle de confiance*	[3 210 ; 3 853]	[3 216 ; 3 825]
Écart-type	2 529,9	2 517,9
Minimum	100	100
1er quartile	1 320	1 320
Médiane	3 600	3 600
3ème quartile	4 572	4 560
Maximum	16 464	16 464

* Calculé sous l'hypothèse d'un sondage aléatoire simple

[50] Les ressources prises en compte pour déterminer le montant du RMI sont celles de l'intéressé, de son éventuel conjoint et des personnes à sa charge : indemnités journalières de sécurité sociale, allocations de chômage, retraites et pensions, prestations familiales, allocation aux adultes handicapés, revenus issus de biens mobiliers et immobiliers et de capitaux et, pour une valeur forfaitaire, revenus d'activité ou de stages, ainsi qu'une partie des aides au logement. Si l'intéressé est propriétaire ou logé gratuitement, un « forfait logement » est déduit de l'allocation RMI.

III. VÉRIFICATION DES IMPUTATIONS DES REVENUS : COMPARAISON AVEC D'AUTRES ENQUÊTES AUPRÈS DES MÉNAGES

A l'issue des imputations de revenus, on procède à des vérifications au niveau des types de revenus d'une part, et à un niveau macroéconomique d'autre part. On compare les distributions issues de l'enquête Logement à celles obtenues dans d'autres enquêtes auprès des ménages[51].

3.1. Vérification de la distribution des différents types de revenus

On compare les salaires issus des redressements à ceux de l'enquête revenus Fiscaux 2005 :

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005
Nombre d'individus	24 366 563	26 287 529
Moyenne	18 661	18 618
Intervalle de confiance*	[18 510 ; 18 812]	[18 436 ; 18 800]
Minimum	15	1
1er décile	5 208	3 151
2e décile	9 600	7 694
3e décile	12 300	11 882
4e décile	14 400	14 430
5e décile	16 500	16 492
6e décile	18 300	18 665
7e décile	21 085	21 497
8e décile	25 000	25 377
9e décile	32 000	32 640
Maximum	483 500	847 620

* Calculé sous l'hypothèse d'un sondage aléatoire simple.

La distribution est proche, à l'exception des premiers déciles qui sont un peu plus élevés dans l'enquête Logement.

S'agissant du RMI, on a les résultats suivants :

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005
Nombre de ménages*	631 782	862 976
Moyenne	3 415	4 139
Intervalle de confiance**	[3 305 ; 3 525]	[4 026 ; 4 252]
Minimum	40	665
1er décile	500	1 519
2e décile	1 140	2 296
3e décile	1 800	3 170
4e décile	2 640	4 038
5e décile	3 600	4 644
6e décile	4 400	4 644
7e décile	4 572	4 644
8e décile	5 196	5 174
9e décile	6 300	6 661
Maximum	13 032	17 235

* On est au niveau ménage et non plus au niveau allocataire
** Calculé sous l'hypothèse d'un sondage aléatoire simple

On observe des écarts plus importants, notamment sur les premiers déciles, et les intervalles de confiance associés à la moyenne ne se recouvrent pas. Ces différences peuvent s'expliquer par le fait que les ménages vivant en foyer ne sont pas interrogés à l'enquête Logement (seuls les logements ordinaires sont enquêtés) : il manque d'ailleurs environ 200 000 bénéficiaires par rapport à l'enquête Revenus Fiscaux 2005.

Les comparaisons effectuées pour les autres types de revenus figurent en annexe 13. Les distributions sont généralement comprises entre celles de l'enquête Revenus Fiscaux 2005 et de SRCV 2005.

[51] Les résultats présentés dans cette partie ont été calculés avec les poids définitifs.

3.2. Vérification de la distribution du revenu total disponible

La distribution du revenu total disponible[52] des ménages de l'enquête Logement est dans l'ensemble proche de celle observée sur l'enquête Revenus Fiscaux 2005[53], même si les bornes de l'intervalle de confiance associé à la moyenne sont plus basses dans le cas de l'enquête Logement :

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005
Moyenne Intervalle de confiance*	29 911,2 [26 661 ; 30 161]	30 412,9 [30 116 ; 30 709]
Écart-type	24 425,2	27 936,0
Minimum	0	0
1er décile	8 010	9 565
2e décile	12 996	13 807
3e décile	16 779	17 187
4e décile	20 520	20 850
5e décile	25 150	24 842
6e décile	30 000	29 442
7e décile	35 284	34 761
8e décile	42 450	42 018
9e décile	54 805	54 856
Maximum	837 000	987 372
Rapport interdécile[54]	6,8	5,7

* Calculé sous l'hypothèse d'un sondage aléatoire simple
Champ : ménages ayant des revenus non négatifs

[52] A l'enquête Logement, on ne dispose pas à proprement parler du revenu disponible mais plus précisément du revenu total perçu avant imposition.
[53] S'agissant de la répartition du revenu disponible selon la catégorie sociale, le statut d'occupation du logement ou encore la structure familiale, on observe des tendances comparables dans l'enquête Logement et dans l'ERF : par exemple, les personnes seules et les familles monoparentales sont bien surreprésentées dans les premiers déciles, et les couples dans les derniers déciles.
[54] Ce ratio, qui est souvent utilisé pour mettre en évidence les écarts entre les plus riches et les plus pauvres, rapporte le revenu au-dessus duquel se situent les 10 % de ménages les plus riches (D9) au revenu en-dessous duquel se situent les 10 % les plus pauvres (D1).

Quatrième partie : Comparaison de l'imputation par la méthode des résidus simulés et par hot-deck aléatoire par classes

En prolongement des redressements décrits précédemment, on compare la méthode des résidus simulés à celle du hot-deck aléatoire par classes. On s'intéresse d'une part aux effets des méthodes utilisées sur la distribution des variables d'intérêt à un niveau agrégé et sur de petits domaines, et d'autre part à leur impact sur des ratios susceptibles de comporter des variables redressées au numérateur et au dénominateur.

A cet effet, on a réalisé des imputations par hot-deck pour certaines des variables redressées par les résidus simulés. Il s'agit des montants suivants : salaires de la première catégorie, indemnités de chômage, retraites, dépenses en eau et en énergies.

I. A UN NIVEAU AGRÉGÉ, LES DISTRIBUTIONS SONT TRÈS PROCHES

Bien que la méthode des résidus simulés prenne davantage d'information auxiliaire en compte que le hot-deck aléatoire par classes, les deux méthodes conduisent à des distributions très proches. Elles sont par ailleurs stables par rapport à la distribution observée avant imputation (cf. annexe 14) : l'utilisation de méthodes d'imputation aléatoire permet donc bien d'éviter une distorsion de la répartition de la variable d'intérêt.

L'imputation de certains revenus par hot-deck plutôt que par la méthode des résidus simulés a au final très peu d'impact sur la distribution du revenu total disponible des ménages de l'enquête Logement :

	Avec les méthodes retenues	*Avec hot-deck pour salaires, indemnités de chômage et retraites*
Moyenne	29 911,2	*29 796,3*
Écart-type	24 425,2	*24 241,0*
Intervalle de confiance*	[29 661 ; 30 161]	*[29 548 ; 30 044]*
Minimum	0	*0*
1er décile	8 010	*8 100*
2e décile	12 996	*13 000*
3e décile	16 779	*16 769*
4e décile	20 520	*20 498*
5e décile	25 150	*25 133*
6e décile	30 000	*30 000*
7e décile	35 284	*35 107*
8e décile	42 450	*42 300*
9e décile	54 805	*54 754*
Maximum	837 000	*837 000*
Rapport interdécile	6,8	*6,8*

* Calculé sous l'hypothèse d'un sondage aléatoire simple
Champ : ménages ayant des revenus non négatifs

L'absence de différences majeures entre la méthode des résidus simulés et le hot-deck aléatoire peut s'expliquer par le fait que les taux de non-réponse sont dans l'ensemble faibles et que l'échantillon est de taille importante.

II. SUR DE PETITS DOMAINES, DES DIFFÉRENCES APPARAISSENT SELON LES MÉTHODES UTILISÉES

Afin de voir si les méthodes d'imputation utilisées ont davantage d'effets sur des sous-populations, on a regardé comment se distribuent les salaires de la première catégorie sur des domaines créés par croisement de la catégorie sociale et du sexe, qui sont deux des critères pris en compte pour le redressement des salaires.

Dans la grande majorité des cas, les écarts entre la distribution avant et après imputation sont minimes, et la méthode des résidus simulés aboutit à des résultats proches de ceux obtenus par hot-deck. On constate cependant des différences pour certaines catégories, comme par exemple pour les professions de l'information, des arts et des spectacles :

Sexe	Salaire	Moyenne	Intervalle de confiance*	Écart-type	Minimum	1er quartile	Médiane	3ème quartile	Maximum
Hommes	Avant imputation	21 350,3	[21 120 ; 21 580]	14 982,9	700	9 600	20 000	29 800	98 059
	Après résidus simulés	21 252,2	[21 031 ; 21 474]	15 105,1	700	9 046	19 548	29 000	98 059
	Après hot-deck	*21 527,4*	*[21 293 ; 21 762]*	*15 979,4*	*700*	*8 900*	*18 916*	*30 000*	*98 059*
Femmes	Avant imputation	17 679,1	[17 517 ; 17 841]	10 369,2	142	10 600	16 800	24 000	60 000
	Après résidus simulés	16 689,0	[16 533 ; 16 845]	10 362,2	142	8 685	15 800	23 400	60 000
	Après hot-deck	*16 848,3*	*[16 695 ; 17 002]*	*10 202,3*	*142*	*9 743*	*16 191*	*22 881*	*60 000*

* Calculé sous l'hypothèse d'un sondage aléatoire simple

Cet exemple permet de voir que l'imputation n'est pas neutre : le salaire moyen des femmes après imputation sort de l'intervalle de confiance calculé avant imputation.

Il montre également qu'il est important de tenir compte des intervalles de confiance lors de l'inférence, notamment si l'on souhaite calculer des évolutions par rapport aux enquêtes antérieures. En effet, après l'imputation par les résidus simulés, le salaire moyen des hommes diminue légèrement par rapport à sa valeur avant imputation, tandis que le hot-deck tend au contraire à le faire augmenter ; mais les deux valeurs ne sortent pas de l'intervalle de confiance calculé sur les données renseignées.

III. LES RÉSULTATS PORTANT SUR PLUSIEURS VARIABLES IMPUTÉES SONT À TRAITER AVEC PRÉCAUTION : EXEMPLE DES TAUX D'EFFORT

Les corrélations entre variables peuvent être modifiées par les redressements lorsqu'elles ont été imputées indépendamment : on peut penser que lorsque les résultats portent sur plusieurs variables imputées (analyses multivariées), ils sont sensibles aux méthodes utilisées.

On s'intéresse donc à l'impact des méthodes d'imputation sur des ratios susceptibles de comporter des variables imputées à la fois au numérateur et au dénominateur : c'est précisément le cas des taux d'effort, qui représentent la part de budget du ménage consacrée aux dépenses de logement. Pour estimer le taux d'effort d'une sous-population, on calcule la moyenne des dépenses en logement et la moyenne des revenus sur cette sous-population, puis le rapport entre les deux :

$$\text{taux d'effort (\textit{brut ou net})} = \frac{\text{moyenne des dépenses annuelles en logement (\textit{brutes ou nettes}) des ménages}}{\text{moyenne des revenus annuels des ménages}}$$

charge financière *brute* = loyers ou remboursements d'emprunts, dépenses en eau et en énergies
charge financière *nette* = charge financière brute – aides au logement

Pour évaluer les effets des méthodes d'imputation par les résidus simulés et par hot-deck, on calcule :
- les taux d'effort obtenus avec les méthodes d'imputations qui ont été retenues pour l'enquête
- les taux d'effort obtenus en utilisant pour certains montants des imputations par hot-deck plutôt que par la méthode des résidus simulés (salaires de la première catégorie, indemnités de chômage, retraites, dépenses en eau et en énergies).

Les méthodes d'imputation utilisées ont peu d'impact si l'on travaille sur des domaines de taille suffisante. Lorsque l'on calcule les taux d'effort pour l'ensemble de la population, elles mènent à des intervalles de confiance identiques et de faible amplitude :

Taux brut (%)		**Taux net (%)**	
Avec les méthodes retenues	*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*	Avec les méthodes retenues	*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*
17,8	*17,8*	16,5	*16,6*
[17,6 ; 18,0]*	*[17,6 ; 18,0]**	[16,3 ; 16,7]*	*[16,4 ; 16,8]**

* Intervalles de confiance calculés sous l'hypothèse d'un sondage aléatoire simple (cf. annexe 16)

Les taux d'effort calculés pour des sous-populations de taille importante sont également peu affectés par la méthode d'imputation utilisée (cf. annexe 15). Des écarts apparaissent néanmoins pour certains domaines dont l'effectif est plus réduit. Par exemple, pour les ménages dont la personne de référence a moins de 20 ans, le taux d'effort s'élève à 31,4 % avec les méthodes retenues pour l'enquête contre 31,8 % si l'on utilise les valeurs imputées par hot-deck pour certains montants, et les intervalles de confiance ne se recouvrent pas totalement :

	Taux brut (%)				**Taux net (%)**			
	Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*		Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*	
Taux d'effort	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**
Selon l'âge de la personne de référence :								
Moins de 20 ans	31,4	[30,8 ; 32,0]	*31,8*	*[31,2 ; 32,4]*	27,2	[26,7 ; 27,7]	*27,5*	*[27,0 ; 28,0]*
20 à 29 ans	24,2	[23,9 ; 24,5]	*24,1*	*[23,8 ; 24,4]*	22,3	[22,0 ; 22,6]	*22,3*	*[22,0 ; 22,6]*
30 à 39 ans	18,5	[18,2 ; 18,8]	*18,5*	*[18,2 ; 18,8]*	17,3	[17,0 ; 17,6]	*17,3*	*[17,0 ; 17,6]*
40 à 49 ans	12,9	[12,6 ; 13,2]	*13,0*	*[12,7 ; 13,3]*	12,2	[11,9 ; 12,5]	*12,3*	*[12,0 ; 12,6]*
50 ans et plus	13,5	[13,1 ; 13,9]	*13,7*	*[13,3 ; 14,1]*	12,9	[12,5 ; 13,3]	*13,0*	*[12,6 ; 13,4]*
Selon le statut d'occupation :								
Accédant à la propriété	24,2	[23,9 ; 24,5]	*24,1*	*[23,8 ; 24,4]*	23,8	[23,5 ; 24,1]	*23,8*	*[23,5 ; 24,1]*
Propriétaire non accédant	6,5	[6,4 ; 6,6]	*6,5*	*[6,4 ; 6,6]*	6,5	[6,4 ; 6,6]	*6,5*	*[6,4 ; 6,6]*
Usufruitier	7,8	[7,2 ; 8,4]	*7,8*	*[7,2 ; 8,4]*	7,8	[7,2 ; 8,4]	*7,8*	*[7,2 ; 8,4]*
Locataire ou sous-locataire	30,5	[30,2 ; 30,8]	*30,8*	*[30,5 ; 31,1]*	26,0	[25,8 ; 26,2]	*26,2*	*[26,0 ; 26,4]*
Logé gratuitement	5,9	[5,5 ; 6,3]	*5,8*	*[5,4 ; 6,2]*	5,9	[5,5 ; 6,3]	*5,8*	*[5,4 ; 6,2]*

***** Calculé sous l'hypothèse d'un sondage aléatoire simple

A un niveau encore plus fin, on constate davantage de différences. Par exemple, pour certaines petites catégories issues du croisement du statut d'occupation et de l'âge de la personne de référence, les taux varient selon les méthodes d'imputation utilisées et les intervalles de confiance ne se recouvrent que partiellement :

Statut d'occupation	**Age de la personne de référence**	**Taux brut (%)**				**Taux net (%)**			
		Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*		Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*	
		Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**
Locataire ou sous-locataire	Moins de 20 ans	35,1	[34,4 ; 35,8]	*35,6*	*[34,9 ; 36,3]*	29,1	[28,5 ; 29,7]	*29,5*	*[28,9 ; 30,1]*
	50 à 59 ans	36,3	[35,1 ; 37,5]	*37,3*	*[36,1 ; 38,5]*	32,4	[31,3 ; 33,5]	*33,3*	*[32,2 ; 34,4]*
Logé gratuitement	Moins de 20 ans	6,5	[5,6 ; 7,4]	*6,7*	*[5,8 ; 7,6]*	6,5	[5,6 ; 7,4]	*6,7*	*[5,8 ; 7,6]*
	30 à 39 ans	4,9	[4,4 ; 5,4]	*4,6*	*[4,1 ; 5,1]*	4,9	[4,4 ; 5,4]	*4,6*	*[4,1 ; 5,1]*
	40 à 49 ans	4,7	[3,9 ; 5,5]	*4,9*	*[4,1 ; 5,7]*	4,7	[3,9 ; 5,5]	*4,9*	*[4,1 ; 5,7]*
	50 à 59 ans	10,9	[9,2 ; 12,6]	*10,5*	*[8,8 ; 12,2]*	10,9	[9,2 ; 12,6]	*10,9*	*[9,2 ; 12,6]*

* Calculé sous l'hypothèse d'un sondage aléatoire simple

Les estimations sont donc de moins bonne qualité lorsque l'on travaille sur de petites sous-populations. Par conséquent, si l'on souhaite calculer des taux d'effort sur des catégories de taille réduite, il est préférable d'utiliser des méthodes d'estimation sur petits domaines qui font appel à de l'information auxiliaire supplémentaire et à des modèles spécifiques. A défaut, il faut au moins prendre en compte les intervalles de confiance, surtout si l'on souhaite procéder à des comparaisons temporelles ou entre de petites sous-populations.

Conclusion

Au terme des redressements de la non-réponse partielle à l'enquête Logement 2006, on peut dresser plusieurs constats.

Il est important de penser à la prévention des valeurs manquantes et aberrantes dès la conception du questionnaire. La programmation de contrôles de cohérence sous CAPI peut éviter des erreurs de saisie et de compréhension : elle permet non seulement d'améliorer la qualité des données mais aussi de passer moins de temps à repérer les valeurs aberrantes en aval de la collecte. Pour les prêts notamment, de nombreuses erreurs de saisie auraient pu être évitées si davantage de contrôles avaient été programmés. S'agissant des revenus individuels, il aurait été utile de prévoir un contrôle pour vérifier la cohérence entre les tranches renseignées et les montants en clair donnés par certains individus du ménage.

La correction de la non-réponse partielle nécessite beaucoup de temps lorsqu'on cherche à utiliser au maximum les informations présentes dans le questionnaire, surtout quand elles diffèrent de celles qui étaient disponibles dans l'enquête précédente. En prévision de la prochaine enquête Logement qui devrait avoir lieu en 2011, on peut émettre plusieurs recommandations sur le futur questionnaire. Certains redressements seraient facilités si l'on revoyait les questions correspondantes. Par exemple, comme certains ménages sont réticents à décrire leurs placements financiers, il serait utile d'autoriser les « Ne Sait Pas » et « Refus » à la question sur la détention afin de ne pas avoir à remettre en cause les réponses négatives : cela limiterait le risque d'imputer à tort des placements à des ménages réellement non détenteurs. En ce qui concerne les revenus individuels, disposer de tranches au niveau de chaque individu plutôt que du ménage permettrait de simplifier les imputations. Il serait également utile de demander sur quelle durée portent les montants de revenus perçus par les ménages. L'ordre de certaines questions devrait par ailleurs être revu. Pour les allocations RMI, les prestations familiales et les aides à la scolarité, l'emplacement actuel des questions en tranches prête en effet à confusion, non seulement pour les ménages enquêtés mais aussi pour les redressements : elles portent sur le montant perçu par l'ensemble du ménage mais ne sont pas posées dans tous les cas de non-réponse à un montant en clair.

Au final, les méthodes de redressement qui ont été utilisées pour l'enquête Logement 2006 permettent de ne pas modifier artificiellement la distribution des variables concernées. Dans le cas des placements financiers, elles aboutissent à des résultats qui sont cohérents avec ceux de l'enquête Patrimoine. A l'issue des imputations des revenus, on obtient des distributions proches de celles issues d'autres enquêtes auprès des ménages.

En prolongement des redressements, on a comparé les imputations obtenues par la méthode des résidus simulés et par hot-deck aléatoire par classes. Sur des domaines de taille importante, on constate peu de différences selon les méthodes utilisées car les taux de non-réponse sont dans l'ensemble faibles et l'échantillon est de taille importante. Par contre, lors de l'exploitation de l'enquête, il faut être prudent si l'on travaille sur des sous-populations peu nombreuses car des écarts peuvent apparaître selon les méthodes employées, notamment pour des ratios. Si l'on souhaite faire des comparaisons entre des sous-populations ou entre deux enquêtes par exemple, il semble préférable d'utiliser des méthodes d'estimation sur petits domaines, ou du moins de tenir compte des intervalles de confiance. Il est donc important de fournir des indications sur la précision des résultats lors de leur diffusion.

Bibliographie

Correction de la non-réponse

Caron Nathalie (2005) : « La correction de la non-réponse par repondération et par imputation », *document de travail Insee*, n°M0502

Caron Nathalie (1999) : « Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen », *Document de travail Insee*, n°9902

Chopin Nicolas, Masse Emmanuel (2002) : « Imputation de l'enquête Budget de famille 2000 », *Actes des Journées de Méthodologie Statistique*

Contencin Didier, Verger Daniel (1996) : « Les imputations économétriques : un mal nécessaire ? L'exemple de l'enquête Revenus Fiscaux 1992 », *Actes des Journées de Méthodologie Statistique*

Deville Jean-Claude (2005) : « Imputation par prédiction paramétrique et équations estimantes : un essai de mise en cohérence », *Colloque francophone sur les sondages, Québec*

Dupont Françoise : « Imputation Procedures for Quantitative and Qualitative Variables », *document de travail Insee*, n°F9406

Favre Anne-Catherine, Matei Alina et Tillé Yves (février 2005) : « Calibrated random imputation for qualitative data », *Journal of Statistical Planning and Inference, Volume 128, Issue 2*, p.411-425
Ford Barry L. (1983) : « An Overview of Hot Deck Procedures », *Incomplete Data in Sample Surveys*, volume 2, *Theory and bibliographies*, Academic Press

Gautier Éric (2005) : « Éléments sur les mécanismes de sélection dans les enquêtes et sur la non-réponse non-ignorable », *Actes des Journées de Méthodologie Statistique*

Gautier Éric : « La simulation de montants manquants ou en fourchette avec la macro SAS %simul »

Gautier Éric : « Utilisation de la macro SAS %hotdeck »

Gautier Éric, Houdré Cédric (2008) : « Approche multivariée de l'estimation des inégalités dans l'enquête Patrimoine 2004 », *document de travail Insee*, n°F0801

Haziza David (juin 2005) : « Traitement de la non-réponse », *notes de cours FCDA*

Haziza David (2002) : « Inférence en présence d'imputation : un survol », *Actes des Journées de Méthodologie Statistique*

Lagarenne Christine, Lorgnet Jean-Paul (2004) : « Imputation des revenus du patrimoine financier dans l'enquête Revenus Fiscaux », *document de travail Insee*, n°F2004

Rousseau Sylvie, Tardieu Frédéric (2004) : « La macro SAS CUBE d'échantillonnage équilibré : documentation de l'utilisateur », *document de travail Insee*, n°F0402

Vanderschelden Mélanie (2005) : « Homogamie et choix du conjoint : traitement de la non-réponse, imputation de variables qualitatives corrélées », *document de travail Insee*, n°F0505

Estimations de la précision des estimateurs

Caron Nathalie, Ravalet Philippe et Sautory Olivier (1996) : « Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises », *document de travail Insee*, n°F9602

Haziza David, Rancourt Éric (2004) : « Estimation de la variance sous l'approche deux-phases avec modèle d'imputation », *Le bulletin d'imputation*, vol. 4, n°1, 3-8, Statistique Canada

Le Blanc David (1998) : « Utilisation du logiciel Poulpe pour le calcul de la précision d'estimateurs tirés de l'enquête Logement 1996 », *Insee Méthodes*, n°84-85-86

Tillé Yves (2000) : « Théorie des sondages », *notes de cours ENSAI*

- **Correction du biais de sélection**

Heckman James J. (1979) : « Sample Selection Bias as a Specification Error », *Econometrica*, 47, p.153-161

Heckman James J. (1990) : « Varieties of Selection Bias », *American Economic Review, Papers & Proceedings*, 80, p.313-318

- **Modélisation**

Chen Colin (2002) : « Robust Regression and Outlier Detection with the ROBUSTREG Procedure », *SUGI 27*

Confais Josiane, Le Guen Monique (1996) : « La régression linéaire sous SAS », *document de travail Insee*, n°F9605

Derquenne Christian (2003) : « Données atypiques, méthodes robustes et implémentation sous SAS », *présentation au Club SAS Stat*

Jacquot Alain (mars 2000) : « Les modèles économétriques : logit - probit - tobit », *dossier d'étude CNAF*, n°6

Le Blanc David, Lollivier Stéphan, Marpsat Maryse, Verger Daniel (2000) : « L'économétrie et l'étude des comportements - Présentation et mise en œuvre de modèles de régression qualitatifs : les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) », *document de travail Insee*, n°F001.

Lollivier Stéphan (1997) : « Modèles univariés et modèles de durée sur données individuelles », *document de travail Insee*, n°9702

Perretti-Watel Patrick (juin 2000) : « Régression sur variables catégorielles », *notes de cours FCDA*

- **Quelques références sur les variables financières redressées**

Arrondel Luc (1996) : « Patrimoine des ménages : toujours le logement, mais aussi les actifs de précaution », *Économie et Statistique*, n°296-297

Cordier Marie, Rougerie Catherine (septembre 2004) : « Patrimoine des ménages début 2004 - Le déploiement de l'épargne salariale », *Insee Première*, n°985

Debreu Pierre (février 2004) : « Rapport sur la production d'une distribution de référence des revenus des ménages », Direction des Statistiques Démographiques et Sociales

Evain Franck (juillet 2007) : « Le salaire des chefs d'entreprises, moyennes et grandes », *Insee Première*, n°1150

Guillemin Olivier, Goutard Luc (février 2006) : « Guide d'utilisation de l'enquête Revenus fiscaux 2002 rétropolée », Direction des Statistiques Démographiques et Sociales

Kramarz Francis (2003) : « Mobilité et salaires : une longue tradition de recherche », *Économie et Statistique*, n°369-370, p.113-118

Koubi Malik (2003) : « Les carrières salariales par cohorte de 1967 à 2000 », *Économie et Statistique*, n°369-370, p.149-171

Lagarenne Christine, Le Blanc David (1998) : « Redressement des revenus à l'enquête Logement 1996 », *document de travail Insee*, n°F9802

Meurs Dominique, Ponthieux Sophie (2006) : « L'écart des salaires entre les femmes et les hommes peut-il encore baisser ? », *Économie et Statistique*, n°398-399, p.99-129

- **Études réalisées à partir de l'enquête Logement 2002**

Bessière Sabine (janvier 2003) : « La proportion de logements vacants la plus faible depuis trente ans », *Insee Première*, n°880

Daubresse Marion (juillet 2003) : « La reprise de l'accession à la propriété », *Insee Première*, n°913

Driant Jean-Claude, Rieg Christelle (février 2004) : « Les conditions de logement des ménages à bas revenus », *Insee Première*, n°950

Driant Jean-Claude, Rieg Christelle (avril 2004) : « Les ménages à bas revenus et le logement social », *Insee Première*, n°962

Driant Jean-Claude, Castéran Bénédicte et O'Prey Sophie (avril 2008) : « Les conditions de logement des ménages jeunes », *Le rapport de l'Observatoire national de la pauvreté et de l'exclusion sociale 2007-2008*, La documentation française

Jacquot Alain (février 2003) : « De plus en plus de maisons individuelles », *Insee Première*, n°885

Jacquot Alain, Jezequel Blandine et Minodier Christelle (novembre 2004) : « Les charges dans le budget des locataires », *Insee Première*, n°990

Lincot Liliane, Rieg Christelle (octobre 2003) : « Les conditions de logement des ménages en 2002 », *Insee Résultats*, n°20

Minodier Christelle (avril 2004) : « Le parc locatif récent : davantage de maisons et de petits immeubles », *Insee Première*, n°957

Minodier Christelle, Rieg Christelle (septembre 2004) : « Le patrimoine immobilier des retraités », *Insee Première*, n°984

Minodier Christelle (mars 2005) : « Portrait de locataires », *Insee Première*, n°1010

Annexes

Annexe 1 : Les méthodes de prévention de la non-réponse partielle à l'enquête Logement 2006

La prévention de la non-réponse partielle lors de la conception du questionnaire

Lors de la conception du questionnaire, en collaboration avec des spécialistes du logement (ministère de l'équipement, CEREN, ANPEEC…), une attention particulière a été portée à l'ordre des questions : le questionnaire commence par des aspects non financiers et aborde très rapidement la description physique du logement. Pour éviter de décourager le ménage, les premières questions financières sont les plus simples (dépenses d'eau et d'électricité) tandis que la partie sur les revenus, qui se révèle la plus délicate à passer et peut susciter des refus, est posée tout à la fin du questionnaire. Elle se déroule en deux temps afin d'éviter les oublis et de retarder le moment de demander les montants perçus : on recense dans une première partie les différents types de revenus dont ont bénéficié l'ensemble des membres du ménage au cours de l'année écoulée (stratégie de balayage), ce qui permet de lister de manière exhaustive les sources de revenus ; puis on s'intéresse dans une seconde partie aux montants correspondants.

Pour les variables dont le recueil est plus délicat, on cherche à faciliter la tâche de l'enquêteur et du ménage afin de limiter les réponses manquantes et les erreurs de calculs. En ce qui concerne les charges par exemple, même si ce sont au final des montants annuels qui nous intéressent, on essaye de se rapprocher le plus possible des modalités concrètes de paiement : on demande la périodicité des versements et, si le montant est variable selon les échéances, les différents montants payés pour chacune (CAPI permet de filtrer les questions d'une manière « transparente » pour l'enquêteur). L'expérience des enquêtes précédentes a également conduit à alléger certaines parties comme celle sur les prêts.

Il est par ailleurs important de bien savoir ce que l'on mesure. Par exemple, on souhaite si possible collecter un loyer « pur » hors charges et taxes locatives afin de disposer de données homogènes et donc comparables. Cependant, un certain nombre de locataires payent un montant global sans pouvoir distinguer les montants des loyers, des charges et des taxes. Pour éviter de trop nombreuses non-réponses, ce cas a été prévu : l'enquêteur doit reporter le montant donné le ménage et on lui demande ensuite de préciser si ce montant inclut ou non des charges (et si oui, de quelles charges il s'agit). Cela permet d'estimer un loyer hors charges et taxes lors de l'exploitation.

Enfin, les remarques formulées par les enquêteurs ont permis d'améliorer plusieurs aspects du questionnaire et notamment de reformuler certaines questions jugées peu claires. La possibilité de répondre en Francs ou en Euros a été prévue afin de limiter des erreurs de conversion et des séparateurs de milliers ont ajoutés sous CAPI pour éviter les erreurs de saisie dans les montants très élevés. Enfin, les points les plus délicats des instructions de collecte ont été ajoutés en commentaire des questions correspondantes sous CAPI. Les tests ont également permis d'évaluer la durée moyenne de passation du questionnaire, que l'on a essayé de limiter à 50 minutes en moyenne pour limiter les abandons en cours d'entretien.

Le rôle primordial des enquêteurs pour la qualité des données et la prévention de la non-réponse

L'enquête se déroulant en face à face, les enquêteurs participent activement à la prévention de la non-réponse durant l'entretien, d'où l'importance des formations qui leur sont dispensées et des instructions de collecte qui leur sont remises. Elles visent à leur préciser la nature de l'information demandée, pour les questions techniques ou d'ordre réglementaire notamment. On leur suggère aussi d'inciter les enquêtés à se référer à des documents (quittances de loyers, échéanciers de remboursement de prêt, factures d'eau…), ce qui permet de gagner en temps et en qualité des réponses ; avant l'entretien, les enquêteurs envoyaient d'ailleurs une lettre-avis informant le ménage du thème de l'enquête et l'invitant à préparer des documents financiers.

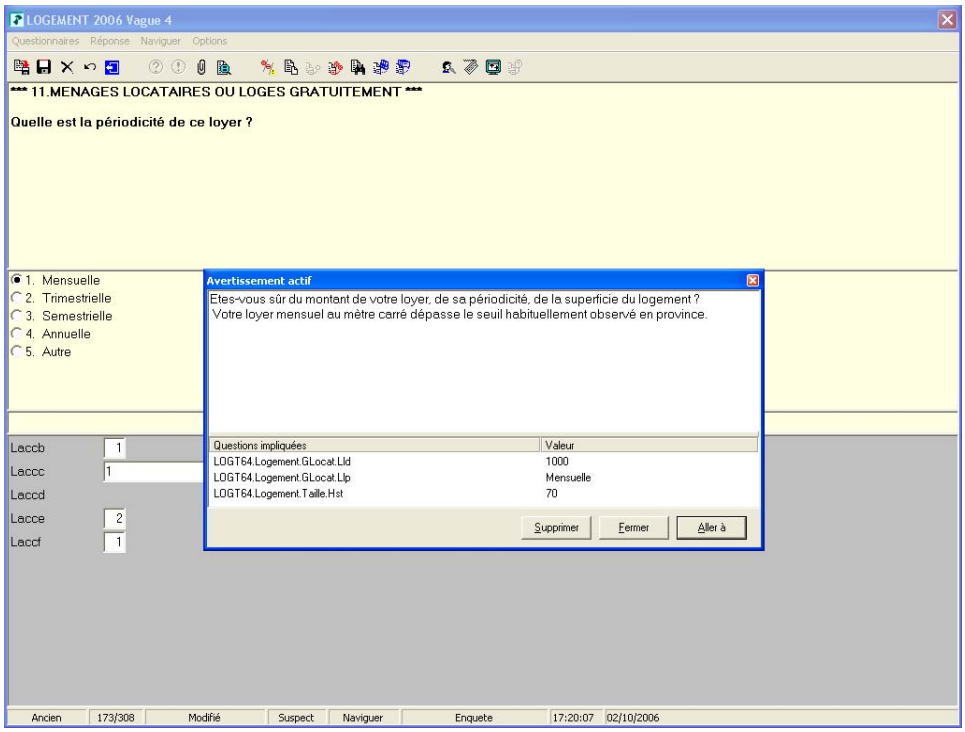
Les enquêteurs devaient interroger de préférence la personne de référence du ménage, qui est souvent le « gestionnaire financier » et donc la plus à même de répondre au questionnaire. Pour faciliter l'entretien, on leur recommandait d'éviter de la faire répondre devant d'autres personnes et on leur proposait dans l'instruction de collecte des arguments sur la confidentialité des données afin de vaincre

d'éventuelles réticences à répondre aux questions sensibles comme celles sur les revenus. On leur fournit également des éléments pour justifier la présence de questions qui ne semblent pas directement liées au thème de l'enquête : par exemple, des questions sont posées sur les revenus même si l'enquête est consacrée au logement car on souhaite déterminer la part de budget consacrée aux dépenses de logement. Enfin, on conseille aux enquêteurs d'indiquer de préférence un montant même imprécis plutôt qu'un « Ne Sait Pas » car le montant imputé en aval risque d'être plus éloigné de la vraie valeur qu'une réponse approximative fournie de mémoire par le ménage.

La programmation de contrôles sous CAPI

On demande par ailleurs aux enquêteurs d'être attentifs à la cohérence des données financières entre elles. Cet examen est facilité par des contrôles programmés sous CAPI qui signalent durant l'entretien de possibles valeurs aberrantes. La formulation du message est adaptée au cas rencontré et donne des indications sur les réponses à corriger éventuellement (le contrôle permet d'accéder directement aux questions susceptibles d'être modifiées). S'il ne s'agit pas d'une erreur mais d'un cas particulier, l'enquêteur peut noter sous CAPI des éléments d'explication : par exemple, un montant de loyer peut être anormalement faible s'il y a des liens de parenté entre le propriétaire et le locataire. Ces remarques sont ensuite prises en compte lors des traitements post-collecte réalisés par les gestionnaires en DR ou par l'équipe conceptrice ; elles facilitent le repérage des valeurs aberrantes et atypiques.

Certains contrôles sont bloquants : ils visent à signaler des incohérences graves (par exemple, on ne peut pas s'installer dans un logement avant sa construction) et l'enquêteur doit modifier sa réponse pour pouvoir continuer. En revanche, les contrôles non-bloquants visent à attirer l'attention de l'enquêteur sur des situations anormales mais pas nécessairement fausses (par exemple, certains chômeurs ne perçoivent pas d'indemnités de chômage) : s'il n'y a pas d'erreur, le contrôle doit alors simplement être confirmé par l'enquêteur. Exemple de message non bloquant apparaissant pour un ménage en province qui occupe un logement de 70 m^2 et déclare un loyer mensuel de 1 000 € :



C'est naturellement la partie sur les revenus qui comporte le plus de contrôles, à la fois sur les types de revenus perçus (par exemple, on s'attend à ce qu'un actif salarié ait bien perçu un salaire au cours des 12 derniers mois) et sur les montants déclarés (on demande à l'enquêteur une confirmation si le salaire ne semble pas en rapport avec le type de profession exercée).

Enfin, pour éviter des pertes d'information liées aux traitements informatiques, une « répétition générale » a permis de tester la remontée des informations collectées sous CAPI vers des tables SAS.

Annexe 2 : Les taux de non-réponse partielle à l'enquête Logement 2006

Les taux de non-réponse partielle à l'enquête Logement 2006 sont les suivants :

Revenus perçus au niveau individuel :

- salaires et primes (8,4 % de non-réponses)
- indemnités de chômage (7,6 %)
- retraites et pensions (11,7 %)
- revenus non salariaux (24,2 %)

Revenus perçus au niveau du ménage :

- aides à la scolarité (9,6 %)
- prestations familiales et liées au handicap (7,0 %)
- revenus fonciers (23,9 %)
- allocations RMI perçues par la personne de référence et/ou son conjoint (4,6 %)
- allocations RMI perçues par une autre personne du ménage (9,5 %)
- revenus de type « autre » (5,3 %)
- tranche du montant des placements financiers (11,2 %)
- tranche du montant du patrimoine (11,9 %)

Variables relatives aux ménages locataires et logés gratuitement :

- charges locatives (18,9 %)
- loyers (0,5 %)
- aides au logement (1,8 %)

Variables relatives aux propriétaires :

- prix du logement (8,1 %)
- variables relatives aux prêts : taux d'intérêt (4,7 %), montant emprunté (2,8 %), montant de la dernière mensualité (1,7 %)
- aides au logement (4,4 %)
- charges de copropriété (2,9 %)

Dépenses en eau et en énergies :

- dépenses en eau (13,2 %)
- dépenses en énergies (16,6 %)

Annexe 3 : Modélisation par la régression robuste et la PROC LIFEREG

La régression robuste

La régression robuste permet d'analyser des données contaminées par des erreurs et des valeurs atypiques. Alors que les coefficients estimés par les MCO peuvent être biaisés en présence d'observations atypiques, la régression robuste fournit des estimations stables, c'est-à-dire peu sensibles aux valeurs aberrantes, car elle atténue l'influence des individus extrêmes (ce sont ceux pour lesquels les résidus sont élevés). Elle repère aussi les points aberrants.
La PROC ROBUSTREG, disponible sous SAS v9, permet d'effectuer les techniques les plus couramment utilisées de régressions robustes[55]. Sa syntaxe est la suivante (on utilise ici la méthode d'estimation M[56]) :

```
proc robustreg data = tab METHOD=m
OUTEST=est /*table contenant les estimations des paramètres*/;
MODEL salaire_log = &varexpl / diagnostics leverage;
OUTPUT OUT = res PREDICTED = xb RESIDUAL = residus
OUTLIER=atyp /*les valeurs atypiques sont celles pour lesquelles atyp=1/;
run;
```

La PROC LIFEREG

La PROC LIFEREG permet d'ajuster des modèles sur variables en tranches, et notamment des modèles pour lesquels la variable d'intérêt est observée (au moins en partie) sous forme d'un système de tranches[57]. L'idée est de postuler l'existence d'une variable latente y* non observable à laquelle on applique un modèle linéaire (modèle à variable latente discrétisée). Les paramètres sont estimés par la méthode du maximum de vraisemblance.

Soit le modèle $\ln(Y) = X'\beta + \sigma U$, où X est la matrice de variables auxiliaires, U une variable aléatoire qui suit une loi centrée réduite (de densité f et de fonction de répartition F), et σ l'écart-type des résidus (σU vecteur des résidus). Soit E l'ensemble des individus i qui ont fourni un montant exact y_i, et T l'ensemble des individus i qui ont donné une tranche (correspondant à un intervalle $[b_i,h_i]$). La vraisemblance du modèle s'écrit alors :

$$L(\beta,\sigma) = \prod_{i\in E} f\left(\frac{\ln(y_i) - x'_i\,\beta}{\sigma}\right) \prod_{i\in T}\left[F\left(\frac{\ln(h_i) - x'_i\,\beta}{\sigma}\right) - F\left(\frac{\ln(b_i) - x'_i\,\beta}{\sigma}\right)\right]$$

La syntaxe de la PROC LIFEREG est la suivante (la loi du modèle est ici la loi log-normale) :

```
proc lifereg data = tab
OUTEST=est /*table contenant les estimations des paramètres*/;
MODEL (sal_min, sal_max) = &varexpl / D=LNORMAL;
/*on définit pour chaque observation une limite basse et une limite haute de
tranche*/
OUTPUT OUT=res XBETA=xb;
run;
```

Notons que pour la tranche la plus élevée, il n'y a pas de valeur maximale (seule la borne inférieure de la tranche est indiquée). Si l'individu a donné une réponse en clair, les sal_min et sal_max prennent cette valeur.
La PROC LIFEREG fournit par ailleurs l'écart-type des résidus (variable SCALE).

[55] On pourrait aussi utiliser le CALL LMS, qui consiste à minimiser la médiane des carrés des résidus estimés.
[56] Alors que l'estimateur des MCO repose sur la minimisation de la somme des carrés des résidus, le M-estimateur est obtenu en minimisant la somme de fonctions des résidus qui augmentent moins rapidement et sont donc plus robustes (la fonction des résidus doit être choisie de manière à limiter l'influence des valeurs atypiques). La résolution se fait par itérations de l'algorithme IRLS (« iteratively reweighted least-squares ») ; l'idée est d'estimer les coefficients avec une régression par les moindres carrés dans laquelle on donne aux résidus faibles un poids plus fort qu'aux résidus élevés.
[57] Voir Lollivier (1997) pour plus de détails.

Annexe 4 : Hétéroscédasticité : tests de White et de Breusch-Pagan

L'idée de ces tests est de regarder si le carré des résidus peut être expliqué par les variables explicatives du modèle (si c'est le cas, il y a hétéroscédasticité). L'hypothèse nulle H_0 est celle de la constance de la variance des résidus dans le modèle.

Test de White :

Le test de White permet de tester l'hypothèse d'homoscédasticité. Il consiste à régresser les carrés des résidus sur les variables explicatives, leurs carrés et tous leurs croisements, puis à faire le test du chi-deux de la nullité globale des coefficients de la régression en calculant le nR^2 de la régression (n étant le nombre d'observations).

On peut considérer qu'on est en présence d'hétéroscédasticité si la probabilité de ne pas rejeter H_0 est inférieure à 10 %. Toutefois, le test est délicat à interpréter quand il y a beaucoup d'observations car il conduit trop souvent à rejeter l'hypothèse nulle (si n est grand, il faut en effet un R^2 très petit pour accepter l'hypothèse d'homoscédasticité).

Le test de White peut être mis en œuvre avec la PROC MODEL, donc la syntaxe est la suivante :
soit le modèle y = a + b*x + c*z + e :

```
PROC MODEL data = tab ;
PARMS a b c ;
y = a + b*x + c*z ;
FIT y / WHITE ;
QUIT ;
```

Test de Breusch-Pagan :

Le test de Breusch-Pagan peut être utilisé pour identifier les variables les plus en cause dans l'hétéroscédasticité. Il consiste à régresser le carré des résidus sur des variables explicatives choisies.

Sous SAS, sa syntaxe est la suivante dans le cas où on cherche à savoir si la variable z est en cause dans l'hétéroscédasticité :

```
PROC MODEL data = tab;
PARMS a b c ;
y = a + b*x + c*z ;
FIT y / BREUSCH = (z) ;
QUIT ;
```

En pratique, on peut faire des tests de Breusch-Pagan sur toutes les variables, puis retenir les deux ou trois variables qui causent le plus d'hétéroscédasticité et regarder la variance des résidus selon les valeurs prises par ces variables.

Annexe 5 : La macro %simul

La macro %simul[58] permet de mettre en œuvre la deuxième partie de la méthode des résidus simulés, c'est-à-dire la simulation ou le tirage des résidus et le calcul des valeurs imputées, tout en tenant compte de l'information en tranches éventuellement disponible pour la variable d'intérêt.

Les méthodes de simulation ou de tirage des résidus

Trois méthodes sont proposées pour simuler ou tirer les résidus.

- **Simulation des résidus dans la loi normale : méthodes 1 ou 2**

Les deux premières méthodes permettent de simuler les résidus dans la loi normale.

On utilise généralement la **méthode 1** (méthode classique de simulation), qui est la plus simple et la plus rapide. Elle consiste à tirer des résidus gaussiens dans la loi conditionnelle à l'information en tranche par inversion de la fonction de répartition.

Soit le modèle $\ln Y = X'\beta + \sigma U$, avec X matrice de variables auxiliaires, U une variable gaussienne centrée réduite et σ l'écart-type des résidus (σU vecteur des résidus). Lorsqu'on dispose d'information en tranche, on tire les résidus dans une loi normale tronquée, c'est-à-dire restreinte à l'intervalle correspondant à la tranche déclarée. Pour l'individu i, si la tranche observée correspond à un intervalle $[b_i,h_i]$, on simule U dans la loi N(0,1) tronquée où l'intervalle de troncature est :

$$\left[\frac{\ln(b_i) - x'_i\,\hat{\beta}}{\hat{\sigma}}, \frac{\ln(h_i) - x'_i\,\hat{\beta}}{\hat{\sigma}}\right].$$

Pour calculer U, on simule une variable V dans une loi uniforme sur [0,1] et on construit :

$$U = \Phi^{-1}\left\{(1-V)\Phi\left(\frac{\ln(b_i) - x'_i\,\hat{\beta}}{\hat{\sigma}}\right) + V\Phi\left(\frac{\ln(h_i) - x'_i\,\hat{\beta}}{\hat{\sigma}}\right)\right\}$$

Dans de rares cas, SAS peut ne pas pouvoir calculer U pour des raisons numériques (problèmes d'arrondis liés au fait que Φ et Φ^{-1} sont obtenus numériquement). On utilise alors la **méthode 2**, qui est une méthode d'acceptation-rejet : on tire plusieurs réalisations dans la loi normale N(0,1)[59] (non conditionnelle, donc plus simple à simuler que la loi conditionnelle à l'information en tranches) et on regarde si le résidu est tel que l'imputation se situe dans la tranche. Notons que la méthode 2 n'est pas forcément très longue à mettre en œuvre car les critères d'acceptation sont optimisés.

- **Tirage des résidus dans la loi empirique des résidus observés : méthode 3**

On utilise cette méthode si les résidus s'écartent beaucoup de la loi normale. Elle s'apparente à du hot-deck : on construit une table des résidus observés puis on tire pour chaque observation à imputer un résidu dans cette table. Il s'agit, comme la méthode 2, d'une méthode d'acceptation-rejet : le résidu imputé est le premier résidu qui permette au montant imputé de se situer dans la tranche éventuellement déclarée.
En cas d'hétéroscédasticité des résidus observés, il est possible de faire du tirage stratifié des résidus dans la loi empirique.

[58] Elle remplace la macro %impute de Stéfan Lollivier. Pour plus de détails, voir Gautier : « La simulation de montants manquants ou en fourchette avec la macro SAS %simul »
[59] Le programme tire en fait dans la loi exponentielle plutôt que dans la loi normale pour des raisons pratiques.

Utilisation de la macro %simul :

- En pratique, il faut préparer en entrée de %simul une table comprenant :

- une variable XB contenant les prédictions calculées lors de la première étape de la méthode des résidus simulés.
Si on a utilisé une PROC LIFEREG pour estimer le modèle, on les a directement grâce au paramètre XBETA. Par contre, avec une PROC ROBUSTREG, il faut les calculer (par une PROC SCORE par exemple).

- une variable SIGMA contenant l'écart-type des résidus observés sur les répondants.
Si on a estimé l'équation avec une PROC LIFEREG, l'écart-type est donné par le paramètre _SCALE_[60]. Si on a utilisé la PROC ROBUSTREG, on n'a pas directement l'écart-type des résidus mais un paramètre d'échelle qui n'est pas exactement l'écart-type ; pour l'obtenir, il faut effectuer une modélisation supplémentaire par les MCO en retirant les valeurs atypiques du modèle.
En cas d'hétéroscédasticité des résidus observés sur les répondants, on doit stratifier par l'intermédiaire de SIGMA, en calculant les écarts-types des résidus dans chaque strate.

- une variable DET qui vaut 1 si l'individu est censé percevoir un montant non nul pour la variable d'intérêt, 0 sinon.

- des variables indiquant les limites des tranches éventuellement renseignées (si la variable à imputer est SALAIRE, elles doivent s'intituler SALAIREMIN et SALAIREMAX). Ces variables sont naturellement à valeur manquante si l'individu n'a pas donné de montant exact ni de tranche.

- On applique ensuite la macro, dont la syntaxe est la suivante[61] :

```
%simul(variable=salaire,tableinput=entree,tableoutput=imp,nbsimul=1,
nbiter=5,pred=0,methode=1,ppv=0);
```

On indique à VARIABLE la variable à simuler, à TABLEINPUT la table en entrée décrite ci-dessus et à TABLEOUTPUT la table de sortie contenant les imputations (la variable imputée est suffixée par IMP1). On peut préciser à NBSIMUL le nombre de simulations (dans le cas où l'on souhaite faire de l'imputation multiple) et à NBITER le nombre d'itérations de l'algorithme d'acceptation-rejet (pour les méthodes 2 et 3).

La macro %simul permet aussi de faire du hot-deck après la simulation des résidus (avec le paramètre PPV=1) : elle cherche des valeurs observées proches de ces valeurs imputées et produit les plus proches voisins. Cela peut être intéressant s'il y a beaucoup de valeurs basses ou au contraire très élevées.

La macro permet aussi de faire également de l'imputation par la moyenne (PRED=1).

[60] Pour l'obtenir, la syntaxe est la suivante : `data entree; set est end=fin; if _n_=1 then set est (keep=_scale_); rename _scale_=sigma; run;` (la table EST est obtenue en sortie de la PROC LIFEREG grâce à l'instruction OUTEST).
[61] Elle ne permet pas d'utiliser les poids mais si la corrélation entre les variables utilisées dans le plan de sondage et la variable d'intérêt n'est pas trop grande, l'imputation non-pondérée mène à des estimateurs approximativement sans biais.

Annexe 6 : La macro %hotdeck

La macro %hotdeck permet de faire du hot-deck aléatoire par classes tout en tenant compte d'information en tranches : elle procède à des simulations dans la loi empirique conditionnellement à l'appartenance à une tranche, en mettant en œuvre un algorithme d'acceptation-rejet[62].

Sa syntaxe est la suivante :

```
%hotdeck(variable=acti,varstrat=cat cs statut2,tableinput=ent,
tableoutput=res,nbsimul=1,nbiter=50,pred=0,tranche=0,nbdon=1);
```

Le paramètre VARIABLE permet d'indiquer la variable à simuler et TABLEINPUT la table en entrée (dans laquelle il faut avoir ajouté des variables de limites des tranches éventuellement renseignés). TABLEOUTPUT correspond à la table de sortie, qui contient les imputations (la variable imputée est suffixée par IMP1). Si on utilise des classes d'imputation, les variables utilisées pour les constituer sont précisées à VARSTRAT. Le paramètre TRANCHE permet d'indiquer si la variable d'intérêt peut être ou non renseignée en tranches (il vaut 1 si c'est le cas, 0 sinon). On indique à NBSIMUL le nombre de simulations (la macro peut faire de l'imputation multiple) et à NBITER le nombre d'itérations de l'algorithme d'acceptation-rejet. Enfin, si l'on fixe le paramètre NBDON à 1, la macro permet de savoir *a posteriori* combien de fois chaque donneur a été utilisé.

Pour que la macro puisse imputer une valeur manquante, il faut qu'au moins un donneur se situe dans la même classe que le non-répondant. En revanche, dans le cas où un non-répondant a indiqué une tranche mais où aucun des répondants de la classe n'a indiqué de valeur se situant dans cette tranche, la macro impute le milieu de tranche ou, si une seule des bornes est connue, laisse la valeur manquante. On peut alors faire tourner un second hot-deck simplifié pour imputer ces observations (en retenant moins de variables et de modalités que dans le premier hot-deck), en retirant des donneurs de ce nouveau hot-deck les donneurs et receveurs du hot-deck précédent.

La macro %hotdeck permet également d'effectuer de la prédiction par la moyenne par classes (si le paramètre PRED vaut 1). En revanche, elle ne fait pas de hot-deck métrique ni de hot-deck séquentiel.

[62] Pour plus de détails, voir Gautier : « Utilisation de la macro SAS %hotdeck ».

Annexe 7 : Le redressement par étapes de variables liées entre elles : les variables relatives aux prêts immobiliers

En 2006, les questions sur les prêts immobiliers ont été modifiées par rapport aux enquêtes Logement précédentes, ce qui nécessite de revoir la manière de les redresser. A cette occasion, on essaie de prendre en compte le maximum d'information auxiliaire disponible.

Un certain nombre de questions ont en effet été abandonnées pour les prêts renégociés ou remboursés par anticipation, ainsi que pour les prêts souscrits par des accédants anciens[63]. Alors que les ménages devaient auparavant décrire l'ensemble des caractéristiques initiales puis actuelles (en cas de renégociation) de chacun de leurs prêts, on ne cherche plus en 2006 à reconstituer l'historique : on demande d'une part la description des prêts contractés au moment de l'achat (montant emprunté, année de départ, durée et taux d'intérêt) dans une optique « plan de financement », et d'autre part les remboursements actuels dans une optique « taux d'effort », ce qui réduit la quantité d'informations disponibles pour les imputations et nécessite de revoir la procédure de redressements utilisée. A cette occasion, on essaie de prendre en compte davantage d'information auxiliaire.

Trois variables doivent être redressées : le taux d'intérêt, le montant du prêt et le montant du dernier remboursement. Elles sont naturellement liées entre elles mais comme la non-réponse n'était pas autorisée pour les autres caractéristiques des prêts, on dispose d'éléments toujours renseignés : les ménages étaient en effet obligés de d'indiquer l'année de départ du prêt et sa durée, éventuellement approximativement. De plus, il y a peu de prêts pour lesquels les trois variables sont manquantes simultanément donc on abandonne l'idée initialement envisagée de les imputer en même temps. On met donc en œuvre un nouveau redressement par étapes.

Bilan de la non-réponse :

Accédants récents	Nombre de prêts décrits	Effectif renseigné	Nombre de non-réponses	% de non-réponses
Montant du remboursement mensuel*	5 812	5 684	128	2,2
Montant du taux d'intérêt**	5 932	5 751	181	3,1
Montant emprunté**	5 932	5 802	130	2,2

* y compris les prêts de périodicité « autre » que l'on peut convertir à une périodicité mensuelle
** y compris les prêts dont le remboursement mensuel est nul (c'est-à-dire les prêts totalement remboursés)

Accédants anciens	Nombre de prêts décrits	Effectif renseigné	Nombre de non-réponses	% de non-réponses
Montant du remboursement mensuel	7 910	7 800	110	1,4
Montant du taux d'intérêt	7 910	7 436	474	6,0
Montant emprunté	7 910	7 651	259	3,3

Pour tenir compte des liens qui relient les variables relatives aux prêts, on procède à un redressement en trois étapes :

i) imputation des taux d'intérêt : on utilise les méthodes d'imputation par la moyenne par classes et par le plus proche voisin. Pour les prêts « classiques »[64] des accédants récents et pour les prêts des accédants anciens, on peut alors déduire la mensualité à partir du montant emprunté (et, réciproquement, le montant emprunté à partir de la mensualité)

ii) imputation des montants empruntés restant à imputer : par la méthode des résidus simulés

iii) imputation de la mensualité : on la déduit à partir des taux d'intérêt et des montants totaux empruntés imputés à l'étape précédente.

[63] Les accédants anciens sont les ménages qui ont acheté leur logement quatre ans au moins avant la date de collecte. En 2006, on ne demande plus si leurs prêts ont fait l'objet d'une renégociation ou d'un remboursement anticipé : à défaut d'informations, on fait pour les imputations l'hypothèse simplificatrice que cela n'a pas été le cas.
[64] On entend par prêts « classiques » les prêts non renégociés, non remboursés par anticipation et dont les remboursements sont constants.

I. PREMIÈRE ÉTAPE : IMPUTATION DES TAUX D'INTÉRÊT ET DÉDUCTION DE LA MENSUALITÉ OU DU MONTANT TOTAL EMPRUNTÉ

1.1. Imputation des taux d'intérêt par la moyenne par classes ou par le plus proche voisin

On commence par imputer les taux d'intérêt non renseignés par une méthode déterministe (par la moyenne des taux des prêts de même type ou, à défaut, par le plus proche voisin). On raisonne séparément pour les accédants anciens et récents et, parmi les accédants récents, selon que les remboursements sont constants ou variables.

On impute les taux manquants par le taux moyen observé pour les prêts de même nature, c'est-à-dire les prêts de même type, souscrits la même année et de même durée. Si l'on ne trouve pas de prêt de même nature, on impute par le taux du prêt de même type dont la durée (ou à défaut l'année de souscription) est la plus proche. Au final, l'imputation ne modifie pas la distribution des taux d'intérêt :

	Avant	Après
Nombre de prêts	8 321 371	8 321 371
Effectif renseigné	7 923 114	8 321 371
Effectif non renseigné	398 257	-
% de non-réponses	4,8	-
Moyenne	4,26	4,31
Intervalle de confiance*	[4,22 ; 4,30]	[4,27 ; 4,35]
Écart-type	2,37	2,35
Minimum	0,00	0,00
1er quartile	3,16	3,25
Médiane	4,30	4,37
3ème quartile	5,54	5,55
Maximum	19,00	19,00

* Calculé sous l'hypothèse d'un sondage aléatoire simple

1.2. Pour les prêts récents « classiques » et les prêts anciens, déduction de la dernière mensualité ou du montant emprunté

Grâce aux taux d'intérêt imputés précédemment, on dispose à présent de toutes les informations nécessaires pour calculer le montant de la dernière mensualité lorsque l'on connaît le montant total emprunté (et réciproquement), pour les prêts « classiques » des accédants récents ainsi que pour les prêts des acquéreurs anciens.

a) Déduction de la mensualité si le montant total emprunté est connu

Pour les prêts qui ne sont pas de taux nul ni de durée indéterminée (et qui n'ont pas été renégociés ni remboursés par anticipation), on peut déduire la mensualité du montant emprunté grâce à la formule suivante :

$$12 \times PMR = \frac{TILD \times PMP}{1-(1+\frac{PTI}{100})^{-PDU}}$$

où PDU est la durée du prêt, PMP le montant total emprunté, PMR le montant de la mensualité, PTI le taux d'intérêt (fixe) et $TILD = (1+\frac{PTI}{100})^{\frac{1}{12}} - 1$ (taux d'intérêt de longue durée).

Pour les prêts de taux nuls, on calcule le montant remboursé comme le rapport du capital emprunté sur la durée du prêt (il s'agit de prêts à taux zéro sans différé d'amortissement).

b) Déduction du montant total emprunté si la mensualité est connue

Pour les prêts qui ne sont pas de taux nul ni de durée indéterminée, on déduit le montant emprunté de la dernière mensualité à partir de la formule précédente.

Pour les prêts à taux zéro, on multiplie la dernière mensualité par la durée du prêt pour obtenir le montant total emprunté (les prêts concernés par le redressement n'ont pas de différé d'amortissement).

II. DEUXIÈME ÉTAPE : IMPUTATION DU MONTANT EMPRUNTÉ PAR LA MÉTHODE DES RÉSIDUS SIMULÉS

La deuxième étape consiste à imputer, par la méthode des résidus simulés, les montants empruntés non renseignés. On modélise séparément les accédants anciens et récents car on dispose de variables auxiliaires différentes (le prix du logement n'était pas demandé aux acquéreurs anciens). A l'occasion de l'enquête 2006, on inclut davantage d'information auxiliaire que lors des enquêtes précédentes.

2.1. Imputation du montant total emprunté pour les prêts des accédants récents

Pour les accédants récents, on régresse le logarithme du montant du prêt sur les variables explicatives suivantes : année de départ du prêt, durée, type de prêt, taux d'intérêt, niveau de vie, prix du logement (déclaré ou imputé), montant de l'apport personnel[65], nombre de prêts, taille d'unité urbaine, nature de l'apport personnel. Ce modèle a un pouvoir explicatif satisfaisant : $R^2 = 0,71$.

On effectue ensuite une PROC LIFEREG en utilisant le prix du logement (déclaré ou imputé) comme borne maximum du montant du prêt, puis on applique la macro %SIMUL.

Au final, la distribution des montants empruntés par les accédants récents est peu modifiée :

	Avant	Après
Nombre de prêts	3 299 182	3 299 182
Effectif renseigné	3 226 293	3 299 182
Effectif non renseigné	72 889	-
% de non-réponses	2,2	-
Moyenne	73 226,5	73 259,4
Intervalle de confiance*	[71 391 ; 75 063]	[71 448 ; 75 072]
Écart-type	70 805,5	70 731,9
Minimum	150	97
1er quartile	15 785	16 000
Médiane	60 000	60 000
3ème quartile	110 000	110 000
Maximum	870 000	870 000

* Calculé sous l'hypothèse d'un sondage aléatoire simple

2.2. Imputation du montant total emprunté pour les prêts des accédants anciens

Pour les accédants anciens, on régresse le logarithme du montant du prêt sur les variables explicatives suivantes : année de départ du prêt, durée, type de prêt, taux d'intérêt, niveau de vie, taille d'unité urbaine, catégorie sociale de la personne de référence du ménage, type d'habitat (individuel ou collectif), surface du logement, nombre de prêts en cours de remboursement. Le coefficient de détermination est satisfaisant : $R^2 = 0,59$. On applique ensuite la macro %SIMUL.

L'imputation modifie peu la distribution des montants empruntés par les accédants anciens :

	Avant	Après
Nombre de prêts	5 022 189	5 022 189
Effectif renseigné	4 881 766	7 910
Effectif non renseigné	140 423	-
% de non-réponses	2,8	-
Moyenne	52 811,3	53 107,4
Intervalle de confiance*	[51 278 ; 54 344]	[51 599 ; 54 616]
Écart-type	65 781,3	66 447,6
Minimum	125	125
1er quartile	15 000	15 244
Médiane	39 000	39 332
3ème quartile	71 650	71 651
Maximum	1 000 000	1 000 000

* Calculé sous l'hypothèse d'un sondage aléatoire simple

[65] Vente d'un ou plusieurs logements (hors prêt-relais), vente de produits financiers, dons de particuliers et/ou héritage, épargne courante du ménage.

III. TROISIÈME ÉTAPE : DEDUCTION DU MONTANT DE LA MENSUALITÉ GRÂCE AU MONTANT TOTAL EMPRUNTÉ IMPUTÉ PRÉCÉDEMMENT

Enfin, on déduit le montant de la mensualité à partir du montant emprunté imputé à l'étape précédente en utilisant la même formule que précédemment, y compris pour les prêts renégociés ou remboursés par anticipation (seuls cinq prêts de ce type sont concernés[66]).
Les prêts à taux zéro nécessitent un traitement différent selon qu'il y a ou non un différé d'amortissement. Si le début des remboursements est postérieur à l'enquête, la mensualité est nulle. Par contre, en l'absence de différé d'amortissement, on calcule le montant remboursé comme le rapport du capital emprunté sur la durée du prêt. Enfin, les prêts familiaux à taux zéro restant à imputer ont une durée indéterminée : on considère que le ménage n'a rien remboursé au cours de l'année car c'est la situation la plus fréquente parmi les répondants ayant souscrit des prêts de mêmes caractéristiques.

La distribution des mensualités est très peu modifiée par l'imputation :

	Avant	**Après**
Nombre de prêts	8 303 936	8 303 936
Effectif renseigné	8 124 530	8 303 936
Effectif non renseigné	179 406	-
% de non-réponses	2,2	-
Moyenne	435,2	433,7
Intervalle de confiance*	[428,2 ; 442,2]	[426,6 ; 440,8]
Écart-type	422,2	425,0
Minimum	0	0
1er quartile	106	104
Médiane	363	360
3ème quartile	617	616
Maximum	8 000	8 000

* Calculé sous l'hypothèse d'un sondage aléatoire simple

[66] On a regardé ce que l'utilisation de la formule donnait pour les répondants ayant souscrit des prêts comparables : il n'y a en moyenne que 100 € de différence entre la mensualité donnée par le ménage et celle déduite par la formule.

Annexe 8 : Détail des redressements des placements financiers

Nombre d'utilisations des donneurs dans les imputations par hot-deck aléatoire par classes :

- pour les ménages ayant déclaré détenir des placements financiers mais n'ayant pas donné le montant associé :

Nombre d'utilisations	Fréquence	%
1	2 617	92,7
2	197	7,0
3	8	0,3
4	1	0,0
Moyenne	1,1	

- pour les ménages auxquels on a affecté la détention de placements : imputation simultanée de types de placements et de tranches :

Nombre d'utilisations	Fréquence	%
1	4 280	84,3
2	685	13,5
3	101	2,0
4	8	0,2
5	3	0,1
Moyenne	1,2	

Annexe 9 : Détail des redressements des charges locatives

Redressement des charges locatives dans l'habitat collectif

Résultats de la régression par les MCO :

The GLM Procedure

Number of Observations Read 14031
Number of Observations Used 11120

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	40	3725.200445	93.130011	323.77	<.0001
Error	11079	3186.807060	0.287644		
Corrected Total	11119	6912.007506			

R-carré	Coeff Var	Racine MSE	log_cmm1 Moyenne
0.538946	12.60322	0.536324	4.255455

Paramètre	Estimation	Erreur standard	Valeur du test t	Pr > |t|
Intercept	3.576406698	0.03059425	116.90	<.0001
eaufroide	0.316391075	0.01335534	23.69	<.0001
eauchaude	0.094308205	0.01456416	6.48	<.0001
chauffcoll	0.494874440	0.01504063	32.90	<.0001
ascenseur	0.125425300	0.01577845	7.95	<.0001
gardien	0.101626902	0.01221886	8.32	<.0001
autreschar	0.186091488	0.01594229	11.67	<.0001
surf_29	-0.152153482	0.03124294	-4.87	<.0001
surf30_54	-0.060081492	0.01842342	-3.26	0.0011
surf75_94	0.048859053	0.01657183	2.95	0.0032
surf95_	0.162832754	0.02415185	6.74	<.0001
tu_m50000	-0.100789194	0.01864775	-5.40	<.0001
tu_m200000	-0.036764876	0.01548802	-2.37	0.0176
tu_paris	0.237722370	0.01405510	16.91	<.0001
dec1	0.008864501	0.02064122	0.43	0.6676
dec2	-0.031009672	0.02065746	-1.50	0.1333
dec3	-0.033643059	0.02148737	-1.57	0.1174
dec4	0.006462017	0.02206316	0.29	0.7696
dec6	-0.003374375	0.02331631	-0.14	0.8849
dec7	-0.022961067	0.02445997	-0.94	0.3479
dec8	0.033735436	0.02544542	1.33	0.1849
dec9	0.073642999	0.02422714	3.04	0.0024
iaa_48	-0.201563321	0.01705880	-11.82	<.0001
iaa75_81	0.025597111	0.01825177	1.40	0.1608
iaa82_89	0.075384881	0.02259332	3.34	0.0009
iaa90_98	0.037738923	0.02276173	1.66	0.0973
iaa99_	0.071619212	0.02122052	3.37	0.0007
etag0_2	-0.200857208	0.01567355	-12.82	<.0001
etag5_6	0.017222198	0.01682071	1.02	0.3059
etag7_	-0.027987565	0.01800367	-1.55	0.1201
maa1at1	0.042800242	0.01586685	2.70	0.0070
maa1at3	0.045410199	0.01473023	3.08	0.0021
maa1at4	0.038929018	0.01813061	2.15	0.0318
maa1at5	0.017283656	0.01567479	1.10	0.2702
libre	-0.162704153	0.01301826	-12.50	<.0001
pieces1	-0.219308163	0.02751265	-7.97	<.0001
pieces2	-0.083498758	0.01840980	-4.54	<.0001
pieces4_	0.101459265	0.01665839	6.09	<.0001
hab1	-0.085036038	0.01378628	-6.17	<.0001
hab3	0.052485211	0.01571614	3.34	0.0008
hab4_	0.099559079	0.01655012	6.02	<.0001

Résultats de la PROC LIFEREG :

The LIFEREG Procedure
Model Information

Data Set	WORK.COLL
Dependent Variable	Log(cmm1min)
Dependent Variable	Log(cmm1max)
Number of Observations	13842
Noncensored Values	11120
Right Censored Values	0
Left Censored Values	2722
Interval Censored Values	0
Name of Distribution	Lognormal
Log Likelihood	-9336.339224
Number of Observations Read	14031
Number of Observations Used	13842
Missing Values	189

Algorithm converged.

Analyse des Résultats estimés du paramètre

Paramètre	DF	Estimation	Erreur standard	95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	3.4581	0.0294	3.4004	3.5157	13818.7	<.0001
eaufroide	1	0.3213	0.0130	0.2958	0.3468	610.94	<.0001
eauchaude	1	0.0763	0.0141	0.0486	0.1040	29.21	<.0001
chauffcoll	1	0.4793	0.0147	0.4505	0.5081	1063.61	<.0001
ascenseur	1	0.1082	0.0154	0.0780	0.1385	49.18	<.0001
gardien	1	0.0829	0.0119	0.0596	0.1061	48.81	<.0001
autreschar	1	0.3111	0.0144	0.2828	0.3393	464.88	<.0001
surf_29	1	-0.1664	0.0304	-0.2260	-0.1068	29.95	<.0001
surf30_54	1	-0.0612	0.0181	-0.0966	-0.0258	11.50	0.0007
surf75_94	1	0.0494	0.0161	0.0178	0.0810	9.41	0.0022
surf95_	1	0.1518	0.0233	0.1061	0.1974	42.40	<.0001
tu_m50000	1	-0.1078	0.0182	-0.1434	-0.0721	35.14	<.0001
tu_m200000	1	-0.0445	0.0150	-0.0740	-0.0150	8.74	0.0031
tu_paris	1	0.2360	0.0137	0.2091	0.2629	296.42	<.0001
dec1	1	-0.0082	0.0201	-0.0476	0.0313	0.17	0.6846
dec2	1	-0.0440	0.0202	-0.0835	-0.0045	4.76	0.0292
dec3	1	-0.0468	0.0210	-0.0879	-0.0057	4.99	0.0255
dec4	1	0.0033	0.0216	-0.0389	0.0456	0.02	0.8780
dec6	1	-0.0031	0.0228	-0.0478	0.0416	0.02	0.8906
dec7	1	-0.0179	0.0240	-0.0650	0.0293	0.55	0.4575
dec8	1	0.0468	0.0251	-0.0023	0.0959	3.49	0.0618
dec9	1	0.0736	0.0238	0.0270	0.1202	9.59	0.0020
iaa_48	1	-0.1998	0.0167	-0.2325	-0.1670	142.52	<.0001
iaa75_81	1	0.0287	0.0177	-0.0060	0.0634	2.62	0.1054
iaa82_89	1	0.0758	0.0221	0.0324	0.1192	11.73	0.0006
iaa90_98	1	0.0368	0.0224	-0.0070	0.0806	2.71	0.0999
iaa99_	1	0.0644	0.0209	0.0236	0.1053	9.55	0.0020
etag0_2	1	-0.1691	0.0151	-0.1986	-0.1395	125.84	<.0001
etag5_6	1	0.0312	0.0165	-0.0012	0.0636	3.57	0.0590
etag7_	1	-0.0068	0.0176	-0.0413	0.0277	0.15	0.6988
maa1at1	1	0.0434	0.0155	0.0129	0.0738	7.81	0.0052
maa1at3	1	0.0464	0.0144	0.0182	0.0746	10.41	0.0013
maa1at4	1	0.0401	0.0177	0.0054	0.0748	5.13	0.0235
maa1at5	1	0.0111	0.0152	-0.0187	0.0410	0.53	0.4652
libre	1	-0.1518	0.0127	-0.1767	-0.1269	143.11	<.0001
pieces1	1	-0.2200	0.0268	-0.2726	-0.1674	67.19	<.0001
pieces2	1	-0.0809	0.0181	-0.1163	-0.0455	20.07	<.0001
pieces4_	1	0.1032	0.0162	0.0715	0.1348	40.81	<.0001
hab1	1	-0.0851	0.0135	-0.1115	-0.0587	39.83	<.0001
hab3	1	0.0554	0.0154	0.0253	0.0855	13.04	0.0003
hab4_	1	0.1015	0.0161	0.0699	0.1331	39.62	<.0001
Scale	1	0.5356	0.0035	0.5287	0.5426		

Redressement des charges locatives dans l'habitat individuel

Résultats de la régression par les MCO :

The GLM Procedure

Number of Observations Read 1582
Number of Observations Used 1537

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	33	426.175700	12.914415	23.26	<.0001
Error	1503	834.557604	0.555261		
Corrected Total	1536	1260.733303			

R-carré	Coeff Var	Racine MSE	log_cmm1 Moyenne
0.338038	24.48687	0.745159	3.043094

Paramètre	Estimation	Erreur standard	Valeur du test t	Pr > |t|
Intercept	2.954823674	0.10482888	28.19	<.0001
eaufroide	0.983748560	0.04991679	19.71	<.0001
autreschar	0.220145936	0.04718003	4.67	<.0001
surf_60	0.081314254	0.08124732	1.00	0.3171
surf85_109	0.028687944	0.05074989	0.57	0.5720
surf110_	0.153415569	0.06682543	2.30	0.0218
maa1at1	0.065015493	0.05782396	1.12	0.2610
maa1at3	-0.026455329	0.05585767	-0.47	0.6358
maa1at4	-0.100455541	0.07156426	-1.40	0.1606
maa1at5	0.017403970	0.06106857	0.28	0.7757
tu_rural	-0.308074791	0.06216787	-4.96	<.0001
tu_m10000	-0.137826594	0.06655048	-2.07	0.0385
tu_m50000	-0.095303748	0.05949772	-1.60	0.1094
tu_m200000	0.010036058	0.05860932	0.17	0.8641
tu_paris	0.307609890	0.07752039	3.97	<.0001
dec1	-0.033377696	0.08069308	-0.41	0.6792
dec2	-0.111308700	0.07681129	-1.45	0.1475
dec3	-0.026789500	0.07266714	-0.37	0.7124
dec4	-0.017912028	0.07147718	-0.25	0.8022
dec6	0.027998297	0.07821113	0.36	0.7204
dec7	0.115559439	0.08012967	1.44	0.1495
dec8	0.060289317	0.08784903	0.69	0.4926
dec9	0.169952260	0.08118647	2.09	0.0365
iaa_48	-0.101278904	0.05764937	-1.76	0.0792
iaa75_81	-0.019047423	0.07719555	-0.25	0.8051
iaa82_89	0.101019617	0.07574394	1.33	0.1825
iaa90_98	0.134233124	0.07581224	1.77	0.0768
iaa99_	0.097916672	0.06389103	1.53	0.1256
libre	-0.289809717	0.04949847	-5.85	<.0001
pieces_2	-0.169623787	0.09420349	-1.80	0.0720
pieces3	-0.074212554	0.05477536	-1.35	0.1757
hab1	-0.224168985	0.06603483	-3.39	0.0007
hab2	-0.174200214	0.05630134	-3.09	0.0020
hab4_	0.057903455	0.05544802	1.04	0.2965

Annexe 10 : Les facteurs liés à la non-réponse au salaire en clair

Pour essayer de déterminer quelques caractéristiques des non-répondants aux salaires en clair, on a modélisé par une régression logistique la probabilité de donner un montant de salaire en clair parmi les salariés.

The LOGISTIC Procedure

Informations sur le modèle

Data Set	WORK.LOGIT
Response Variable	NR
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	37632
Number of Observations Used	37632

Profil de réponse

Valeur ordonnée	NR	Fréquence totale
1	0	34467
2	1	3165

Probability modeled is NR=0.

État de convergence du modèle

Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	21729.210	21334.157
SC	21737.745	21539.012
-2 Log L	21727.210	21286.157

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	441.0523	23	<.0001
Score	441.7499	23	<.0001
Wald	432.8549	23	<.0001

Analyse des estimations de la vraisemblance maximum

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept	1	2.5661	0.0656	1531.2442	<.0001
etranger	1	0.1760	0.0813	4.6914	0.0303
celib	1	-0.2949	0.0519	32.3103	<.0001
veu_div	1	0.1250	0.0793	2.4866	0.1148
age15_24	1	-0.3054	0.0742	16.9435	<.0001
age25_34	1	0.0539	0.0557	0.9361	0.3333
age45_54	1	-0.2457	0.0529	21.5736	<.0001
age55_99	1	-0.5868	0.0658	79.5688	<.0001
tu_rural	1	0.1564	0.0597	6.8582	0.0088
tu_m10000	1	-0.0571	0.0693	0.6777	0.4104
tu_m50000	1	-0.2131	0.0684	9.7060	0.0018
tu_m200000	1	-0.1958	0.0584	11.2288	0.0008
tu_paris	1	0.2159	0.0567	14.5054	0.0001
cs_cadre	1	-0.4720	0.0572	68.0650	<.0001
cs_ouvrier	1	0.0104	0.0529	0.0383	0.8448
cs_autre	1	-0.1945	0.0514	14.3515	0.0002
cdd	1	-0.2387	0.0603	15.6689	<.0001
cdi_partiel	1	0.0458	0.0584	0.6137	0.4334
autre_type	1	0.1827	0.0729	6.2879	0.0122
sal_etat	1	0.3364	0.0560	36.1353	<.0001

sal_coll_loc	1	0.3926	0.0724	29.3717	<.0001
non_propri	1	0.2080	0.0436	22.7991	<.0001
pers_seule	1	0.3237	0.0708	20.9017	<.0001
autre	1	-0.2111	0.0604	12.2311	0.0005

Estimations des rapports de cotes

Effet	Point Estimate	95% Limites de confiance de Wald	
etranger	1.192	1.017	1.398
celib	0.745	0.673	0.824
veu_div	1.133	0.970	1.324
age15_24	0.737	0.637	0.852
age25_34	1.055	0.946	1.177
age45_54	0.782	0.705	0.868
age55_99	0.556	0.489	0.633
tu_rural	1.169	1.040	1.315
tu_m10000	0.945	0.825	1.082
tu_m50000	0.808	0.707	0.924
tu_m200000	0.822	0.733	0.922
tu_paris	1.241	1.110	1.387
cs_cadre	0.624	0.558	0.698
cs_ouvrier	1.010	0.911	1.121
cs_autre	0.823	0.744	0.910
cdd	0.788	0.700	0.886
cdi_partiel	1.047	0.934	1.174
autre_type	1.200	1.041	1.385
sal_etat	1.400	1.254	1.562
sal_coll_loc	1.481	1.285	1.707
non_propri	1.231	1.130	1.341
pers_seule	1.382	1.203	1.588
autre	0.810	0.719	0.911

Association des probabilités prédites et des réponses observées

Percent Concordant	59.8	Somers' D	0.215
Percent Discordant	38.3	Gamma	0.219
Percent Tied	1.9	Tau-a	0.033
Pairs	109088055	c	0.607

Annexe 11 : Détail des redressements des salaires

Imputation de l'activité de l'établissement pour les catégories 1 et 2 de salariés

Nombre d'utilisations des donneurs :

	Nombre d'utilisations	Fréquence	%
Pour les non-retraités	1	1 587	96,9
	2	49	3,0
	3	1	0,1
	Moyenne	1,0	
	Nombre d'utilisations	Fréquence	%
Pour les retraités	1	422	99,5
	2	2	0,5
	Moyenne	1,0	

Seuls 13 donneurs sont utilisés une fois dans chacun des deux hot-deck.

Imputation des salaires des hommes de la première catégorie

► Résultats de la régression robuste :

The ROBUSTREG Procedure

Model Information

Data Set	WORK.CAT1H2
Dependent Variable	rmsaltote_log
Number of Independent Variables	31
Number of Observations	16972
Missing Values	1657
Method	M Estimation

Number of Observations Read	18629
Number of Observations Used	16972
Missing Values	1657

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
log_age	3.4340	3.6636	3.8712	3.6294	0.2933	0.3078
etranger	0	0	0	0.0979	0.2972	0
celib	0	0	1.0000	0.3733	0.4837	0
veu_div	0	0	0	0.0597	0.2370	0
dip0	0	0	0	0.1766	0.3814	0
dipBrev	0	0	0	0.0636	0.2441	0
dipBac	0	0	0	0.1537	0.3606	0
dip1cycl	0	0	0	0.1183	0.3229	0
dip23cycl	0	0	0	0.1699	0.3756	0
sal_	0	0	0	0.0966	0.2954	0
cdd	0	0	0	0.0967	0.2955	0
cdi_partiel	0	0	0	0.0675	0.2508	0
autre_type	0	0	0	0.0748	0.2630	0
sal_etat	0	0	0	0.1403	0.3473	0
sal_coll_loc	0	0	0	0.0653	0.2471	0
classh1	0	0	0	0.1117	0.3150	0
classh3	0	0	0	0.0903	0.2866	0
classh4	0	0	0	0.0505	0.2190	0
classh5	0	0	0	0.0672	0.2503	0
classh6	0	0	0	0.0585	0.2347	0
classh7	0	0	0	0.1455	0.3526	0
classh8	0	0	0	0.0653	0.2471	0
classh9	0	0	0	0.1420	0.3491	0
acthd_indus	0	0	0	0.2093	0.4069	0

acthd_tert_n	0	0	0	0.2182	0.4130	0
acthd_autre	0	0	0	0.1726	0.3779	0
tu_rural	0	0	0	0.1804	0.3845	0
tu_m10000	0	0	0	0.0913	0.2881	0
tu_m50000	0	0	0	0.0858	0.2801	0
tu_m200000	0	0	0	0.1321	0.3386	0
tu_paris	0	0	0	0.2184	0.4131	0
rmsaltote_lo	9.4921	9.7981	10.1266	9.7223	0.7659	0.4598

Parameter Estimates

Parameter	DF	Estimation	Erreur standard	95% Confidence Limits		Khi 2	Pr > Khi 2
Intercept	1	8.2359	0.0407	8.1562	8.3156	41009.2	<.0001
log_age	1	0.4151	0.0108	0.3941	0.4362	1491.14	<.0001
etranger	1	-0.1149	0.0092	-0.1329	-0.0969	156.95	<.0001
celib	1	-0.0704	0.0062	-0.0825	-0.0582	128.69	<.0001
veu_div	1	-0.0428	0.0106	-0.0635	-0.0220	16.33	<.0001
dip0	1	-0.0838	0.0076	-0.0988	-0.0689	121.23	<.0001
dipBrev	1	-0.0106	0.0108	-0.0317	0.0105	0.97	0.3239
dipBac	1	0.0495	0.0079	0.0340	0.0650	39.34	<.0001
dip1cycl	1	0.0883	0.0090	0.0706	0.1059	95.76	<.0001
dip23cycl	1	0.1507	0.0095	0.1320	0.1693	249.82	<.0001
sal_	1	-0.2462	0.0094	-0.2646	-0.2279	691.20	<.0001
cdd	1	-0.2271	0.0091	-0.2451	-0.2092	617.15	<.0001
cdi_partiel	1	-0.1069	0.0099	-0.1262	-0.0875	117.07	<.0001
autre_type	1	-0.4238	0.0107	-0.4448	-0.4028	1561.44	<.0001
sal_etat	1	0.0932	0.0169	0.0600	0.1264	30.29	<.0001
sal_coll_loc	1	0.0253	0.0169	-0.0078	0.0584	2.25	0.1334
classh1	1	-0.1350	0.0091	-0.1529	-0.1171	219.20	<.0001
classh3	1	0.1369	0.0099	0.1175	0.1563	190.88	<.0001
classh4	1	0.2100	0.0187	0.1734	0.2466	126.37	<.0001
classh5	1	0.2213	0.0110	0.1997	0.2430	401.67	<.0001
classh6	1	0.4319	0.0194	0.3939	0.4699	496.45	<.0001
classh7	1	0.5830	0.0102	0.5630	0.6029	3280.25	<.0001
classh8	1	0.0305	0.0175	-0.0038	0.0647	3.04	0.0812
classh9	1	-0.0308	0.0089	-0.0482	-0.0135	12.10	0.0005
acthd_indus	1	0.0899	0.0069	0.0763	0.1035	167.94	<.0001
acthd_tert_n	1	-0.0471	0.0098	-0.0664	-0.0278	22.97	<.0001
acthd_autre	1	0.0268	0.0076	0.0120	0.0417	12.54	0.0004
tu_rural	1	0.0148	0.0075	0.0002	0.0294	3.94	0.0472
tu_m10000	1	0.0345	0.0093	0.0162	0.0528	13.70	0.0002
tu_m50000	1	0.0025	0.0095	-0.0162	0.0212	0.07	0.7928
tu_m200000	1	-0.0073	0.0081	-0.0233	0.0086	0.81	0.3691
tu_paris	1	0.0516	0.0071	0.0376	0.0656	52.30	<.0001
Scale	1	0.2937					

Diagnostics Summary

Observation Type	Proportion	Cutoff
Outlier	0.0724	3.0000
Leverage	0.0000	6.9449

Goodness-of-Fit

Statistic	Value
R-Square	0.3953
AICR	22569.44
BICR	22830.14
Deviance	1942.190

► **Résultats de la régression par les MCO des salaires non atypiques :**

The REG Procedure
Model: MODEL1
Dependent Variable: rmsaltote_log

Number of Observations Read	18534
Number of Observations Used	16877
Number of Observations with Missing Values	1657

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	31	4334.99921	139.83868	524.07	<.0001
Error	16845	4494.78023	0.26683		
Corrected Total	16876	8829.77943			

Root MSE	0.51656	R-Square	0.4910
Dependent Mean	9.72804	Adj R-Sq	0.4900
Coeff Var	5.30998		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	7.83964	0.06625	118.33	<.0001
log_age	1	0.51579	0.01752	29.44	<.0001
etranger	1	-0.15623	0.01494	-10.45	<.0001
celib	1	-0.08236	0.01008	-8.17	<.0001
veu_div	1	-0.07127	0.01719	-4.15	<.0001
dip0	1	-0.11154	0.01239	-9.00	<.0001
dipBrev	1	-0.01763	0.01749	-1.01	0.3136
dipBac	1	0.03884	0.01282	3.03	0.0025
dip1cycl	1	0.08165	0.01466	5.57	<.0001
dip23cycl	1	0.13135	0.01553	8.46	<.0001
sal_	1	-0.29684	0.01530	-19.41	<.0001
cdd	1	-0.39811	0.01489	-26.74	<.0001
cdi_partiel	1	-0.18962	0.01606	-11.81	<.0001
autre_type	1	-0.59345	0.01759	-33.74	<.0001
sal_etat	1	0.07956	0.02752	2.89	0.0038
sal_coll_loc	1	0.01274	0.02743	0.46	0.6423
classh1	1	-0.20230	0.01485	-13.62	<.0001
classh3	1	0.12118	0.01611	7.52	<.0001
classh4	1	0.24790	0.03036	8.16	<.0001
classh5	1	0.25118	0.01794	14.00	<.0001
classh6	1	0.47666	0.03153	15.12	<.0001
classh7	1	0.62473	0.01657	37.70	<.0001
classh8	1	0.06579	0.02840	2.32	0.0205
classh9	1	-0.08056	0.01440	-5.59	<.0001
acthd_indus	1	0.09766	0.01129	8.65	<.0001
acthd_tert_n	1	-0.04805	0.01598	-3.01	0.0026
acthd_autre	1	0.01953	0.01232	1.59	0.1130
tu_rural	1	0.02152	0.01212	1.78	0.0758
tu_m10000	1	0.04204	0.01518	2.77	0.0056
tu_m50000	1	0.01877	0.01553	1.21	0.2268
tu_m200000	1	-0.00770	0.01325	-0.58	0.5609
tu_paris	1	0.04837	0.01162	4.16	<.0001

Durbin-Watson D	1.887
Number of Observations	16877
1st Order Autocorrelation	0.057

► **Vérification de l'homoscédasticité des résidus**

i) Test de White

```
proc model data= indh_sal_reg;
parms a b c d e f g h j k l m n o p q r s t u v w x y z aa ab ac ad ae af ag;
rmsaltote_log = a + b*log_age + c*etranger +
d*dip0 + e*dipBrev + f*dipBac + g*dip1cycl + h*dip23cycl +
j*sal + k*cdd + l*cdi_partiel + m*autre_type + n*sal_etat + o*sal_coll_loc +
p*classh1 + q*classh3 + r*classh4 + s*classh5 + t*classh6 + u*classh7 +
v*classh8 + w*classh9 + x*acthd_indus + y*acthd_tert_n + z*acthd_autre +
aa*celib + ab*veu_div + ac*tu_rural + ad*tu_m10000 + ae*tu_m50000 +
af*tu_m200000 + ag*tu_paris;
fit rmsaltote_log / white;
quit;
```

```
                                The MODEL Procedure
                                  Model Summary
                              Model Variables          1
                              Parameters              32
                              Equations                1
                              Number of Statements     1

                              The Equation to Estimate is
 rmsaltote_log =  F(a(1), b(log_age), c(etranger), d(dip0), e(dipBrev), f(dipBac), g(dip1cycl),
h(dip23cycl), j(sal_), k(cdd), l(cdi_partiel), m(autre_type), n(sal_etat),
                 o(sal_coll_loc), p(classh1), q(classh3), r(classh4), s(classh5),
                 t(classh6), u(classh7), v(classh8), w(classh9), x(acthd_indus),
                 y(acthd_tert_n), z(acthd_autre), aa(celib), ab(veu_div), ac(tu_rural),
                 ad(tu_m10000), ae(tu_m50000), af(tu_m200000), ag(tu_paris))
             NOTE: At OLS Iteration 1 CONVERGE=0.001 Criteria Met.

                              OLS Estimation Summary
                                 Data Set Options
                                DATA=    INDH_SAL_REG

                                Minimization Summary
                           Parameters Estimated           32
                           Method                      Gauss
                           Iterations                      1

                              Final Convergence Criteria
                              R                 8.85E-12
                              PPC               9.46E-11
                              RPC(a)            77619.22
                              Object            0.997201
                              Trace(S)          0.266832
                              Objective Value   0.266326

                                Observations Processed
                                   Read       18534
                                   Solved     18534
                                   Used       16877
                                   Missing     1657

                 Non linéaire OLS Récapitulatif des erreurs résiduelles
                Modèle    Erreur                                 Racine              R carré
 Équation           DF        DF         SSE         MSE         MSE    R-carré        aj.
rmsaltote_log       32     16845      4494.8      0.2668      0.5166     0.4910     0.4900

                          Estimations OLS Parameter non linéaires
                                       Erreur
                                      standard                          Approx
          Parameter       Estimation       appr.    Valeur du test t     Pr > |t|
```

a	7.839641	0.0663	118.33	<.0001
b	0.515794	0.0175	29.44	<.0001
c	-0.15623	0.0149	-10.45	<.0001
d	-0.11154	0.0124	-9.00	<.0001
e	-0.01763	0.0175	-1.01	0.3136
f	0.038839	0.0128	3.03	0.0025
g	0.081649	0.0147	5.57	<.0001
h	0.131352	0.0155	8.46	<.0001
j	-0.29684	0.0153	-19.41	<.0001
k	-0.39811	0.0149	-26.74	<.0001
l	-0.18962	0.0161	-11.81	<.0001
m	-0.59345	0.0176	-33.74	<.0001
n	0.079558	0.0275	2.89	0.0038
o	0.012742	0.0274	0.46	0.6423
p	-0.2023	0.0149	-13.62	<.0001
q	0.121175	0.0161	7.52	<.0001
r	0.247897	0.0304	8.16	<.0001
s	0.251185	0.0179	14.00	<.0001
t	0.476662	0.0315	15.12	<.0001
u	0.624725	0.0166	37.70	<.0001
v	0.065794	0.0284	2.32	0.0205
w	-0.08056	0.0144	-5.59	<.0001
x	0.097663	0.0113	8.65	<.0001
y	-0.04805	0.0160	-3.01	0.0026
z	0.019532	0.0123	1.59	0.1130
aa	-0.08236	0.0101	-8.17	<.0001
ab	-0.07127	0.0172	-4.15	<.0001
ac	0.021525	0.0121	1.78	0.0758
ad	0.042039	0.0152	2.77	0.0056
ae	0.018769	0.0155	1.21	0.2268
af	-0.0077	0.0132	-0.58	0.5609
ag	0.048374	0.0116	4.16	<.0001

Nombre d'observations		Statistiques pour Système	
Used	16877	Objective	0.2663
Missing	1657	Objective*N	4495

Test d'hétéroscédasticité

Équation	Test	Statistique	DF	Pr > Khi 2	Variables
rmsaltote_log	White's Test	1466	432	<.0001	Cross of all vars

D'après le test de White, on rejette l'hypothèse d'homoscédasticité.

ii) Tests de Breusch-Pagan (pour repérer les variables les plus en cause dans l'hétéroscédasticité)

C'est le type d'emploi qui semble être la variable la plus en cause :

Variable	Statistique	Pr > Khi 2
CDD	285,5	<,0001
Autre type d'emploi	242,6	<,0001

Edition des écart-types des résidus des répondants selon le type d'emploi :

te	Écart-type
1	0.7697167
2	0.4282339
3	0.6001127
4	0.7844144

► **Résultats de la PROC LIFEREG :**

The LIFEREG Procedure
Model Information

Data Set	WORK.INDH_SAL	
Dependent Variable	Log(rmsaltotemin)	
Dependent Variable	Log(rmsaltotemax)	
Number of Observations	17371	
Noncensored Values	16877	
Right Censored Values	9	*(9 observations n'ont pas de borne maximale)*
Left Censored Values	0	*(toutes les observations ont une borne minimale)*
Interval Censored Values	485	
Zero or Negative Response	25	*(485 + 25 + 9 = 519 ont une tranche renseignée)*
Name of Distribution	Lognormal	
Log Likelihood	-13542.41796	

Number of Observations Read	18534
Number of Observations Used	17371
Missing Values	1138

Algorithm converged.

Analyse des effets de Type III

Effet	DF	Khi 2 de Wald	Pr > Khi 2
log_age	1	901.6385	<.0001
etranger	1	114.9148	<.0001
celib	1	71.6871	<.0001
veu_div	1	17.1785	<.0001
dip0	1	86.9618	<.0001
dipBrev	1	1.0890	0.2967
dipBac	1	9.3238	0.0023
dip1cycl	1	31.8308	<.0001
dip23cycl	1	77.5626	<.0001
sal_	1	394.2690	<.0001
cdd	1	712.5758	<.0001
cdi_partiel	1	135.7825	<.0001
autre_type	1	1142.0543	<.0001
sal_etat	1	8.6906	0.0032
sal_coll_loc	1	0.2556	0.6132
classh1	1	188.3764	<.0001
classh3	1	57.7094	<.0001
classh4	1	67.0607	<.0001
classh5	1	212.5566	<.0001
classh6	1	233.6937	<.0001
classh7	1	1515.8577	<.0001
classh8	1	5.3860	0.0203
classh9	1	29.2454	<.0001
acthd_indus	1	80.4510	<.0001
acthd_tert_n	1	9.7673	0.0018
acthd_autre	1	2.7032	0.1001
tu_rural	1	3.0690	0.0798
tu_m10000	1	7.1843	0.0074
tu_m50000	1	1.3348	0.2480
tu_m200000	1	0.1493	0.6992
tu_paris	1	17.7782	<.0001

Analyse des Résultats estimés du paramètre

Paramètre	DF	Estimation	Erreur standard	95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	7.8388	0.0651	7.7112	7.9665	14486.7	<.0001
log_age	1	0.5169	0.0172	0.4832	0.5507	901.64	<.0001
etranger	1	-0.1576	0.0147	-0.1865	-0.1288	114.91	<.0001
celib	1	-0.0839	0.0099	-0.1033	-0.0645	71.69	<.0001
veu_div	1	-0.0695	0.0168	-0.1024	-0.0366	17.18	<.0001
dip0	1	-0.1140	0.0122	-0.1380	-0.0901	86.96	<.0001
dipBrev	1	-0.0180	0.0172	-0.0518	0.0158	1.09	0.2967

dipBac	1	0.0386	0.0126	0.0138	0.0633	9.32	0.0023
dip1cycl	1	0.0813	0.0144	0.0531	0.1095	31.83	<.0001
dip23cycl	1	0.1342	0.0152	0.1043	0.1640	77.56	<.0001
sal_	1	-0.3001	0.0151	-0.3297	-0.2704	394.27	<.0001
cdd	1	-0.3923	0.0147	-0.4211	-0.3635	712.58	<.0001
cdi_partiel	1	-0.1840	0.0158	-0.2149	-0.1530	135.78	<.0001
autre_type	1	-0.5870	0.0174	-0.6210	-0.5529	1142.05	<.0001
sal_etat	1	0.0802	0.0272	0.0269	0.1335	8.69	0.0032
sal_coll_loc	1	0.0137	0.0271	-0.0394	0.0669	0.26	0.6132
classh1	1	-0.2017	0.0147	-0.2305	-0.1729	188.38	<.0001
classh3	1	0.1204	0.0159	0.0894	0.1515	57.71	<.0001
classh4	1	0.2460	0.0300	0.1871	0.3049	67.06	<.0001
classh5	1	0.2562	0.0176	0.2217	0.2906	212.56	<.0001
classh6	1	0.4763	0.0312	0.4153	0.5374	233.69	<.0001
classh7	1	0.6319	0.0162	0.6001	0.6637	1515.86	<.0001
classh8	1	0.0652	0.0281	0.0101	0.1203	5.39	0.0203
classh9	1	-0.0766	0.0142	-0.1044	-0.0489	29.25	<.0001
acthd_indus	1	0.0993	0.0111	0.0776	0.1210	80.45	<.0001
acthd_tert_n	1	-0.0492	0.0157	-0.0800	-0.0183	9.77	0.0018
acthd_autre	1	0.0200	0.0121	-0.0038	0.0438	2.70	0.1001
tu_rural	1	0.0209	0.0119	-0.0025	0.0443	3.07	0.0798
tu_m10000	1	0.0400	0.0149	0.0108	0.0693	7.18	0.0074
tu_m50000	1	0.0176	0.0152	-0.0122	0.0474	1.33	0.2480
tu_m200000	1	-0.0050	0.0130	-0.0305	0.0205	0.15	0.6992
tu_paris	1	0.0483	0.0114	0.0258	0.0707	17.78	<.0001
Scale	1	0.5149	0.0028	0.5095	0.5203		

Résultats de l'imputation des salaires restants (catégories 2 et 3) par hot-deck aléatoire par classes

Comparaison de la distribution des salaires avant et après imputation :

- pour la catégorie 2 :

	Avant	Après
Nombre d'individus	609 335	609 335
Effectif renseigné	522 001	609 335
Effectif non renseigné	87 334	-
% de non-réponses	14,3	-
Moyenne	27 525,1	27 474,6
Intervalle de confiance*	[25 222 ; 29 829]	[25 314 ; 29 635]
Écart-type	31 946,2	30 756,3
Minimum	15	15
1er quartile	11 301	11 039
Médiane	19 741	20 000
3ème quartile	33 657	33 657
Maximum	483 500	483 500

- pour la catégorie 3 :

	Avant	Après
Nombre d'individus	542 631	542 631
Effectif renseigné	502 029	542 631
Effectif non renseigné	40 602	-
% de non-réponses	7,5	-
Moyenne	5 998,3	5 974,7
Intervalle de confiance*	[5 442 ; 6 554]	[5 262 ; 6 387]
Écart-type	6 906,6	6 934,3
Minimum	80	80
1er quartile	1 900	1 800
Médiane	4 000	4 000
3ème quartile	8 000	8 000
Maximum	65 858	65 858

* Les intervalles de confiance sont calculés sous l'hypothèse d'un sondage aléatoire simple.

Annexe 12 : Comparaison de la PROC LIFEREG et de la PROC ROBUSTREG

Pour mesurer l'apport de l'information en tranches pour les salaires, on peut comparer les résultats des régressions du logarithme du salaire par les MCO et par la PROC ROBUSTREG à ceux obtenus avec la PROC LIFEREG.
Comme seuls les salaires de la première catégorie sont modélisables, on restreint les régressions aux individus de cette catégorie, en estimant séparément les modèles pour les hommes et les femmes. Afin que les tranches soient au même niveau que les revenus en clair, on restreint les régressions aux individus appartenant à un ménage dans lequel ils étaient à la fois le seul salarié et la personne de référence ou son conjoint.

Résultats pour les hommes :

		Régression par les MCO	PROC ROBUSTREG	PROC LIFEREG
Effectif de répondants pris en compte dans la régression		5873	5873	6265
R^2		0,44	0,38	
	Constante	8,37	8,55	8,35
	Log (âge)	**0,37**	**0,33**	**0,37**
Nationalité *(français)*	étranger	**-0,18**	**-0,12**	**-0,19**
Statut matrimonial *(marié)*	célibataire	**-0,07**	**-0,07**	**-0,08**
	veuf ou divorcé	-0,06	**-0,04**	**-0,07**
Diplôme *(CAP BEP)*	aucun diplôme	**-0,13**	**-0,10**	**-0,13**
	brevet	ns	ns	ns
	baccalauréat	ns	0,03	ns
	1er cycle	ns	**0,07**	ns
	2e ou 3e cycle	0,06	**0,11**	0,07
Types de revenus perçus *(salaires uniquement)*	salaire et autres revenus	**-0,31**	**-0,32**	**-0,32**
Type d'emploi *(CDI à temps complet)*	CDD	**-0,35**	**-0,20**	**-0,34**
	CDI à temps partiel	**-0,23**	**-0,15**	**-0,21**
	autre	**-0,53**	**-0,30**	**-0,51**
Statut *(salarié d'une collectivité locale)*	salarié de l'État	ns	**0,10**	ns
	salarié d'une entreprise ou d'un particulier	ns	ns	ns
Classification dans l'emploi *(ouvrier qualifié)*	manœuvre, ouvrier spécialisé	**-0,16**	**-0,13**	**-0,16**
	technicien	**0,16**	**0,15**	**0,16**
	catégorie B	**0,39**	**0,25**	**0,38**
	agent de maîtrise	**0,31**	**0,27**	**0,32**
	catégorie A	**0,62**	**0,52**	**0,62**
	ingénieur, cadre, directeur	**0,73**	**0,65**	**0,75**
	catégorie C-D	0,13	ns	**0,13**
	employé	**-0,09**	**-0,05**	**-0,08**
Activité de l'établissement *(tertiaire marchand)*	industrie	**0,11**	**0,08**	**0,11**
	tertiaire non marchand	ns	**-0,06**	ns
	autre	ns	**0,04**	ns
Taille de la tranche d'unité urbaine *(200 000 habitants et plus)*	rural	ns	**0,04**	ns
	moins de 10 000 hab.	0,07	**0,05**	0,07
	10 000 à 49 999 hab.	ns	ns	ns
	50 000 à 199 999 hab.	ns	ns	ns
	agglomération parisienne	*0,04*	0,03	0,04

Les coefficients sont présentés selon leur seuil de significativité : **en gras 1 %**, en romain 5 % et *en italique 10 %.* Au-delà, ils sont marqués ns (non significatif).

Résultats pour les femmes :

		Régression par les MCO	PROC ROBUSTREG	PROC LIFEREG
Effectif de répondants pris en compte dans la régression		5797	5797	6 091
R^2		0,45	0,37	
Constante		7,91	8,04	7,90
Log (âge)		**0,40**	**0,40**	**0,41**
Nationalité *(français)*	étranger	**-0,28**	**-0,22**	**-0,28**
Statut matrimonial *(marié)*	célibataire	**0,07**	**0,05**	**0,06**
	veuf ou divorcé	**0,07**	**0,06**	**0,07**
Diplôme *(CAP BEP)*	aucun diplôme	**-0,20**	**-0,15**	**-0,21**
	brevet	ns	ns	ns
	baccalauréat	**0,11**	**0,09**	**0,11**
	1er cycle	**0,18**	**0,16**	**0,18**
	2e ou 3e cycle	**0,16**	**0,18**	**0,15**
Types de revenus perçus *(salaires uniquement)*	salaire et autres revenus	**-0,35**	**-0,30**	**-0,34**
Type d'emploi *(CDI à temps complet)*	CDD	**-0,62**	**-0,41**	**-0,61**
	CDI à temps partiel	**-0,36**	**-0,30**	**-0,35**
	autre	**-0,67**	**-0,44**	**-0,65**
Statut *(salarié d'une collectivité locale)*	salarié de l'État	ns	ns	ns
	salarié d'une entreprise ou d'un particulier	ns	ns	ns
Classification dans l'emploi *(employé)*	manœuvre, ouvrier spécialisé	ns	*-0,03*	ns
	technicien, agent de maîtrise	**0,30**	**0,25**	**0,30**
	catégorie B	**0,37**	**0,27**	**0,36**
	catégorie A	**0,60**	**0,46**	**0,60**
	ingénieur, cadre, directeur	**0,68**	**0,61**	**0,68**
	catégorie C-D	**0,23**	0,09	**0,23**
Activité de l'établissement *(tertiaire marchand)*	tertiaire non marchand	ns	ns	ns
	autre	ns	ns	ns
Taille de la tranche d'unité urbaine *(200 000 habitants et plus)*	rural	ns	ns	ns
	moins de 10 000 hab.	ns	ns	ns
	10 000 à 49 999 hab.	*-0,05*	*-0,04*	ns
	50 000 à 199 999 hab.	ns	ns	ns
	agglomération parisienne	**0,11**	**0,07**	**0,11**

Les coefficients sont présentés selon leur seuil de significativité : **en gras 1 %**, en romain 5 % et *en italique 10 %.* Au-delà, ils sont marqués ns (non significatif).

Annexe 13 : Vérification des revenus imputés : comparaison avec d'autres enquêtes auprès des ménages

Pour les revenus, on peut comparer les résultats de l'enquête Logement 2006 avec ceux obtenus avec d'autres enquêtes auprès des ménages :

- l'enquête Revenus Fiscaux 2005 : issue d'un appariement de l'enquête Emploi avec les fichiers de la DGI, l'ERF est la source de référence pour les études sur la distribution des revenus. En 2005, elle porte sur près de 35 000 ménages.

- le dispositif SRCV 2005 (statistiques sur les ressources et les conditions de vie) : il s'intègre dans un projet européen (EU-SILC) visant à produire des statistiques harmonisées en matière de revenus et de conditions de vie des ménages et comporte près de 10 000 ménages répondants.

- l'enquête Patrimoine 2004 : elle vise à décrire les biens immobiliers, financiers et professionnels des ménages ainsi que les facteurs explicatifs des comportements patrimoniaux (revenus et situation financière notamment). Elle compte près de 10 000 ménages répondants.

Revenus perçus par les individus[67]

Salaires et primes

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre d'individus	24 366 563	26 287 529	25 214 188
Moyenne	18 661	18 618	17 361
Intervalle de confiance*	[18 513 ; 18 808]	[18 435 ; 18 801]	
Minimum	15	1	12
1er décile	5 208	3 151	3 800
2e décile	9 600	7 694	8 000
3e décile	12 300	11 882	11 595
4e décile	14 400	14 430	13 542
5e décile	16 500	16 492	15 477
6e décile	18 300	18 665	17 574
7e décile	21 085	21 497	20 068
8e décile	25 000	25 377	23 837
9e décile	32 000	32 640	30 267
Maximum	483 500	847 620	225 791

Indemnités de chômage

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre d'individus	2 749 950	4 749 820	3 851 910
Moyenne	6 454	5 765	6 077
Intervalle de confiance*	[6 301 ; 6 607]	[5 616 ; 5 913]	
Minimum	10	1	5
1er décile	1 000	680	992
2e décile	1 900	1 344	1 800
3e décile	2 800	2 157	2 816
4e décile	3 870	3 141	3 760
5e décile	4 990	4 226	4 810
6e décile	5 920	5 255	5 800
7e décile	7 650	6 902	7 317
8e décile	9 636	9 128	9 180
9e décile	12 257	12 227	11 776
Maximum	75 000	61 737	60 840

* Les intervalles de confiance sont calculés sous l'hypothèse d'un sondage aléatoire simple.

[67] Les résultats concernant les revenus sont calculés avec les poids définitifs.

Retraites et pensions

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre d'individus	13 140 276	13 154 488	12 857 372
Moyenne	13 967	14 203	14 686
Intervalle de confiance*	[13 774 ; 14 160]	[14 059 ; 14 346]	
Minimum	15	17	50
1er décile	2 400	3 429	3 905
2e décile	5 600	6 387	7 017
3e décile	7 800	8 560	8 704
4e décile	9 670	10 603	10 578
5e décile	11 900	12 576	12 416
6e décile	14 048	14 687	14 400
7e décile	16 680	17 004	16 738
8e décile	20 092	20 265	20 098
9e décile	26 277	25 174	25 711
Maximum	302 000	191 999	280 000

Revenus perçus par les ménages

Revenus non salariaux

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre de ménages	2 199 505	2 023 223	1 968 536
Moyenne	26 005	25 965	25 643
Intervalle de confiance*	[24 564 ; 27 446]	[24 213 ; 27 717]	
Minimum	- 90 000	- 984 461	2
1er décile	1 200	367	2 877
2e décile	5 300	2 346	6 318
3e décile	9 781	5 447	9 600
4e décile	13 000	9 195	12 000
5e décile	17 250	14 018	15 833
6e décile	20 500	19 852	21 203
7e décile	28 000	26 509	29 193
8e décile	37 000	39 832	37 500
9e décile	60 000	62 950	60 000
Maximum	687 000	742 739	360 000

Prestations familiales et aides à la scolarité

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre de ménages	6 599 739	6 808 005	6 611 193
Moyenne	3 145	3 681	2 989
Minimum	13	24	24
1er décile	311	833	338
2e décile	998	1 381	1 351
3e décile	1 380	1 381	1 351
4e décile	1 627	1 769	1 731
5e décile	1 916	2 033	1 864
6e décile	2 400	3 364	2 026
7e décile	3 600	4 578	3 291
8e décile	5 487	6 477	4 908
9e décile	7 395	8 283	6 849
Maximum	46 825	37 632	22 800

Allocations RMI

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005	SRCV 2005
Nombre de ménages	631 782	862 976	800 530
Moyenne	3 415	4 139	4 153
Intervalle de confiance*	[3 308 ; 3 522]	[4 025 ; 4 253]	
Minimum	40	665	4
1er décile	500	1 519	640
2e décile	1 140	2 296	1 661
3e décile	1 800	3 170	2 659
4e décile	2 640	4 038	3 405
5e décile	3 600	4 644	4 272
6e décile	4 400	4 644	4 488
7e décile	4 572	4 644	5 000
8e décile	5 196	5 174	6 276
9e décile	6 300	6 661	7 430
Maximum	13 032	17 235	13 172

Revenus fonciers des ménages non déficitaires

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005
Nombre de ménages	2 167 075	3 205 324
Moyenne Intervalle de confiance*	5 968 [5 662 ; 6 274]	5 807 [5 498 ; 6 116]
Minimum	1	1
1er décile	526	389
2e décile	1 300	952
3e décile	2 122	1 680
4e décile	3 300	2 412
5e décile	3 800	3 240
6e décile	5 000	4 082
7e décile	6 098	5 347
8e décile	8 305	7 598
9e décile	12 600	12 979
Maximum	100 000	183 057

Niveau de vie

Le niveau de vie permet de comparer les revenus de ménages de compositions différentes (taille et structure par âge). Il correspond au revenu disponible du ménage corrigé de façon à tenir compte du nombre de personnes qui le composent : c'est le revenu disponible du ménage par unité de consommation[68].

Sa distribution est la suivante parmi les ménages ayant des revenus non négatifs :

	Enquête Logement 2006	Enquête Patrimoine 2004
Moyenne Intervalle de confiance*	18 465,1 [18 323 ; 18 607]	16 824,9 [16 555 ; 17 094]
Ecart-type	13 756,9	11 070,6
Minimum	0	0
1er décile	6 400	7 552
2e décile	9 403	9 759
3e décile	11 808	11 643
4e décile	13 800	13 255
5e décile	16 000	14 775
6e décile	18 272	16 538
7e décile	21 000	18 807
8e décile	24 936	21 755
9e décile	32 362	27 115
Maximum	558 000	493 134

Allocations logement perçues par les locataires et propriétaires

	Enquête Logement 2006	Enquête Revenus Fiscaux 2005
Nombre de ménages	4 223 478	4 422 198
Moyenne Intervalle de confiance*	2 291,3 [2 259 ; 2 324]	2 136,5 [2 104 ; 2 169]
Ecart-type	1 228,3	1 216,8
Minimum	288	289
1er quartile	1 320	1 123
Médiane	2 220	2 035
3e quartile	3 024	2 876
Maximum	10 152	8 361

[68] Il est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation du ménage (il est donc le même pour tous les individus d'un même ménage). Les unités de consommation sont généralement calculées selon l'échelle d'équivalence dite de l'OCDE : on compte le nombre d'« équivalents adultes » qui composent chaque ménage en affectant à chaque individu un coefficient selon son poids dans la consommation du ménage. Le système de pondération est le suivant : le premier adulte (la personne de référence) a un poids de 1, les autres adultes et les enfants de 14 ans ou plus comptent pour 0,5 unité de consommation, les enfants de moins de 14 ans ont un poids de 0,3.
Remarque : on n'a calculé ici le niveau de vie que pour les logements dans lesquels il y a un seul ménage, c'est-à-dire dans lesquels il n'y a pas de budget séparé.

Annexe 14 : Comparaison de l'imputation par la méthode des résidus simulés et par hot-deck aléatoire par classes : effets sur la distribution

Pour imputer les salaires des individus de la première catégorie, les indemnités de chômage, les retraites ainsi que les dépenses en eau et en énergies, c'est la méthode des résidus simulés qui a été retenue. Cependant, à des fins de comparaison, on a aussi procédé à des imputations par hot-deck aléatoire par classes pour ces variables.

Salaires et primes

Hommes de la première catégorie

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre d'individus	11 497 939	11 497 939	*11 497 939*
Effectif renseigné	10 438 032	11 497 939	*11 497 939*
Effectif non renseigné	1 059 907	-	-
% de non-réponses	9,2	-	-
Moyenne	21 534,4	21 679,5	*21 607,2*
Intervalle de confiance*	[21 298 ; 21 771]	[21 453 ; 21 906]	*[21 381 ; 21 833]*
Écart-type	15 392,6	15 461,4	*15 380,7*
Minimum	15	15	*15*
1er quartile	14 000	13 940	*14 000*
Médiane	18 200	18 294	*19 294*
3ème quartile	25 005	25 527	*25 200*
Maximum	370 000	370 000	*370 000*

Femmes de la première catégorie

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre d'individus	11 061 050	11 061 050	*11 061 050*
Effectif renseigné	10 228 930	11 061 050	*11 061 050*
Effectif non renseigné	832 120	-	-
% de non-réponses	7,5	-	-
Moyenne	15 847,0	15 869,4	*15 852,9*
Intervalle de confiance*	[15 677 ; 16 017]	[15 711 ; 16 027]	*[15 696 ; 16 010]*
Écart-type	10 235,7	10 368,0	*10 328,7*
Minimum	15	15	*15*
1er quartile	9 840	9 600	*9 605*
Médiane	14 556	14 500	*14 500*
3ème quartile	20 000	20 028	*20 028*
Maximum	330 000	330 000	*330 000*

Indemnités de chômage

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre d'individus	2 619 488	2 619 488	*2 619 488*
Effectif renseigné	2 396 142	2 619 488	*2 619 488*
Effectif non renseigné	223 346	-	-
% de non-réponses	8,5	-	-
Moyenne	6 417,1	6 413,9	*6 355,8*
Intervalle de confiance*	[6 262 ; 6 572]	[6 261 ; 6 567]	*[6 206 ; 6 506]*
Écart-type	6 530,6	6 568,4	*6 432,2*
Minimum	10	10	*10*
1er quartile	2 400	2 310	*2 380*
Médiane	5 000	4 930	*4 926*
3ème quartile	8 565	8 520	*8 552*
Maximum	75 000	75 000	*75 000*

Retraites et pensions

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre d'individus	13 027 306	13 027 306	*13 027 306*
Effectif renseigné	11 466 412	13 027 306	*13 027 306*
Effectif non renseigné	1 560 894	-	-
% de non-réponses	12,0	-	-
Moyenne	13 592,0	13 934,2	*13 799,9*
Intervalle de confiance*	[13 399 ; 13 785]	[13 741 ; 14 127]	*[13 616 ; 13 982]*
Écart-type	11 560,3	12 372,7	*11 735,6*
Minimum	15	15	*15*
1er quartile	6 940	6 804	*7 020*
Médiane	11 832	11 845	*11 919*
3ème quartile	18 000	18 012	*18 000*
Maximum	302 000	302 000	*302 000*

Niveau de vie disponible des ménages (parmi les ménages ayant des revenus non négatifs)

	Après résidus simulés	*Après hot-deck*
Moyenne	18 465,1	*18 383,7*
Intervalle de confiance*	[18 323 ; 18 607]	*[18 243 ; 18 524]*
Ecart-type	13 756,9	*13 579,0*
Minimum	0	*0*
1er décile	6 400	*6 442*
2e décile	9 403	*9 419*
3e décile	11 808	*11 817*
4e décile	13 800	*13 800*
5e décile	16 000	*16 000*
6e décile	18 272	*18 269*
7e décile	21 000	*21 000*
8e décile	24 936	*24 833*
9e décile	32 362	*32 000*
Maximum	558 000	*558 000*

Dépenses en eau

Dans l'habitat individuel

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre de ménages	14 609 383	14 609 383	*14 609 383*
Effectif renseigné	13 184 737	14 609 383	*14 609 383*
Effectif non renseigné	1 424 646	-	-
% de non-réponses	9,8	-	-
Moyenne	309,0	307,7	*308,0*
Intervalle de confiance*	[305 ; 313]	[304 ; 311]	*[304 ; 312]*
Écart-type	234,1	231,8	*233,4*
Minimum	10	10	*10*
1er quartile	169	166	*167*
Médiane	258	254	*254*
3ème quartile	391	390	*390*
Maximum	5 112	5 112	*5 112*

Dans l'habitat collectif

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre de ménages	2 963 271	2 963 271	*2 963 271*
Effectif renseigné	2 231 996	2 963 271	*2 963 271*
Effectif non renseigné	731 275	-	-
% de non-réponses	24,7	-	-
Moyenne	265,2	262,6	*266,6*
Intervalle de confiance*	[259 ; 271]	[257 ; 268]	*[261 ; 272]*
Écart-type	183,2	182,4	*199,2*
Minimum	10	10	*10*
1er quartile	140	137	*140*
Médiane	216	210	*211*
3ème quartile	350	342	*340*
Maximum	2 400	2 400	*2 400*

Dépenses en énergies

	Avant imputation	Après résidus simulés	*Après hot-deck*
Nombre de ménages	24 876 207	24 876 207	*24 876 207*
Effectif renseigné	20 872 734	24 876 207	*24 876 207*
Effectif non renseigné	4 003 473	-	-
% de non-réponses	16,1	-	-
Moyenne	1 275,4	1 270,5	*1 270,0*
Intervalle de confiance*	[1 266 ; 1 285]	[1 261 ; 1 279]	*[1 261 ; 1 279]*
Écart-type	898,6	934,1	*929,4*
Minimum	13	13	*13*
1er quartile	650	620	*635*
Médiane	1 154	1 120	*1 129*
3ème quartile	1 162	1 659	*1 650*
Maximum	13 913	13 913	*13 913*

Annexe 15 : Impact sur le taux d'effort des méthodes d'imputation utilisées

Tranche de taille de l'unité urbaine :

	Taux brut (%)				Taux net (%)			
	Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*		Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*	
	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**
Commune rurale	15,5	[15,1 ; 15,9]	*15,6*	*[15,2 ; 16,0]*	14,9	[14,5 ; 15,3]	*14,9*	*[14,5 ; 15,3]*
Unité urbaine de moins de 5 000 habitants	16,5	[15,6 ; 17,4]	*16,6*	*[15,7 ; 17,5]*	15,5	[14 ,7 ; 16,3]	*15,6*	*[14,8 ; 16,4]*
Unité urbaine de 5 000 à 9 999 hab.	17,2	[16,3 ; 18,1]	*17,3*	*[16,4 ; 18,2]*	16,1	[15,2 ; 17,0]	*16,2*	*[15,3 ; 17,1]*
Unité urbaine de 10 000 à 19 999 hab.	16,8	[16,0 ; 17,6]	*16,9*	*[16,1 ; 17,7]*	15,4	[14,7 ; 16,1]	*15,5*	*[14,8 ; 16,2]*
Unité urbaine de 20 000 à 49 999 hab.	18,0	[17,8 ; 18,2]	*18,1*	*[17,9 ; 18,3]*	16,4	[16,2 ; 16,6]	*16,5*	*[16,3 ; 16,7]*
Unité urbaine de 50 000 à 99 999 hab.	18,4	[17,9 ; 18,9]	*18,5*	*[18,0 ; 19,0]*	16,4	[15,9 ; 16,9]	*16,5*	*[16,0 ; 17,0]*
Unité urbaine de 100 000 à 199 999 hab.	19,3	[18,6 ; 20,0]	*19,2*	*[18,6 ; 20,0]*	17,3	[16,7 ; 17,9]	*17,3*	*[16,7 ; 17,9]*
Unité urbaine de 200 000 à 1 999 999 hab.	19,2	[18,9 ; 19,5]	*19,2*	*[18,9 ; 19,5]*	17,3	[17,0 ; 17,6]	*17,4*	*[17,1 ; 17,9]*
Unité urbaine de Paris	19,3	[18,9 ; 19,7]	*19,4*	*[19,0 ; 19,8]*	18,4	[18,1 ; 18,7]	*18,5*	*[18,2 ; 18,8]*

Type de ménage :

	Taux brut (%)				Taux net (%)			
	Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*		Avec les méthodes retenues		*Avec hot-deck pour revenus individuels, dépenses en eau et énergies*	
	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**	Taux	Intervalle de confiance*	*Taux*	*Intervalle de confiance**
Personne seule	25,1	[24,7 ; 25,5]	25,4	[25,0 ; 25,8]	22,8	[22,4 ; 23,2]	23,0	[22,6 ; 23,4]
Famille monoparentale	26,5	[25,8 ; 27,2]	26,7	[26,0 ; 27,4]	20,8	[20,3 ; 21,3]	20,9	[20,4 ; 21,4]
Couple	12,6	[12,3 ; 12,9]	12,7	[12,4 ; 13,0]	12,3	[12,0 ; 12,6]	12,4	[12,1 ; 12,7]
Famille	17,6	[17,3 ; 17,9]	17,6	[17,3 ; 17,9]	16,7	[16,5 ; 16,7]	16,7	[16,5 ; 16,9]
Autre	15,9	[15,2 ; 16,6]	16,0	[15,3 ; 16,7]	14,5	[13,8 ; 15,2]	14,6	[13,9 ; 15,3]

* Les intervalles de confiance sont calculés sous l'hypothèse d'un sondage aléatoire simple (cf. annexe 16).

Annexe 16 : Calcul de la précision d'un ratio : le taux d'effort

Soit N le nombre de ménages de la population et n la taille de l'échantillon. Le taux d'effort peut s'écrire comme le ratio des totaux des dépenses et des revenus des ménages de la population :

$$TXEFF = \frac{\sum_{i=1}^{N} dépenses_i}{\sum_{i=1}^{N} revenus_i}$$

On l'estime par le ratio des totaux dans l'échantillon[69], en estimant séparément le numérateur et le dénominateur :

$$txeff = \frac{\sum_{i=1}^{n} dépenses_i}{\sum_{i=1}^{n} revenus_i}$$

Pour estimer sa variance, on utilise la technique de linéarisation appliquée à un ratio70.

On crée une variable artificielle U définie pour un ménage i par :

$$u_i = \frac{1}{\frac{N}{n}\sum_{i=1}^{n} revenus_i}\left(dépenses_i - txeff \times revenus_i\right)$$

Sous l'hypothèse d'un sondage aléatoire simple sans remise, une estimation de la variance du taux d'effort est :

$$\hat{V}(txeff) = (1-f)\frac{N^2}{n}\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2 = (1-f)\frac{N^2}{n}s_u^2$$

où s_u^2 est la variance empirique de la variable U sur l'échantillon et f = n/N le taux de sondage.

L'intervalle de confiance du taux d'effort de niveau 0,95 est alors :

$$IC = \left[txeff - 1{,}96\sqrt{\hat{V}(txeff)}\,;\, txeff + 1{,}96\sqrt{\hat{V}(txeff)}\right]$$

[69] Il n'est pas sans biais mais lorsque la taille de l'échantillon est suffisamment grande, le biais est négligeable.
[70] Voir Caron, Ravalet et Sautory (1996) pour plus de détails.