Méthodologie statistique

M2016/08

Exploitation de l'enquête expérimentale Logement internet/papier

Tiaray Razafindranovona (DMCSI)

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2016/08

Exploitation de l'enquête expérimentale Logement internet/papier

Tiaray Razafindranovona (DMCSI)

Ce document regroupe différents travaux autour de l'expérimentation Logement internet/papier et de sa comparaison avec l'enquête nationale Logement. Je remercie Laure Crusson, Céline Arnold, Erwan Pouliquen et Catherine Rougerie pour leurs contributions à l'exploitation de cette expérimentation.

Je remercie également Gaël de Peretti, Olivier Sautory et Sylvie Lagarde pour leur relecture attentive et leurs commentaires constructifs. Je reste seul responsable des erreurs ou omissions qui pourraient rester.

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101
18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 33 - Fax : 33 (1) 41 17 61 97 - CEDEX - E-mail :-DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Exploitation de l'enquête expérimentale Logement internet/papier

Tiaray Razafindranovona *

Résumé

Afin d'accumuler de la connaissance et de bien cerner les difficultés méthodologiques que pose la collecte par Internet ou multimode auprès des ménages, l'Insee a programmé sur plusieurs années un plan d'expérimentations. L'enquête expérimentale web/papier Logement est l'une de ces expérimentations, réalisée en 2014, en parallèle de l'enquête nationale Logement 2013 (ENL). L'objectif principal de cette expérimation est la comparaison des estimations de montant de loyer payé par les ménages. Après traitement de la non-réponse et calage sur les totaux de la base de sondage, le loyer mensuel moyen estimé à partir de l'expérimentation est supérieur d'environ 30 euros à celui de l'ENL. Après contrôle de la sélection sur caractéristiques observables, l'écart est de 25 euros. L'écart de loyer peut s'expliquer par un effet de mode lié à la mesure mais peut aussi relever de la sélection non contrôlée.
Outre cette question centrale, différents points sont abordés dans ce document comme le déroulement de la collecte, la qualité des données recueillies et le recours aux documents externes.

Mots-clés : enquête multimode, collecte par internet, effets de mode, effets de mesure, effets de sélection, logement, loyer, méthodologie d'enquête, calage, score de propension, arrondi.

Abstract

Insee decided to launch a series of experimental households surveys to increase knowledge in the field of internet data collection and to identify methodological issues. One of these experimental surveys (web and paper), entitled "Expérimentation Logement (Housing Experiment)" took place in 2014, in parallel of the face-to-survey "Enquête Nationale Logement (Housing National Survey)". The major aim of the experiment is the comparison of rentals paid by households. After non-response correction and calibration, the monthly average rental is 30 euros higher in the experiment than in the face-to-face survey. After controlling the selection on observables, the difference is about 25 euros. These differences are due to measurement effects or also possibly to uncontrolled selection.
Besides this main question, this document tackles subjects such as progress of data collection, quality of collected data and use of external documents.

Keywords : mixed-mode survey, web data collection, mode effects, measurement effects, selection effects, housing, rent, survey methodology, calibration, propensity score, rounding.

* INSEE, Département des méthodes statistiques, tiaray.razafindranovona@insee.fr

Table des matières

Introduction

Les enquêtes nationales Logement constituent la source privilégiée pour décrire, au niveau national, le parc de logements et les conditions d'occupation de ces logements par les ménages. Ces enquêtes, généralistes mais très détaillées, permettent, par exemple, de fournir des éléments structurels sur les données financières telles que le prix des logements ou les montants de loyer payés par les ménages qui occupent leur logement en tant que locataires. Un suivi régulier des niveaux de loyer est néanmoins difficile à réaliser à partir de ces enquêtes : ainsi, la dernière édition a eu lieu en 2013 et la précédente en 2006. Un besoin peut ainsi se faire ressentir de disposer de données alimentées plus régulièrement sur les montants de loyer pour éclairer le débat public. En complément du dispositif Loyers et Charges qui porte sur l'évolution conjoncturelle des loyers à un niveau agrégé et à qualité constante, des données régulières sur les montants de loyers apporteraient un éclairage sur les évolutions des loyers selon les caractéristiques des ménages et des logements. L'autre enjeu serait de disposer de données plus finement localisées.
En l'état actuel, les enquêtes nationales Logement peuvent difficilement répondre à ces défis. Pourrait-on imaginer des dispositifs plus légers et ciblés sur les montants de loyer ? Une option envisageable, car moins lourde et moins coûteuse, serait de réaliser une courte enquête auto-administrée par internet ou par papier qui interrogerait les occupants de leur logement principalement sur les montants de loyer. Une telle opération a été mise en oeuvre dans le cadre d'une expérimentation en 2014. Quelle est la qualité de l'information recueillie en l'absence d'enquêteur alors que les concepts en jeu sont plutôt complexes et nécessitent un effort conséquent du répondant ? Et dans quelle mesure les estimations de loyer à partir d'une telle opération sont proches des estimations de référence, produites à partir des enquêtes nationales Logement ? L'exploitation de l'enquête expérimentale Logement web/papier va nous fournir des éléments de réponse quant à la pertinence de ce type de dispositif ciblé sur les montants de loyer.

Chapitre 1

Présentation de l'expérimentation et objectifs

L'expérimentation Logement web/papier est une enquête auto-administrée dont le questionnement est largement inspiré de l'enquête nationale Logement 2013. Cette expérimentation vise principalement à collecter de l'information sur les montants de loyer payés par les ménages qui occupent leur logement en tant que locataires. Elle fait partie du programme d'expérimentations de l'Insee qui vise à accumuler de la connaissance sur les enquêtes ménages réalisées via des protocoles qui offrent la possibilité de répondre par internet.

1.1 Les enquêtes nationales Logement (ENL)

L'enquête Logement est l'une des principales enquêtes de l'Insee par son ancienneté (1955) et la taille de son échantillon (54 000 logements en 2013). Cette enquête est réalisée périodiquement par l'Insee : la dernière édition a été réalisée en 2013 et la précédente en 2006. C'est la source statistique majeure pour décrire le parc de logements et les conditions d'occupation par les ménages de leur résidence principale. La passation du questionnaire s'effectue en face à face par un enquêteur, en une seule visite, par collecte assistée par ordinateur. Elle s'adresse à l'occupant en titre du logement ou son conjoint éventuel.

Les recensements ont bien sûr l'avantage de couvrir un échantillon plus important mais ne permettent pas de connaître les loyers, les charges, les plans de financement, les revenus, et bien des dimensions de la qualité de l'habitat des Français (en particulier des plus mal logés) qui sont au contraire abordées en détail par l'enquête Logement.

Ses usages sont multiples : données de cadrage structurelles, étude de sous-populations fines et modélisation des comportements, analyses semi-conjoncturelles ou en pseudo-panels basées sur des comparaisons chronologiques entre enquêtes successives, etc.

1.2 L'ENL

Suite aux travaux du Conseil national de l'information statistique sur le mal-logement, l'ENL s'intéresse plus particulièrement aux épisodes sans domicile personnel dans le passé, aux statuts d'occupation individuels des occupants du logement et à l'hébergement chez un tiers faute de logement personnel. L'enquête est menée auprès d'un échantillon de 43 000 logements en métropole et 11 250

logements dans les DOM.

1.3 Le plan d'expérimentations d'enquêtes multimode à l'Insee

L'utilisation d'internet comme mode de recueil des données privilégié ou complémentaire est une solution envisagée à plus ou moins long terme par les Instituts nationaux de statistique pour répondre à la demande toujours plus exigeante en termes de qualité et de diversité des enquêtes auprès des ménages, dans un contexte général de restriction budgétaire (de Peretti et Razafindranovona, 2014). En particulier, le recours à la collecte multimode pourrait apparaître comme la solution théorique à mettre en oeuvre que l'on se place dans le cadre de l'erreur d'enquête totale (Groves et Lyberg, 2010) ou dans celui, plus général, de la qualité totale (Lyberg, 2012).
Cependant, si la collecte par Internet est un mode peu coûteux, elle pose des problèmes méthodologiques non négligeables : couverture, auto-sélection ou biais de sélection, non-réponse et les difficultés de sa correction, *satisficing*, etc. Aussi, avant de développer ou généraliser l'utilisation du multimode, l'Insee s'est lancé dans une vaste opération d'expérimentations afin d'étudier ces différentes questions méthodologiques. En particulier, la nécessité d'expérimenter spécifiquement pour chaque type d'enquête auprès des ménages s'est imposée car les résultats de la littérature ne sont pas toujours facilement généralisables.

1.3.1 Enquêtes expérimentales de l'Insee

Une première expérimentation de type multimode incluant une collecte par internet a été réalisée en 2010 et portait sur le thème du logement (Amiel et Denoyelle, 2012). Depuis 2013, plusieurs expérimentations de collecte multimode ont été réalisées ou sont programmées : Qualité de la vie au travail (2013), Vols, violences et sécurité (2013), Patrimoine (2015), etc.

1.4 L'enquête expérimentale Logement

1.4.1 Protocole de l'enquête expérimentale

Le protocole prévu initialement est le protocole « standard » utilisé pour les enquêtes expérimentales auto-administrées à l'Insee :

— envoi d'une lettre-avis avec les données de connexion à la version en ligne du questionnaire ;
— 1ère relance (3 semaines après) avec envoi d'une lettre contenant les données de connexion, un questionnaire papier ainsi qu'une enveloppe pré-affranchie ;
— 2ème relance (3 semaines après) avec les données de connexion.

La possibilité de répondre sur papier est offerte afin de ne pas exclure d'emblée de l'enquête une partie de la population, d'autant plus quand la réponse est obligatoire, comme dans l'enquête principale. En revanche, cette possibilité n'est offerte que dans un deuxième temps afin de favoriser la réponse par internet sans que cela n'ait d'incidence négative sur le taux de réponse global.

1.4.2 Échantillonnage

La base de sondage est constituée à partir de fichiers fiscaux de 2013 : les logements retenus sont les résidences principales (au sens fiscal) occupées par des locataires en France métropolitaine. 40 000 logements sont ainsi tirés dans cette base de sondage. Comme c'est le cas pour l'enquête en face à face, deux extensions régionales ont été tirées en Île-de-France et en Nord-Pas-de-Calais. La

base de tirage a été triée selon le code commune, la date d'achèvement et le type de logement.

1.4.3 Objectif

L'objectif principal de l'expérimentation est de voir dans quelle mesure une enquête, *a priori* légère et auto-administrée, permet de recueillir de l'information de qualité sur les loyers. Alors qu'en face à face, l'enquêteur peut demander au répondant de se munir de documents (quittances de loyer, relevés de charges, baux, etc.), il n'y a pas de motivation extrinsèque, en auto-administré, à faire l'effort de recherche pour reporter l'information pertinente. Par ailleurs, alors que l'enquêteur peut jouer un certain rôle pour faciliter la compréhension de concepts assez complexes et pas forcément explicites, en auto-administré, l'enquêté doit s'auto-motiver à bien comprendre ce qui lui est demandé. Il s'agit ici, en effet, d'un questionnement qui demande à la fois un effort de recherche ainsi qu'un effort cognitif important pour parvenir à répondre correctement aux questions posées. L'exploitation de l'expérimentation devra ainsi évaluer la qualité de l'information recueillie et, à cette fin, les estimations issues de l'expérimentation seront comparées avec celles produites à partir de l'ENL.

Chapitre 2

Bilan de la collecte et premières exploitations brutes de l'enquête expérimentale Logement

Cette partie décrit, dans un premier temps, le déroulement de la collecte. Elle mentionne les difficultés pratiques rencontrées qui ont conduit à une modification de protocole expliquant par ailleurs la répartition des réponses collectées entre les différents modes. De premières exploitations brutes autour du questionnement sont ensuite présentées avant d'analyser les réponses collectées en termes de qualité.

2.1 Déroulement de la collecte et taux de réponse

2.1.1 Déroulement de la collecte

Le calendrier de l'enquête initialement prévu a connu quelques décalages.
Suite au report de la fin de collecte de l'enquête en face à face, la période de collecte de l'expérimentation Logement a été repoussée de quelques mois. Cela afin d'éviter, dans la mesure du possible [1], que des personnes contactées par des enquêteurs pour l'enquête en face à face soient simultanément contactées pour l'expérimentation internet-papier. Ce report de quelques mois a également permis d'utiliser un millésime plus récent des fichiers fiscaux pour la base de sondage, ce qui a toute son importance, en particulier car la population cible (les locataires) est plutôt mobile.
Par ailleurs des contraintes opérationnelles lors de la collecte de l'expérimentation ont également eu des conséquences sur le calendrier : suite à des difficultés rencontrées lors de l'envoi des courriers de 1ère relance [2], le choix a été fait de repousser la fin de la collecte d'environ un mois.
Ainsi la collecte s'est finalement déroulée entre mars et juin 2014.

2.1.2 Taux de réponse

Environ 12 000 unités parmi les 40 000 enquêtées ont répondu au questionnaire. La répartition entre les deux modes de collecte proposés est plutôt en faveur du questionnaire papier : 7 000 questionnaires ont été retournés par voie postale et 5 000 questionnaires remplis sur internet. Cette part

1. Un effort tout particulier est réalisé pour assurer une disjonction entre les enquêtes, mais celle-ci ne peut être totale car les bases de sondage utilisées diffèrent.
2. Problèmes d'impression et d'envoi de certains courriers.

un peu plus importante du papier en comparaison d'autres expérimentations de l'Insee peut s'expliquer en partie par une modification du protocole initialement prévu : pour remédier aux problèmes d'impression des questionnaires envoyés à la première relance, le questionnaire papier a également été envoyé à la deuxième relance. Ainsi le nombre de questionnaires papier envoyés aux enquêtés a été beaucoup plus important que dans les précédentes enquêtes expérimentales de l'Insee.

2.1.3 Collecte des réponses web

La distribution des réponses validées sur internet est fournie dans le graphique 2.1. Les pics correspondent à la réception des courriers : lettre-avis puis première et deuxième relances éventuelles. Environ la moitié des réponses arrivent avant la réception de la première relance, et l'autre moitié de réponses qui arrivent ultérieurement se répartissent de manière relativement équilibrée sur les deux dernières périodes délimitées par les relances.

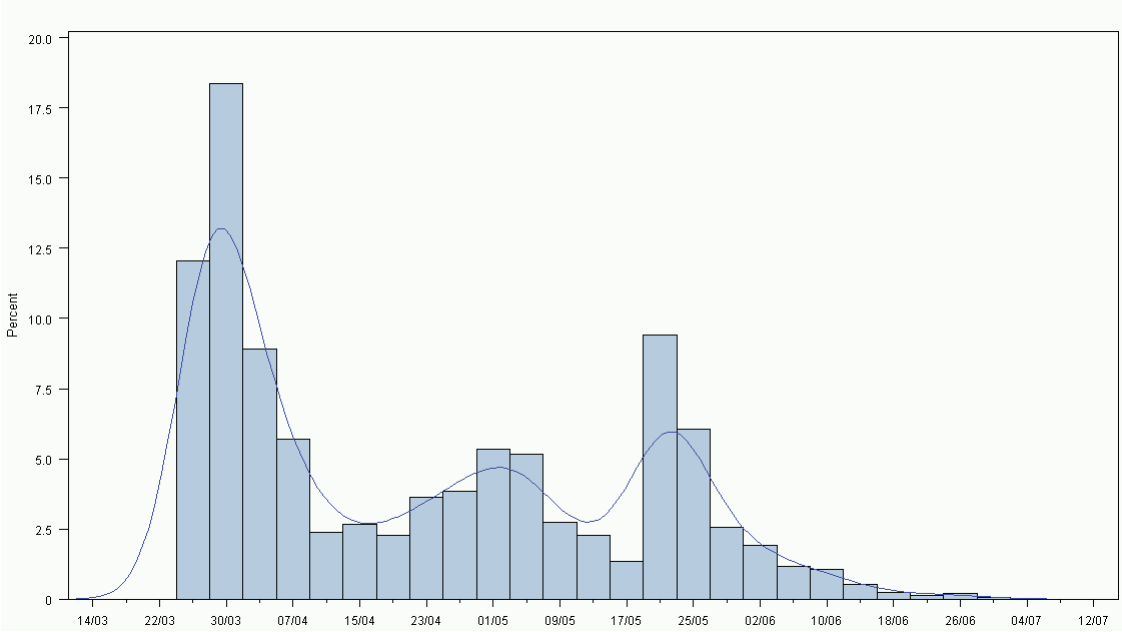

FIGURE 2.1 – Réponses web selon la date de validation

2.1.4 Familiarité des répondants avec l'outil internet

Il n'est guère surprenant d'observer que les personnes qui répondent par internet soient beaucoup plus familières avec ce moyen de communication que celles qui répondent en retournant le questionnaire papier par voie postale. En effet, 84 % des répondants web utilisent internet tous les jours ou presque, une proportion plus de deux fois supérieure à celle des répondants papier et également bien plus importante que dans l'ensemble de la population où cette proportion est de 62 % d'après les résultats de l'enquête Technologies de l'information et de la communication 2014 (voir table 2.1). Un résultat important est que près d'un tiers des répondants sur papier n'ont pas utilisé internet au cours des trois derniers mois : ces personnes n'auraient sans doute pas participé si inter-

net était le seul mode de réponse possible [3].

	Répondants web	Répondants papier	Référence TIC 2014
Utilisation d'internet (3 derniers mois, %)			
Tous les jours ou presque	84,1	42,9	62,4
Au moins une fois par semaine	9,5	14,7	11,4
Moins d'une fois par semaine	2,3	7,0	3,6
Jamais	3,9	35,4	22,7

TABLE 2.1 – Utilisation d'internet

2.1.5 Ressenti des répondants sur le questionnement

Quelques questions posées en fin d'interrogation dans la version web du questionnaire portent sur le ressenti des répondants en ce qui concerne la longueur et la difficulté du questionnaire. 80 % des répondants jugent le questionnaire comme étant de longueur raisonnable alors qu'ils sont 18 % à le trouver trop long et 2 % qui, au contraire, le trouvent trop court. Pour 51 % des répondants, les questions sont faciles à comprendre et pour 45 % des répondants, les questions sont jugées "normales" ; la modalité "difficiles à comprendre" est quant à elle choisie par 4 % des répondants. Pour comparaison, dans l'enquête expérimentale Vols, violences et sécurité (VVS), cette même question était posée : pour les deux tiers des répondants, les questions étaient jugées faciles, l'autre tiers les considérant "normales" et moins d'1 % les trouvaient difficiles. Ceci semble indiquer que le questionnaire de l'expérimentation Logement est jugé comme un peu plus difficile que celui de VVS, ce qui n'est guère étonnant au vu des concepts en jeu.

2.1.6 Sur les commentaires libres

La zone de commentaire libre est par ailleurs utilisée par un nombre non négligeable de répondants : 21 % des répondants (soit 2 500 zones de commentaire) écrivent au moins un caractère dans cette zone. La proportion est un peu plus élevée pour les répondants sur internet (22 %) que pour les répondants papier (20 %). Pour certains d'entre eux, le commentaire précise qu'il n'y a rien à signaler : l'acronyme RAS ou sa version développée est utilisée dans une centaine de commentaires.
Sans faire une analyse textuelle approfondie de ces commentaires, nous pouvons relever les mots les plus fréquemment cités. Les mots autour du questionnement, du protocole de l'enquête reviennent assez souvent : questions (291 occurences), internet (219), questionnaire (198), répondre et ses déclinaisons (243), réponse (98), ou encore enquête (104) et courrier (45). Divers mots autour de la thématique de l'enquête sont utilisés : charges (260), logement (223), loyer (184), construction (74), chauffage (49) ou encore immeuble (67), maison (43), appartement (40). Notons pour finir que l'adjectif indiscret est cité dans une cinquantaine de commentaires.

2.2 Recours aux documents externes

L'un des objectifs de l'expérimentation est d'apprécier le recours aux documents en l'absence d'enquêteur pour motiver ce recours. Ainsi, plusieurs questions sont posées à cette fin lors de l'in-

3. Néanmoins, nous pouvons remarquer que 4 % des répondants web sont des personnes qui n'utilisent jamais internet. Un certain nombre de ces personnes indiquent dans la zone de commentaire libre qu'une personne de leur entourage a été sollicitée pour aider à remplir le questionnaire.

terrogation. Environ 50 % des répondants qui sont locataires de leur logement déclarent avoir eu recours à leur quittance pour répondre au questionnaire. Ils sont environ 15 % à déclarer avoir recours à leur bail et sont 53 % à avoir recours à l'un ou l'autre de ces documents. Le recours aux documents externes est légèrement plus fréquent chez les répondants papier que chez ceux qui répondent sur internet : sur papier, 52 % ont recours à la quittance, 16 % au bail et 55 % à l'un de ces deux documents contre respectivement 46 %, 13 % et 51 %. Ces proportions semblent légèrement moindres que dans l'ENL où la proportion du recours à l'un de ces documents est de 62 % pour les locataires de leur logement lorsque cette information est collectée : la comparaison doit néanmoins être effectuée avec précaution car pour un nombre non négligable de logements (environ 40 %) l'information n'est pas collectée dans l'ENL. Par ailleurs, alors que dans l'expérimentation c'est le répondant qui répond directement à la question du recours au document, dans l'ENL, c'est l'enquêteur qui collecte lui-même l'information.

Enfin, même si ce n'est *a priori* pas prévu par le protocole, quelques répondants envoient à l'Insee des documents pour appuyer, justifier ou préciser leurs réponses au questionnaire. En particulier une quinzaine de personnes joignent à leur courrier une quittance de loyer ou un avis d'échéance et une dizaine de personnes font parvenir un document de régularisation de charges.

2.3 Non-réponse partielle, qualité de réponse

La qualité de réponse est généralement considérée comme meilleure sur internet que sur papier en raison des divers contrôles en ligne qui peuvent être implémentés. En particulier, la non-réponse partielle est beaucoup plus faible car il est souvent plus aisé pour le répondant de sauter une question sur papier que sur un questionnaire en ligne.

2.3.1 Non-réponse partielle

Alors que sur papier, hormis le fait de proposer un questionnaire le plus clair et lisible possible, il n'y a guère de moyens de motiver l'enquêté à répondre à une question, sur internet plusieurs options existent. Des contrôles bloquants peuvent être implémentés afin de forcer la réponse pour les questions les plus importantes : ils doivent néanmoins être utilisés avec parcimonie car le répondant peut choisir d'abandonner définitivement si ces sollicitations l'exaspèrent. Pour limiter la non-réponse partielle, une possibilité est de ne proposer la modalité de non-réponse que dans un deuxième temps, lorsque l'enquêté tente de sauter la question. C'est ce choix qui a été fait pour le questionnaire en ligne de l'expérimentation Logement.

Ainsi, lorsque l'on observe les questions posées à l'ensemble des individus (donc non filtrées), la non-réponse partielle est quasiment inexistante sur les questionnaires en ligne alors qu'une proportion non négligeable de personnes sautent les questions sur papier. Ainsi, sur internet, la non-réponse est quasi-nulle pour les différentes questions sur les caractéristiques sociodémographiques (voir table 2.2). Elle varie en revanche de 0,6 % à 8,0 % pour ces mêmes questions dans la version papier. Sur la version web du questionnaire, l'une des seules questions, communes à tous les répondants, pour laquelle la non-réponse n'est pas négligeable est celle qui porte sur la date d'achèvement du logement. La non-réponse à cette question est néanmoins beaucoup moins élevée sur le web (3,6 %) que sur papier (21,1 %) [4].

4. Cette question est aussi posée dans les enquêtes annuelles de recensement (EAR) et fait partie des questions où le taux de non-réponse partielle est le plus fort : en moyenne 13,7 % sur la période 2009-2014 (questionnaire auto-administré papier). Depuis 2015, il est possible de répondre aussi par internet. Là aussi, les comportements de non-réponse sur cette question sont très différents : 15,1 % sur papier et 0,6 % par internet.

	NR Logement web (%)	NR Logement papier (%)
Caractéristiques sociodémographiques		
Année de naissance	0,1	4,8
Sexe	0,0	0,6
En couple	0,1	2,3
Diplôme	0,1	5,3
Situation vis-à-vis du travail	0,0	3,6
Usage internet	0,1	8,0
Logement		
Résidence principale/secondaire	0,0	4,8
Type de logement	0,2	3,3
Année d'entrée dans le logement	0,0	9,0
Date d'achèvement	3,6	21,1
Transfert courrier	0,1	10,5

TABLE 2.2 – Non-réponse partielle dans les questionnaires web et papier de Logement

2.3.2 Qualité des réponses

Respect des filtres, cohérence

Dans un questionnaire en ligne, le répondant est automatiquement guidé : les filtres des questions sont quasiment indolores pour le répondant. Dans le cas d'un questionnaire papier, les filtres peuvent être problématiques car ils nécessitent une attention particulière du répondant pour correctement naviguer au sein du questionnaire.

Un exemple est emblématique dans le questionnaire de l'expérimentation Logement : il s'agit de la question sur la réception du courrier à l'adresse actuelle de l'enquêté, posée pour des raisons de périmètre. Si la personne répond négativement à cette question, il est demandé de fournir le code postal de son adresse actuelle. Dans le cas contraire, il n'a pas à répondre à cette dernière question.

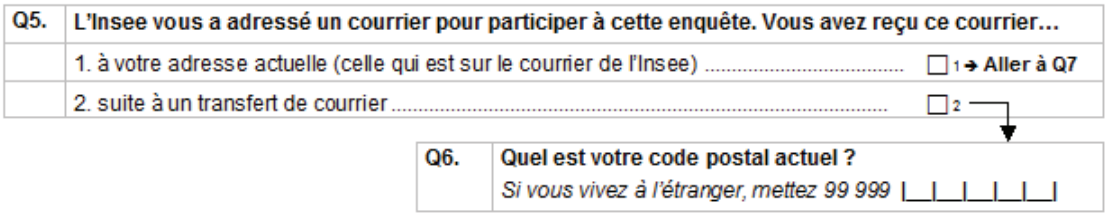

FIGURE 2.2 – Extrait du questionnaire : code postal

Pourtant, dans la version papier du questionnaire, un grand nombre de personnes prend la peine de répondre à cette question alors que le filtre indique que cette question doit être passée : près de 50 % des personnes pour lesquelles la question n'est pas demandée fournissent quand même un code postal. Cet exemple n'est pas en soi problématique : à la marge, cela modifie l'expérience utilisateur

de passation par rapport à ce qui est prévu initialement par les concepteurs de l'enquête.

Un autre exemple est celui d'un enchaînement de 2 questions sur le type de loyer. Dans la version papier du questionnaire, environ 40 % des personnes qui répondent que leur loyer relève de la législation HLM choisissent l'une des modalités de la question suivante alors que cette dernière n'a pas à leur être posée [5]. Ces répondants vont alors se positionner alors que ce n'est pas souhaitable et peut, automatiquement, par construction, soulever des contradictions dans les réponses.

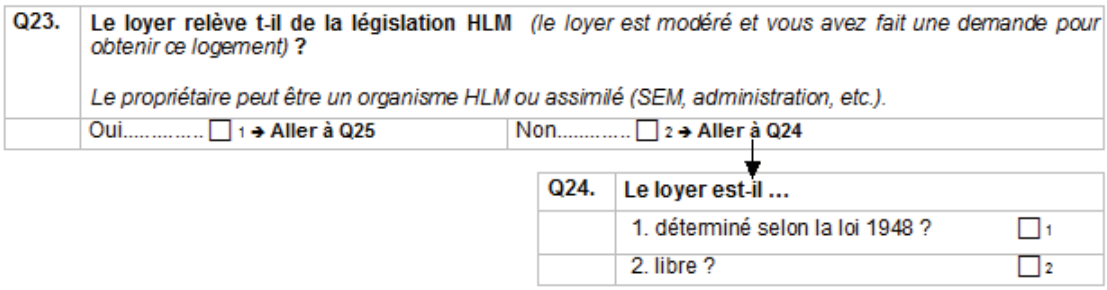

FIGURE 2.3 – Extrait du questionnaire : type de loyer

Un dernier exemple que l'on peut fournir porte sur le nombre d'habitants du logement. Ainsi, dans le questionnaire, outre le nombre total d'habitants du logement, le répondant doit aussi donner le nombre d'habitants selon des tranches d'âge : entre 0 et 13 ans, entre 14 et 17 ans et 18 ans ou plus. Si sur internet les chiffres sont tous cohérents en raison des contrôles implémentés, pour 40 % des questionnaires papier, le nombre d'habitants ne correspond pas à la somme des habitants déclinés par tranche d'âge.

Champs numériques

Plusieurs questions de l'expérimentation demandent à l'enquêté de fournir une valeur numérique : il s'agit de tous les montants financiers (loyers, charges, etc.) mais aussi des données relatives aux caractéristiques du logement comme le nombre de pièces ou encore la surface (demandée en clair dans un premier temps dans la version web).

Là aussi, la qualité des questionnaires remplis sur internet est supérieure à celle des questionnaires sur papier (table 2.3) : la non-réponse partielle y est moindre et les contrôles embarqués contribuent à limiter les erreurs grossières.

Ainsi, sur la question de nombre de pièces, la non-réponse est de 13,6 % sur papier alors qu'elle est inférieure à 1 % sur internet.
L'écart est relativement moindre pour la surface du logement demandée en clair. Par ailleurs, sur internet les personnes qui ne donnent pas la surface en clair doivent choisir une tranche de surface ; ainsi, une information est collectée pour la totalité des répondants web sur cette question.
Enfin, le montant de loyer mensuel est toujours collecté sur internet car la question est bloquante ;

5. 30 % déclarent que leur loyer relève de la loi de 1948 et 10 % que leur loyer est libre.

	NR Logement web (%)	NR Logement papier (%)
Nombre de pièces dans le logement	0,3	13,6
Surface du logement (en clair)	2,5	9,9
Montant de loyer	0,0	6,1

TABLE 2.3 – Non-réponse partielle pour quelques champs numériques demandés

sur papier, la non-réponse s'élève à plus de 5 % pour cette variable. De plus, des valeurs probablement aberrantes car trop élevées sont en plus grand nombre dans les réponses papier : le 99ème percentile de distribution se situe à 35 000 euros (contre 2 200 euros sur internet) et environ 1,5 % des valeurs sont supérieures à 10 000 euros (contre 0,2 % sur internet).

2.3.3 Les questionnaires mal imprimés et remplis ?

Un certain nombre de questionnaires ont été envoyés avec un défaut d'impression (voir plus haut) et un peu plus de 300 de ces questionnaires ont été remplis, provenant très majoritairement (à plus de 90 %) de la région Rhône-Alpes [6]. Une question qui se pose est celle de la qualité de ces questionnaires : peuvent-ils être intégrés lors de la phase d'exploitation ? Des éléments de réponse à ces questions peuvent être donnés en comparant certains des indicateurs de qualité présentés précédemment avec ceux des questionnaires papier au bon format d'impression.

La non-réponse partielle est globalement du même ordre que celle des questionnaires au bon format pour les questions posées à l'ensemble des enquêtés (table 2.4). Une comparaison déclinée par question montre que la non-réponse partielle est globalement plus élevée pour les questionnaires mal imprimés.

	NR Logement "mal imprimés" (%)	NR Logement papier (%)
Caractéristiques sociodémographiques		
Année de naissance	5,5	4,8
Sexe	0,7	0,6
En couple	1,9	2,3
Diplôme	4,8	5,3
Situation vis-à-vis du travail	4,5	3,6
Usage internet	12,6	8,0
Logement		
Résidence principale/secondaire	5,5	4,8
Type de logement	2,9	3,3
Année d'entrée dans le logement	9,4	9,0
Date d'achèvement	22,3	21,1
Transfert courrier	11,3	10,5

TABLE 2.4 – Non-réponse partielle selon le format d'impression

6. Il n'est pas possible de connaître avec certitude la répartition régionale des questionnaires envoyés avec un défaut d'impression. Néanmoins, au vu des retours de questionnaires, la région Rhône-Alpes semble particulièrement touchée avec 287 questionnaires collectés présentant un défaut d'impression ce qui représente 44 % des retours de questionnaires papier pour cette région. L'autre région touchée est le Limousin : si les volumes en jeu sont plus faible (20 questionnaires), cela représente tout de même 32 % des retours papier pour cette région.

La tendance est globalement du même ordre pour les champs numériques (table 2.5) : la non-réponse est un peu plus forte pour les questionnaires mal imprimés.

	NR Logement "mal imprimés" (%)	NR Logement papier (%)
Nombre de pièces dans le logement	15,0	13,6
Surface du logement (en clair)	10,1	9,9
Montant de loyer	6,3	6,1

TABLE 2.5 – Non-réponse partielle (champs numériques) selon le format d'impression

Si l'on reprend les exemples des filtres non suivis (voir plus haut), on constate que cette tendance est moins marquée pour les questionnaires mal imprimés : le code postal est rempli quand il ne devrait pas l'être dans 45 % des cas (contre 50 % pour les autres questionnaires papier) et une réponse à la question sur le type de loyer est donnée alors qu'elle n'est pas demandée dans 30 % des cas (contre 40 % pour les autres questionnaires papier). Ce type de résultats semble indiquer que si l'expérience utilisateur est globalement plus difficile pour les enquêtés qui répondent à ces questionnaires mal imprimés, cette population est en quelque sorte sélectionnée : ceux qui font l'effort de répondre font preuve d'une certaine attention lorsqu'ils remplissent le questionnaire.

Ces différents éléments semblent montrer qu'il n'y a ainsi pas trop de raisons d'écarter définitivement ces questionnaires malgré le défaut d'impression qui aurait pourtant pu être rédhibitoire.

2.3.4 Premières exploitations brutes de la variable de loyer

L'objectif principal de l'expérimentation Logement est de comparer les estimateurs de loyer issus de cette enquête avec ceux de l'enquête nationale Logement. Avant toute phase de corrections et de redressements, il peut être intéressant d'observer quelle est la nature et la qualité de l'information brute récupérée.

Sur les 12 130 questionnaires récupérés, le répondant reporte un montant de loyer dans 10 282 cas. Rappelons qu'il existe plusieurs cas où il est tout à fait normal que le loyer ne soit pas reporté, par exemple si le ménage n'est pas locataire mais propriétaire du logement qu'il occupe.

Sur les 9 900 enquêtés pour lesquels il est demandé de reporter le dernier montant de loyer payé [7], une réponse est fournie dans 9 570 cas [8], soit une non-réponse partielle à cette question de l'ordre de 3,3 %. La non-réponse partielle est exclusive aux questionnaires papier puisqu'il n'est en théorie pas possible de passer cette question qui fait l'objet d'un contrôle bloquant sur internet. Ainsi la non-réponse partielle à cette question est donc nulle sur internet et de 6,1 % sur papier.

Regardons de plus près comment se distribuent ces loyers (table 2.6) [9] lorsque leur périodicité est mensuelle (environ 99 % des cas [10]). Avant corrections, ce n'est pas vraiment pertinent de regarder la

7. Ici, nous n'effectuons pas d'autres restrictions de champ comme nous allons devoir le faire lorsque nous allons comparer les loyers des deux enquêtes.

8. Cela fait près de 700 questionnaires en moins que les 10 282 évoqués plus haut : cette différence correspond à un certain nombre de questionnaires pour lesquels un montant de loyer est reporté alors qu'il n'aurait pas dû l'être en raison du statut d'occupation déclaré.

9. Même si ici l'objectif n'est pas vraiment la représentativité, on pondère quand même pour ce tableau par les pondérations issues du tirage.

10. Pour information, sur un peu plus de 10 000 loyers déclarés, il y a 70 cas de loyers dont la périodicité est trimestrielle et 30 cas où elle est annuelle ; cela est très marginal mais moins que dans l'ENL où pour un nombre d'observations à peu

moyenne, mais pour information, elle est à 987 euros.

	Tous	Internet	Papier
Maximum	354 407	64 220	354 407
P99	5 100	2 012	35 935
P95	1 100	1 160	1 000
P90	867	925	800
P75	650	700	603
P50	491	530	455
P25	373	403	347
P10	296	322	275
P5	252	284	235
P1	162	200	138
Minimum	1	1	3

TABLE 2.6 – Distribution des loyers bruts mensuels

Un apurement sera nécessaire pour corriger certaines valeurs : en particulier sur papier, la très forte valeur du 99ème percentile (comparée à celle sur internet) semble montrer que ces problèmes ne sont sans doute pas négligeables. N'oublions pas que sur papier, une erreur peut se produire lors du report de la valeur par l'enquêté mais aussi lors de la saisie des réponses. Un cas d'erreur typique est celui des décimales quand la virgule ou le point est difficile à repérer pour l'opérateur de saisie.

Comportement d'arrondi

Les invididus ne reportent pas toujours la vraie valeur du montant de loyer : celui-ci peut être arrondi de manière plus ou moins grossière. La préférence pour les chiffres ronds est traditionnellement perçue comme une forme de *satisficing* potentiellement néfaste pour la qualité des données, même si certains auteurs nuancent cela (Holbrook *et al.*, 2014).

Lorsque les mécanismes d'arrondis sont simples et uniformes dans la population, les conséquences sont limitées sur les estimations des moments d'ordre 1 mais cela peut néanmoins biaiser les estimations de la variance (Sheppard, 1897) ainsi que celles des moments d'ordre supérieur (Kendall, 1938). Lorsque les mécanismes d'arrondi sous-jacents sont plus complexes, les estimations des moyennes et des autres paramètres de la distribution peuvent être éloignées de leurs vraies valeurs. Par exemple, Dreschler et Kiesl donnent un exemple des conséquences de comportements d'arrondi [11] sur l'estimation d'un taux de pauvreté (Drechsler et Kiesl, 2015).

Le graphique 2.4 représente la distribution du loyer déclaré : la densité est estimée de manière non paramétrique avec une largeur de fenêtre d'estimation assez petite pour faire apparaître nettement les pics liés aux comportements d'arrondi en particulier ceux à la centaine et à la cinquantaine. Par exemple, entre 500 et 700 euros, on observe des grands pics de densité à 500, 600 et 700 euros indiquant que ces montants de loyer sont très fréquemment déclarés par les répondants. Des pics intermédiaires sont également particulièrement visibles à 550 et 650 euros.

près équivalent 35 loyers sont trimestriels et 5 annuels.

11. Les auteurs génèrent une distribution de revenus qui suit une loi log-normale et simulent des comportements d'arrondis à 1, 10, 100 et 1 000 euros dont les proportions dans la population sont fixées respectivement à 10 %, 40 %, 40 % et 10 %.

Ces points d'accumulation indiquent que les répondants semblent assez nombreux à arrondir d'eux-mêmes leur loyer : en effet, près de 12 % des loyers sont des multiples de 100, 19 % des multiples de 50 et 40 % des multiples de 10 (table 2.7). Le comportement d'arrondi du loyer semble un peu plus fréquent sur internet que sur papier : par exemple dans 14 % des cas, le loyer est arrondi à la centaine sur internet contre 10 % sur papier.
S'il n'est pas impossible qu'il existe également des comportements de préférence pour les chiffres ronds lorsque les montants de loyer sont fixés initialement par les propriétaires, cela ne jouerait que pour les loyers avant toute revalorisation ou indexation automatique. Cette préférence pour les chiffres ronds par les propriétaires semble exister au vu des montants de loyer de ceux ayant emménagé très récemment (ceux qui déclarent avoir emménagé en 2014 ou qui disent avoir reçu le courrier de l'Insee suite à un transfert de courrier, exclus *in fine* de l'analyse, cf 3.2.2) : 21 %, 35 % et 62 % des montants de loyer sont respectivement des multiples de 100, de 50 et de 10.
Et si, globalement, les loyers semblent plus souvent arrondis dans l'expérimentation Logement que dans l'ENL, cela doit être nuancé : ainsi, contredisant la théorie d'un *satisficing* plus important en auto-administré qu'avec intermédiation d'enquêteur, les proportions de loyers multiples de 50 et de 10 sont plus fréquentes dans l'ENL que chez les répondants papier de l'expérimentation. Rappelons par ailleurs que les taux de réponse ne sont pas du même ordre et que les personnes qui participent à l'expérimentation sont plutôt sélectionnées et potentiellement assez sérieuses dans leur comportement de réponse.

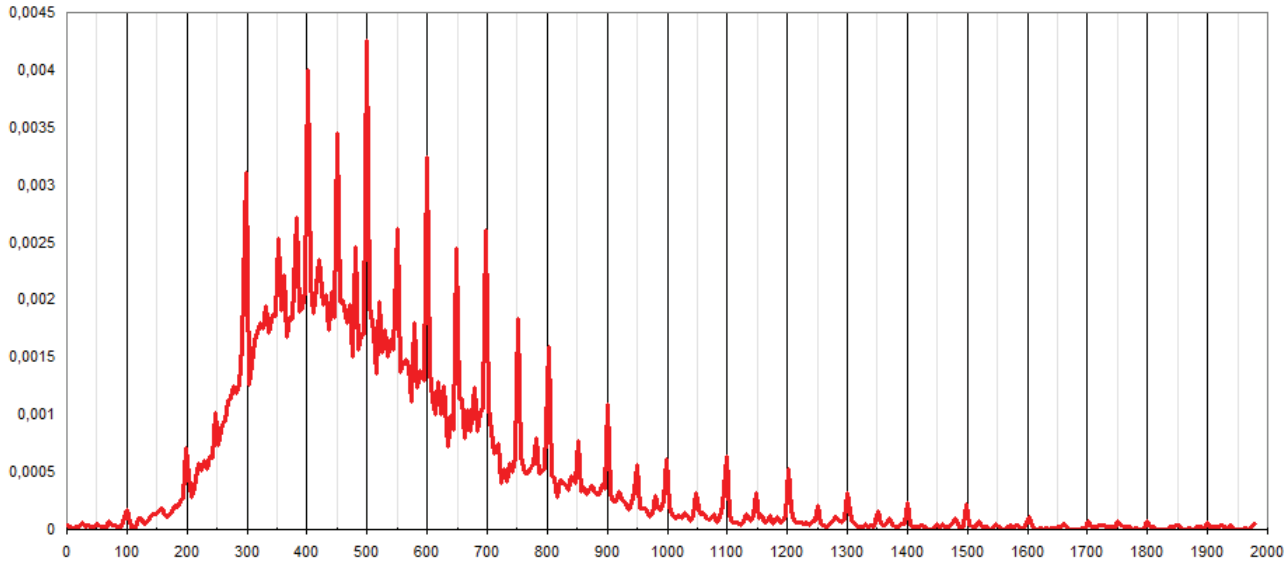

FIGURE 2.4 – Densité du loyer mensuel déclaré dans l'expérimentation Logement

	Expérimentation Internet+Papier (%)	Internet (%)	Papier (%)	*ENL (%)*
Loyer multiple de **100**	11,8	13,8	10,0	*9,8*
Loyer multiple de **50**	19,0	22,6	15,8	*16,8*
Loyer multiple de **10**	40,1	46,3	34,7	*38,0*

TABLE 2.7 – Loyers en centaines, cinquantaines et dizaines d'euros

Le recours à la quittance peut être nécessaire pour apporter des réponses aux questions sur les montants les plus précises possibles. On vérifie en effet, que concernant les réponses à la question sur le montant du loyer, l'utilisation de l'arrondi semble moins fréquente lorsque la personne déclare avoir eu recours à sa quittance (rappelons qu'environ un locataire sur deux déclare avoir eu recours à sa quittance de loyer) : 5 % de ces personnes arrondissent leur loyer à la centaine, 8 % à la cinquantaine et 21 % à la dizaine contre respectivement 16 %, 26 % et 49 % pour ceux qui disent ne pas avoir eu recours à leur quittance (table 2.8).

	Tous (%)	Internet (%)	Papier (%)
Individus déclarant avoir recours à la quittance			
Loyer multiple de **100**	4,7	5,4	4,2
Loyer multiple de **50**	7,8	9,1	6,7
Loyer multiple de **10**	22,1	25,3	19,6
Individus déclarant n'avoir pas recours à la quittance			
Loyer multiple de **100**	18,2	21,0	15,4
Loyer multiple de **50**	28,9	34,2	24,5
Loyer multiple de **10**	54,6	62,1	47,3

TABLE 2.8 – Comportement d'arrondi et recours à la quittance

Par ailleurs, un certain nombre d'individus reportent les chiffres décimaux de leur montant de loyer alors que cela n'est pas spécifiquement demandé. Cela ne concerne que les questionnaires papier car il n'était pas possible de reporter le montant de cette manière sur Internet. Ainsi environ 20 % des individus qui ont reporté un montant de loyer sur le questionnaire l'ont spontanément écrit avec les décimales.

Chapitre 3

Traitement de la non-réponse, calage, apurement et redressement du loyer

Cette partie décrit les principaux traitements post-collecte réalisés avant de produire les estimations de loyer. Ces traitements ne visent pas à rapprocher directement l'expérimentation de l'enquête de référence qu'est l'ENL. Il s'agit plutôt ici de traitements "autonomes", sans références à l'ENL (pour le traitement de la non-réponse et le calage, seules les variables de la base de sondage sont utilisées) et préparatoires à la comparaison des estimations entre expérimentation et ENL, sans contrôle particulier des différences quant à la sélection.

3.1 Traitement de la non-réponse totale

3.1.1 Caractéristiques des répondants

L'exploitation des variables de sondage nous permet d'observer la relative déformation de l'échantillon tiré initialement, suite à la phase de réponse à l'enquête. Cela sera ultérieurement utilisé lors de la correction de la non-réponse totale, mais nous pouvons d'ores et déjà relever quelques points. Comme cela est souvent constaté, la population de répondants sur internet, qu'on la compare aux répondants papier ou à l'échantillon initial, est plutôt jeune et aisée.

	Echantillon	Répondants	Web	Papier
Age (moyenne)	49,2	51,8	44,5	56,9
Revenus annuels (euros, médians)	15 900	17 900	21 000	16 300
Surface (m^2, moyenne)	63,2	65,3	65,2	65,4
Appartement (proportion)	76,0 %	75,7 %	77,1 %	74,6 %
HLM (proportion)	31,9 %	35,4 %	27,6 %	41,0 %

TABLE 3.1 – Caractéristiques des répondants

Le répondant est-il la personne à qui le courrier est adressé ?

La personne qui répond au questionnaire n'est pas forcément celle à qui le courrier est nominativement adressé. En effet, il est indiqué en préambule du questionnaire que "La personne qui doit répondre est de préférence une personne qui vit habituellement dans le logement et qui est locataire ou propriétaire en titre, c'est-à-dire celle dont le nom apparaît sur le bail ou sur le titre de propriété.

Votre conjoint peut également répondre."
Dans les faits, en comparant les réponses au questionnaire et les informations provenant de la base de sondage, la personne à qui le courrier est adressé est bien, le plus souvent, celle qui répond au questionnaire : dans 93 % des cas, âge et sexe déclarés par la personne qui répond coïncident avec les informations de la base de sondage. Ce taux de coïncidence est logiquement moins élevé lorsque la personne déclare vivre en couple (88 %) que lorsque ce n'est pas le cas (98 %).

Réponse et non réception de la lettre-avis

L'expérimentation Logement vise à collecter de l'information sur le montant de loyer payé : le tirage de l'échantillon vise donc des logements dont les occupants sont locataires. Cette sous-population, plus mobile que la sous-population des propriétaires, n'est pas toujours facile à contacter : un nombre relativement important de courriers envoyés (19 %) reviennent en NPAI [1].
Ce fort taux de retours en NPAI [2] explique en partie les taux de réponse relativement faibles observés (table 3.2). Ceci est particulièrement marquant pour les populations les plus jeunes : les moins de 29 ans ne rechignent pas tant que ça à répondre à l'expérimentation dès lors que le contact est établi, mais ce contact n'est pas évident, avec près de 40 % des courriers qui reviennent en NPAI et, *in fine*, 22 % des logements échantillonnés pour lesquels une réponse est obtenue.

	NPAI (%)	Répondants (%)	Répondants hors NPAI (%)
Tranche d'âge			
29 ans et moins	40,4	21,9	36,7
30 - 39 ans	24,1	27,0	35,6
40 - 49 ans	16,1	29,2	34,8
50 - 59 ans	11,5	33,0	37,3
60 - 69 ans	10,4	37,6	41,9
70 ans et plus	11,6	35,7	40,4

TABLE 3.2 – Tranche d'âge et contact/participation

3.1.2 Correction de la non-réponse totale en deux phases

On choisit d'effectuer la correction de la non-réponse en deux phases :

— une première phase modélise le fait de recevoir ou non le courrier invitant à répondre à l'enquête ;
— une deuxième phase modélise le fait, une fois le courrier reçu, de répondre ou non à l'enquête.

Cette manière de procéder permet de bien distinguer ces deux mécanismes qui ne mettent pas forcément en jeu les mêmes phénomènes. En effet, ce sont plutôt des facteurs liés à la mobilité résidentielle qui jouent pour la phase de réception du courrier alors que ce sont des déterminants plus classiques de la non-réponse (revenus par exemple) qui importent lors de la deuxième phase.

Première étape de CNR : réception de la lettre-avis

Le modèle retenu pour cette première étape de la correction fait intervenir les variables suivantes :

1. N'habite pas à l'adresse indiquée.
2. Pour comparaison, pour d'autres expérimentations web/papier de l'Insee visant à la fois des propriétaires et des locataires, le taux de NPAI est plutôt aux alentours de 12 %.

— l'âge ;
— le type de logement ;
— le type de propriétaire ;
— la surface ;
— la date d'achèvement ;
— la taille d'unité urbaine ;
— quelques indicatrices sur la qualité de certaines variables de la base de sondage (revenus, statut matrimonial).

La réception de la lettre-avis est modélisée par une régression logistique. Les odds ratios de cette régression sont présentés en annexe A.

Une fois cette régression effectuée, nous récupérons les probabilités estimées $\hat{p}_{1k}$ à partir de ce modèle de réception du courrier. L'utilisation directe des probabilités estimées pouvant poser des problèmes de robustesse, nous constituons H [3] classes d'individus à partir de ces estimations et nous calculons les proportions observées de réception de courrier au sein de ces classes : pour la classe h, cette proportion observée est $\hat{p}_1(h)$. Un individu k de l'échantillon, ayant comme poids initial w_k et appartenant à la classe h se voit attribuer le poids $w_{k(h)}$ à l'issue de cette première phase, avec :

$$w_{k(h)} = w_k \frac{1}{\hat{p}_1(h)}$$

Deuxième phase de correction : répondre ou non à l'enquête

Pour la deuxième phase de correction, notre analyse préliminaire consiste à observer les fréquences de réponses parmi ceux ayant reçu la lettre-avis. Il en ressort que les variables que nous allons retenir pour l'analyse sont :

— le sexe ;
— l'âge ;
— les revenus ;
— la taille d'unité urbaine ;
— une indicatrice de raccordement à l'eau, au chauffage ...

On effectue la régression logistique sur le champ des individus ayant reçu la lettre-avis, avec comme variable à expliquer le fait de répondre à l'enquête, que cela soit par internet ou en retournant le questionnaire papier. Les odds ratios de l'estimation sont donnés en annexe A.

On déroule le même type de traitements qu'à la phase précédente, et ayant construit L [4] classes d'individus lors de cette étape, le poids à l'issue de la correction de la non-réponse totale est au final, pour un individu répondant à l'enquête, appartenant aux classes h et l :

$$w_{k(h,l)} = w_k \frac{1}{\hat{p}_1(h)} \frac{1}{\hat{p}_2(l)}$$

Ainsi, comme l'estimateur

$$\hat{Y} = \sum_{k \in s} \frac{y_k}{\pi_k p_k}$$

(avec p_k la probabilité de réponse de l'individu k) est sans biais pour Y, Y est estimé par

3. On prend ici $H = 50$.
4. Pour cette phase , nous avons choisi L=20.

$$\hat{Y}_{CNR} = \sum_{k \in s} \frac{y_k w_k}{\hat{p}_k}$$

(asymptotiquement sans biais), avec

$$\hat{p}_k = \hat{p}_1(h)\hat{p}_2(l)$$

3.2 Calages et apurements

3.2.1 Le calage sur les variables de la base de sondage

Après l'étape de correction de la non-réponse totale, un calage est effectué en utilisant les marges suivantes issues de la base de sondage :

— le type de logement ;
— l'unité urbaine ;
— la surface du logement ;
— le nombre de pièces ;
— le type de propriétaire ;
— la date d'achèvement ;
— la région ;
— les revenus.

L'un des objectifs de l'expérimentation est d'estimer les paramètres d'intérêt (par exemple, la moyenne nationale du loyer) indépendamment de l'ENL. Ainsi, l'étape de calage présentée ici ne va pas intégrer de marges externes relatives à l'ENL. Comme les variables ont, pour la plupart, été utilisées à l'une ou l'autre des étapes de la correction de la non-réponse totale, l'étape de calage ne modifie pas substantiellement les pondérations : le premier décile du rapport des poids (après/avant calage) est de 0,87 et le neuvième décile est 1,15.

3.2.2 Apurement et périmètre de comparaison

Valeurs aberrantes

86 valeurs de loyer (73 sur papier, 13 sur internet) sont supérieures à 10 000 euros. Si l'on regarde de plus près, il existe une valeur parmi ces 86 pour laquelle le montant est plausible (480 m^2 à Paris, revenus supérieurs au million d'euros). Pour le reste de ces valeurs, l'observation du loyer par m^2 semble indiquer qu'il s'agit vraisemblablement d'une erreur lors du report ou lors de la saisie liée à l'utilisation de décimales. Diviser par 100 ces montants apparait comme une solution raisonnable.

Entre 1 000 et 9 999 euros, on retire une dizaine d'observations pour lesquelles les montants de loyer semblent trop élevés au vu de la surface du logement.

Périmètre, incohérences avec base de sondage

Les comparaisons avec l'ENL vont porter sur le périmètre des locataires qui payent un loyer mensuel et qui n'ont pas emménagé très récemment : ainsi, on va exclure les observations pour lesquelles

le répondant a déclaré avoir reçu le courrier de l'Insee suite à un transfert de courrier ainsi que les logements pour lesquels le locataire a emménagé en 2014 (environ 600 logements appartiennent à l'une des deux catégories) [5]. Vont également être exclues les observations (environ 500) pour lesquelles il existe une incohérence sur la nature du logement (logement collectif ou individuel) entre ce qui est déclaré par le répondant et le contenu de la base de sondage. Environ 8 200 observations sont au final retenues pour les estimations de loyer et les comparaisons avec l'ENL.

3.3 Redressement du loyer

Les étapes de redressement du loyer visent à permettre l'estimation d'un loyer "pur", purgé des éventuels montants payés pour le service de stationnement, des charges et des aides que les locataires auraient pu inclure en reportant le montant de loyer. Les traitements effectués dans cette partie sont inspirés, avec quelques légères adaptations et simplifications, de ceux effectués pour l'ENL et présentés dans le mémoire de Gwendoline Volat (Volat, 2015).

3.3.1 Imputation de montants de service de stationnement

Pour l'ENL, on souhaite retirer le loyer du service de stationnement du montant de loyer déclaré lorsque le logement est dans un immeuble collectif. Un traitement similaire est effectué pour l'expérimentation : lorsque le répondant déclare qu'il dispose d'un service de stationnement, que celui-ci est inclus dans le loyer déclaré et qu'il n'arrive pas à l'isoler, on va chercher à imputer un montant pour le service de stationnement.

L'imputation est réalisée via une méthode de hot-deck aléatoire par classe. Les 4 classes sont :

— logement avec garage ou box en Île-de-France ;
— logement avec parking extérieur ou souterrain en Île-de-France ;
— logement avec garage ou box hors Île-de-France ;
— logement avec parking extérieur ou souterrain hors Île-de-France.

Les donneurs sont les logements pour lesquels un montant est déclaré pour le service de stationnement et les receveurs sont les logements pour lesquels un service de stationnement est inclus dans le loyer déclaré mais n'a pas pu être isolé par le répondant. Le hot-deck aléatoire par classe consiste alors à tirer, pour un receveur donné, un donneur de la même classe, et à attribuer la valeur du donneur au receveur.

3.3.2 Imputations de charges collectives

L'objectif de cette étape est de retirer les charges qui sont comprises dans le loyer déclaré. Des montants de charges vont ainsi être imputés aux logements pour lesquels le loyer comprend certaines charges et pour lesquels les montants de charges ne sont pas reportés, à partir d'une estimation par régression sur la population des logements pour lesquels on dispose d'un montant. Deux modèles sont estimés : l'un pour le logement collectif, l'autre pour les logements individuels. Le modèle sur le logement collectif fait intervenir : la surface, la présence d'ascenseur, le nombre d'habitants du logement, l'étage, le type d'unité urbaine, le type de propriétaire, la date d'achèvement du logement et le mode de chauffage. Le modèle sur le logement individuel fait lui intervenir : la surface,

5. Il s'agit de personnes qui ont déménagé et répondent ici pour un logement qui n'est pas celui tiré lors de l'échantillonnage des logements.

le nombre d'habitants du logement, le type d'unité urbaine, le type de propriétaire et la date d'achèvement du logement.

3.3.3 Les différentes étapes de redressement du loyer

Ainsi, après avoir apuré la variable du loyer déclaré, et préparé les imputations pour les services de stationnement et les charges, les traitements suivants sont réalisés :

— on retire du loyer déclaré le montant du service de stationnement [6] déclaré (environ 500 logements) ou imputé (environ 1 000 logements) lorsque celui est inclus dans le loyer ;
— on retire du loyer déclaré le montant des charges [7] déclarées (environ 1 100 logements) ou imputées (environ 400 logements) lorsqu'elles sont incluses dans le loyer ;
— on ajoute le montant des aides déclarées au loyer déclaré [8] lorsque le montant des aides est supérieur au coût total du logement (loyer, charges et stationnement) : 80 logements sont concernés.

6. En moyenne, le montant retiré aux loyers des logements concernés par cette étape du redressement est de 45 euros.
7. En moyenne, le montant retiré aux loyers des logements concernés par cette étape du redressement est de 93 euros.
8. En moyenne, le montant ajouté aux loyers des logements concernés par cette étape du redressement est de 339 euros.

Chapitre 4

Estimations du loyer issues de l'expérimentation

L'objectif est ici de voir dans quelle mesure le loyer estimé "directement" à partir de l'expérimentation, à savoir sans contrôle ou calage à partir des données de l'ENL mais avec néanmoins des redressements du loyer similaires, est proche ou non de celui issu de l'ENL. Des pistes quant à l'interprétation des écarts entre les estimations issues de ces deux enquêtes sont également fournies.

Par ailleurs, si le loyer constitue la variable d'intérêt principale de l'enquête, des résultats similaires sont présentés en annexe B sur les loyers rapportés à la surface du logement.

4.1 Synthèse sur les différents traitements aval

Le tableau suivant donne les estimations sur les montants nationaux de loyer aux différentes étapes des traitements :

	Moyenne	Q25	Médiane	Q75
Après apurement	553	372	490	650
+ CNR totale	540	370	480	634
+ calages totaux base de sondage	540	370	480	634
+ correction garages	529	360	470	620
+ correction charges	513	346	457	603
+ correction aides	**518**	**350**	**461**	**610**
répondants Internet	559	383	499	651
répondants Papier	474	314	422	564
ENL	**487**	**330**	**434**	**580**

TABLE 4.1 – Synthèse sur les estimations de loyer

À l'issue des traitements, l'écart entre l'estimateur national de la moyenne des loyers issu de l'expérimentation et celui issu de l'ENL est de 31 euros, ce qui correspond à une différence relative d'environ 6 %. La différence sur les montants médians est du même ordre avec un écart entre expérimentation et ENL de 27 euros.
Pour information, les différents paramètres de distribution sont également donnés pour les deux catégories de répondants de l'expérimentation : les répondants sur internet et les répondants sur papier. Les loyers sont globalement beaucoup plus élevés chez les répondants web : ils sont, en

moyenne, de 85 euros supérieurs à ceux des répondants papier. Ces écarts traduisent principalement des différences de sélection entre les populations qui répondent à ces modes. La régression présentée en annexe C nuance ces écarts, après contrôle de la sélection entre les deux modes de réponse.

Comparaison des distributions

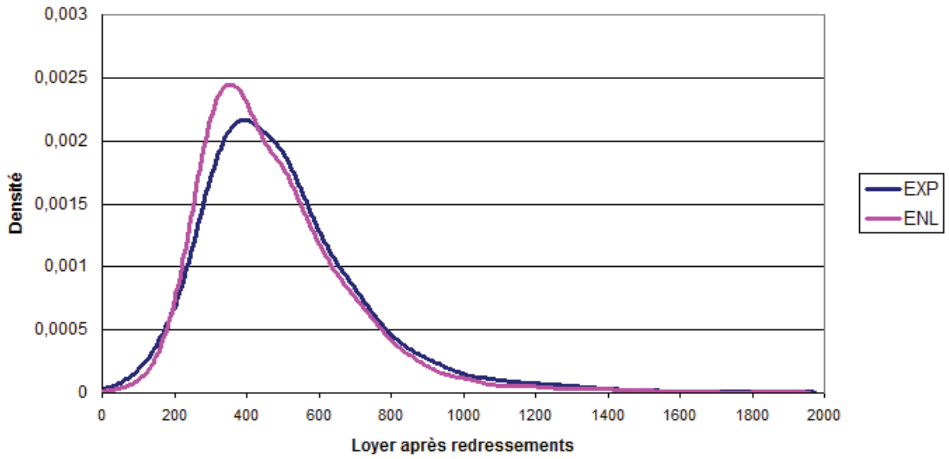

FIGURE 4.1 – Distributions des loyers de l'expérimentaion et de l'ENL

La comparaison des représentations graphiques des distributions des deux enquêtes semble indiquer une concentration plus importante des faibles montants de loyer dans l'ENL entre 250 et 400 euros.

4.2 Calcul de précision

La précision des estimateurs de la moyenne des loyers pour l'expérimentation est calculée en utilisant la macro Sas EVEREST(**E**stimation de **V**ariance dans les **E**nquêtes **R**edressées à **E**chantillon **ST**ratifié) développée au Département des méthodes statistiques de l'Insee. Ce calcul de précision tient compte de la correction de la non-réponse totale ainsi que du calage sur marges. On obtient ainsi pour la moyenne estimée à partir de l'expérimentation l'intervalle de confiance suivant :

	Borne inf de l'IC à 95 %	Moyenne estimée	Borne sup de l'IC à 95 %
Expérimentation Logement	512	518	524
ENL	481	487	494

TABLE 4.2 – Intervalles de confiance des loyers redressés

L'intervalle de confiance (à 95 %) de la moyenne estimée à partir de l'expérimentation ne recoupe pas celui de la moyenne issue de l'ENL : il y a environ une vingtaine d'euros d'écart entre la borne inférieure de l'intervalle calculé pour l'expérimentation et la borne supérieure de l'intervalle de l'ENL.

4.3 Loyer décliné selon diverses caractéristiques du logement

La comparaison des moyennes nationales indique que le montant de loyer est globalement plus élevé d'une trentaine d'euros dans l'expérimentation par rapport à l'ENL. Ce constat d'un loyer plus élevé dans l'expérimentation se retrouve lorsque les loyers sont déclinés suivant les principales caractéristiques des logements.
Les écarts oscillent entre 20 et 40 euros sur les moyennes déclinées par nature du logement (table 4.3), par le caractére privé ou social (table 4.4), par la taille d'unité urbaine (table 4.5) ou encore par la date d'emménagement (table 4.6). Pour ces différentes caractéristiques, les intervalles de confiance à 95 % ne se chevauchent pas.
Il existe quelques cas où les écarts sont bien plus importants comme les logements de 5 pièces et plus (table 4.7) ou ceux dont la surface est supérieure à 120 m^2 (table 4.8) : mais la variabilité des loyers est grande pour ces catégories de logements et les intervalles de confiance se recoupent.
Pour ces enquêtes, rappelons que des extensions régionales ont été réalisées en Île-de-France et en Nord-Pas-de-Calais. Si l'écart est assez réduit pour ce qui est du Nord-Pas-de-Calais (moins de dix euros sur les moyennes), il est, en revanche, assez marqué en Île-de-France, de l'ordre d'une quarantaine d'euros (table 4.9). Pour ces deux régions, sont également fournies les comparaisons par départements (tables 4.10 et 4.11). Ces chiffres sont à prendre avec précaution en raison des effectifs limités, comme en témoignent les intervalles de confiance assez larges. Ainsi, pour l'Île-de-France, le constat d'un montant moyen de loyer plus élevé dans l'expérimentation se retrouve pour la majorité des départements. Une seule exception est celle de la Seine-Saint-Denis, où le montant moyen est inférieur d'une dizaine d'euros dans l'expérimentation par rapport à l'ENL ce qui témoigne peut-être de la difficulté, en face à face, d'enquêter dans les quartiers les plus difficiles du territoire.

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Individuel	601 [588 ;615]	433	549	700	*1 600*
ENL Individuel	560 [536 ;584]	403	515	669	*2 500*
EXP Collectif	498 [491 ;504]	333	438	581	*6 400*
ENL Collectif	465 [457 ;473]	317	406	550	*8 500*

TABLE 4.3 – Comparaison Exp / ENL : nature du logement

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Parc privé	601 [592 ;609]	410	536	700	*4 000*
ENL Parc privé	567 [556 ;577]	400	511	662	*4 600*
EXP Parc social	417 [409 ;423]	297	385	495	*4 100*
ENL Parc social	381 [375 ;388]	285	350	446	*6 400*

TABLE 4.4 – Comparaison Exp / ENL : parc privé ou social

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Rural	490 [476 ;503]	380	480	582	*600*
ENL Rural	466 [432 ;501]	350	450	550	*900*
EXP UU<100 000 hab	461 [451 ;472]	320	421	550	*1 900*
ENL UU<100 000 hab	435 [417 ;454]	306	400	530	*2 600*
EXP UU>100 000 hab	493 [484 ;503]	350	450	583	*2 900*
ENL UU>100 000 hab	471 [455 ;487]	334	429	563	*4 400*
EXP UU Paris	639 [624 ;655]	390	550	750	*2 800*
ENL UU Paris	604 [583 ;624]	366	513	716	*3 100*

TABLE 4.5 – Comparaison Exp / ENL : taille d'unité urbaine

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Avant 2010	496 [487 ;504]	326	440	583	*3 200*
ENL Avant 2010	462 [450 ;475]	312	405	550	*4 500*
EXP 2010 et après	550 [542 ;560]	379	495	650	*4 700*
ENL 2010 et après	516 [502 ;531]	359	468	614	*6 500*

TABLE 4.6 – Comparaison Exp / ENL : date d'emménagement

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP 1 pièce	420 [412 ;432]	301	395	500	*1 200*
ENL 1 pièce	395 [383 ;408]	289	371	480	*1 100*
EXP 2 pièces	480 [470 ;490]	334	431	570	*2 000*
ENL 2 pièces	441 [426 ;455]	308	400	525	*2 300*
EXP 3 pièces	514 [503 ;523]	351	460	601	*2 600*
ENL 3 pièces	472 [454 ;489]	320	420	561	*3 700*
EXP 4 pièces	562 [551 ;576]	390	511	670	*1 700*
ENL 4 pièces	526 [494 ;558]	358	480	630	*2 800*
EXP 5 pièces ou plus	759 [715 ;802]	475	626	850	*600*
ENL 5 pièces ou plus	668 [585 ;751]	439	570	730	*1 000*

TABLE 4.7 – Comparaison Exp / ENL : nombre de pièces

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP 1 à 39 m^2	438 [430 ;450]	309	400	526	*1 400*
ENL 1 à 39 m^2	412 [400 ;425]	299	380	498	*1 500*
EXP 40 à 59 m^2	464 [457 ;473]	320	420	547	*2 000*
ENL 40 à 59 m^2	433 [418 ;447]	300	393	518	*2 500*
EXP 60 à 79 m^2	501 [490 ;510]	355	454	589	*2 400*
ENL 60 à 79 m^2	466 [449 ;482]	326	416	550	*3 700*
EXP 80 à 99 m^2	579 [567 ;594]	413	530	672	*1 500*
ENL 80 à 99 m^2	543 [509 ;577]	385	500	650	*2 200*
EXP 100 à 119 m^2	651 [623 ;678]	462	600	762	*500*
ENL 100 à 119 m^2	631 [521 ;741]	457	572	720	*700*
EXP 120 m^2 et plus	918 [833 ;999]	520	700	1008	*300*
ENL 120 m^2 et plus	834 [583 ;1086]	512	689	880	*400*

TABLE 4.8 – Comparaison Exp / ENL : surface

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Île-de-France	636 [622 ;651]	388	550	750	*2 900*
ENL Île-de-France	598 [580 ;619]	367	513	710	*3 300*
EXP Nord-Pas-de-Calais	471 [455 ;488]	350	430	543	*900*
ENL Nord-Pas-de-Calais	463 [448 ;477]	338	430	550	*1 600*

TABLE 4.9 – Comparaison Exp / ENL : quelques estimations régionales

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Paris	807 [770 ;841]	484	687	961	*1 000*
ENL Paris	758 [720 ;809]	423	634	860	*700*
EXP Seine-et-Marne	539 [499 ;576]	363	490	677	*200*
ENL Seine-et-Marne	502 [467 ;539]	340	464	635	*300*
EXP Yvelines	588 [544 ;631]	379	517	679	*300*
ENL Yvelines	582 [531 ;625]	371	498	700	*300*
EXP Essonne	567 [525 ;608]	368	503	689	*200*
ENL Essonne	506 [467 ;542]	360	461	610	*300*
EXP Hauts-de-Seine	642 [603 ;681]	388	550	770	*400*
ENL Hauts-de-Seine	565 [538 ;595]	343	500	713	*400*
EXP Seine-Saint-Denis	499 [473 ;529]	354	460	605	*300*
ENL Seine-Saint-Denis	507 [472 ;537]	350	459	612	*500*
EXP Val-de-Marne	568 [529 ;614]	341	510	688	*300*
ENL Val-de-Marne	548 [514 ;576]	381	490	664	*400*
EXP Val-d'Oise	530 [502 ;564]	386	506	653	*200*
ENL Val-d'Oise	519 [482 ;544]	360	475	620	*400*

TABLE 4.10 – Comparaison Exp / ENL : estimations départementales (IDF)

	Moyenne	Q25	Médiane	Q75	*Effectif*
EXP Nord	471 [451 ;492]	350	430	550	*600*
ENL Nord	467 [452 ;481]	342	437	562	*1 000*
EXP Pas-de-Calais	471 [443 ;502]	351	430	522	*300*
ENL Pas-de-Calais	456 [432 ;479]	335	424	530	*600*

TABLE 4.11 – Comparaison Exp / ENL : estimations départementales (NPDC)

4.4 Interprétation des écarts

La comparaison directe entre estimateurs de la moyenne des montants de loyer indique que ceux-ci sont plus élevés dans l'expérimentation : les écarts absolus sont de l'ordre de la trentaine d'euros et les écarts relatifs d'environ 6 % sur la moyenne, comme sur la médiane. Plusieurs raisons peuvent être avancées pour expliquer, au moins partiellement, les écarts constatés : certaines tiennent au périmètre de la comparaison, d'autres relèvent plutôt des effets de mode au sens large du terme. L'effet de mode au sens large regroupe ainsi d'un côté, les différences liées à la sélection des répondants qui n'est pas la même pour les deux enquêtes, de l'autre, les différences qui relèvent de la mesure car le mode de collecte n'est pas le même (auto-administré vs face à face).

4.4.1 Décalage temporel

Alors que la collecte de l'expérimentation s'est déroulée entre mars et juin 2014, celle de l'ENL, si elle s'est également terminée en 2014, s'est majoritairement effectuée en 2013 : plus de 90 % des questionnaires de l'ENL ont été recueillis au deuxième semestre de 2013. Ainsi, une partie de l'écart constaté entre les estimateurs de loyer pourrait relever de ce décalage temporel. Au vu des éléments disponibles [1], cela ne peut néanmoins expliquer que très partiellement l'écart constaté entre les estimateurs de loyer des deux enquêtes.

4.4.2 Périmètre des emménagés récents

L'unité statistique que l'on souhaite enquêter dans ces enquêtes est le logement. Ainsi, pour l'expérimentation, ce sont des logements qui sont tirés dans la base de sondage même si le courrier est adressé à une personne. Ce type de protocole fait qu'il n'est guère possible de collecter de l'information sur un logement lorsque celui-ci n'est plus occupé, au moment de l'enquête, par le ménage qui l'occupait au moment du tirage dans la base de sondage [2]. Les personnes qui ont emménagé très récemment [3] sont exclues du périmètre de l'expérimentation [4] ce qui n'est pas le cas dans l'ENL. Une comparaison à périmètre constant serait plutôt de nature à augmenter les écarts constatés : une restriction aux logements dont les locataires n'ont pas emménagé moins d'un an et demi avant la date de l'enquête conduit à un montant moyen de 481 euros pour l'ENL (contre 487 euros sans la restriction) et une médiane de 426 euros (contre 434 euros sans la restriction).

4.4.3 Caractéristiques observables des logements

Le parc de logements sur lequel repose l'estimation à partir de l'ENL n'est pas exactement celui de l'expérimentation : les bases de sondage ne sont pas les mêmes, les marges de calages diffèrent, etc. Le chapitre suivant vise à essayer de contrôler ces différences de caractéristiques ; cela modifie l'écart estimé puisque l'écart passe, après contrôle de la sélection, à moins d'une vingtaine d'euros.

1. D'après la série des loyers effectifs de l'indice des prix à la consommation harmonisé, la hausse est d'environ 0,7 % entre octobre 2013 et mai 2014.

2. Il faudrait, en cas de déménagement, que le ménage qui occupe dorénavant le logement, ouvre le courrier du ménage prédécesseur, ce qui ne semble guère convenable !

3. Lorsque l'emménagement est postérieur à la date du tirage dans la base de sondage.

4. Soit elles ne répondent pas, soit elles répondent car on fait suivre leur courrier mais sont exclues dès lors qu'elles signifient qu'elles ont déménagé ou que leur date d'emménagement indiquée est trop récente.

4.4.4 Caractéristiques inobservables, sélection des répondants

Une modélisation fine du montant de loyer pourrait intégrer certaines caractéristiques qui n'apparaissent pas dans les variables de contrôle, faute de disponibilité : exposition, isolation du logement ou encore caractéristiques du quartier. L'effet estimé est-il pour autant biaisé et si oui, dans quel sens ? Il est difficile d'avoir une idée des corrélations de ces variables omises avec le mode de collecte, d'anticiper le sens et l'ampleur des biais et donc d'en conclure que l'écart s'explique en partie par l'absence de certaines variables de contrôle.

En revanche, une partie de l'écart pourrait s'expliquer par la forte sélection des répondants de l'expérimentation. Un des résultats de l'expérimentation web sur le logement de 2010 était que les répondants sont globalement moins satisfaits de leurs conditions de logement que ceux de l'ENL 2006 (Amiel et Denoyelle, 2012). Ce résultat peut s'interpréter de deux manières. Cela peut relever de l'effet de mode "pur" lié à la mesure : pour des raisons liées à la moindre désirabilité sociale qu'en face à face, les répondants n'hésitent pas à faire part de leur mécontentement sur leurs conditions de logement. L'autre possibilité tiendrait à la forte sélection des répondants : lorsque les taux de réponse sont particulièrement faibles, ceux qui ont tendance à répondre seraient ceux qui ont des choses à dire, et plutôt en mal, sur la thématique en question. Ainsi, il est possible qu'il existe des facteurs inobservables qui participent au processus de sélection des individus et que ces facteurs soient fortement liés aux variables d'intérêt (Razafindranovona, 2015). Cela ne reste qu'une hypothèse, mais il est possible qu'un tel mécanisme opère dans le cadre de l'expérimentation auto-administrée de 2014. Les personnes qui répondent seraient plutôt celles ayant des choses à dire sur leurs conditions de logement et en particulier sur leur montant de loyer, question centrale de l'expérimentation. Ainsi, les personnes, fortement sélectionnées, qui répondent à l'expérimentation pourraient être plutôt celles qui sont insatisfaites de leur logement en particulier de son rapport qualité prix. Cela pourrait expliquer qu'à caractéristiques plus ou moins contrôlées, l'estimation du loyer à partir de l'expérimentation soit plus élevée que celle issue de l'ENL.

4.4.5 Effet de mode ?

Une différence constatée entre les valeurs estimées selon deux modes de collectes différents (face à face et auto-administré) pourrait également relever de l'effet de mode "pur" lié à la mesure. Une même personne pourrait ne pas reporter son loyer de la même manière selon le mode de collecte.

Une première différence que l'on peut évoquer tient au *satisficing*[5] : en auto-administré, faute d'enquêteur pour motiver le répondant, il est possible que les efforts nécessaires pour reporter la valeur correcte du montant de loyer payé ne soient pas mis en oeuvre. Cela est d'autant plus vraisemblable ici que l'effort demandé est assez important. En effet, rappelons que les différentes étapes du processus cognitif qu'engage le répondant face à une question posée lors d'une enquête sont : la compréhension de la question, la récupération de l'information pertinente, l'utilisation de cette information pour élaborer son jugement et enfin la sélection et le report de la réponse (Tourangeau, 2000). Dans le cadre de cette expérimentation qui vise à estimer les montants de loyer, la première [6]

5. Le *satisficing* renvoie à la faible implication du répondant. Ce phénomène peut s'appréhender à travers les modèles théoriques de réponse aux enquêtes. D'après la théorie du *satisficing* de Krosnick (1991), les répondants n'ont pas toujours la motivation et l'attention nécessaires au bon déroulement du processus cognitif qu'engage le répondant face à une question posée lors d'une enquête. Ainsi, les réponses fournies ne sont pas toujours optimales.

6. Le questionnaire est jugé globalement plus compliqué que ceux d'autres expérimentations auto-administrées similaires (2.1.5).

et la deuxième étapes [7] sont assez lourdes. Dans un cadre auto-administré, cela peut ainsi générer un *satisficing* plus important qu'en face à face. L'information recueillie peut être de moindre qualité : cela peut par exemple se matérialiser par un recours à l'arrondi lors du report du montant de loyer. Il n'est en revanche pas évident de savoir *a priori* si cela peut expliquer le fait que les loyers soient globalement plus élevés chez les répondants de l'expérimentation.

Une chose que l'on peut néanmoins constater est que, sur la base des données de l'expérimentation, le fait de ne pas avoir eu recours à des documents externes est plutôt corrélé positivement au montant de loyer : une régression sur le montant de loyer (après redressements) contrôlant de diverses caractéristiques donne un coefficient de +13 euros à la variable "pas de recours aux documents". Ceci indique quel type de conséquence peut avoir la forme de *satisficing* qui consiste à ne pas faire l'effort de recherche optimal pour reporter la réponse pertinente à la question du montant de loyer. Ces résultats sont présentés en annexe C.

Le *satisficing* peut aussi se manifester par un moindre effort des ménages pour s'approprier les concepts abordés dans le questionnement de l'expérimentation. Il est ainsi difficile de repérer et de corriger parfaitement les cas où le ménage n'aurait pas fait l'effort de renseigner un loyer hors charges et se serait contenté d'indiquer le montant de loyer, charges comprises, sans doute plus facile à connaître.

Un autre mécanisme qui pourrait expliquer les différences est celui de la désirabilité sociale : face à un enquêteur, le répondant peut avoir tendance à fournir des réponses plus socialement dans la norme qu'en auto-administré. Cela est sans doute plus marqué pour les questions d'opinion que pour les questions plutôt factuelles comme celles qui portent sur le montant de loyer payé, mais cela peut jouer pour certains montants "hors normes" de loyer [8].
Enfin, travestir la vérité, y compris à des fins de manipulation, est plus aisé en auto-administré que face à un enquêteur qui exerce une sorte de contrôle social. Si le locataire pense que l'expérimentation [9] sera utilisée à des fins de politique publique, dans une période où la loi dite ALUR [10] a été adoptée (le 24 mars 2014), la tentation peut être de reporter un montant de loyer plus élevé.

Ces différents phénomènes étudiés séparément peuvent se conjuguer : une personne qui ne connait pas parfaitement son montant de loyer et qui n'a pas le courage de chercher sa quittance de loyer peut être tentée d'arrondir allègrement son montant de loyer à la centaine supérieure parce qu'il considère que cet arrondi est pertinent même s'il travestit quelque peu la vérité.
Et l'effet de mode peut également accompagner un effet de sélection préalable : l'exemple cité précédemment peut porter sur une personne qui répond à l'enquête parce qu'elle considère qu'elle paie bien trop cher pour son logement et qu'elle veut le faire savoir.

7. La récupération de l'information pertinente n'est pas toujours immédiate et peut demander un recours à des documents externes.

8. Les montants de loyer très élevés sont plus fréquents dans l'expérimentation : environ 0,5 % des loyers sont supérieurs à 2 000 euros dans l'expérimentation contre 0,3 % dans l'ENL.

9. Le répondant n'est pas censé savoir que l'enquête en question est expérimentale.

10. Loi pour l'accès au logement et un urbanisme rénové qui vise pour partie à réguler le marché de la location.

Chapitre 5

Différentes méthodes pour contrôler la sélection

Dans la partie précédente, on cherche à estimer les loyers sans se caler en quoi que ce soit sur les marges de l'ENL (mais en appliquant des traitements similaires). Dans cette partie, on cherche à contrôler les différences de sélection (les caractéristiques des logements pour lesquels une réponse est obtenue ne sont pas forcément les mêmes dans l'expérimentation et l'ENL) selon plusieurs méthodes. Les écarts qui persistent à l'issue de ce contrôle de la sélection peuvent s'interpréter en tant qu'effet de mode, soit lié à la mesure, soit lié à la sélection non contrôlée.

5.1 Principes généraux

Le besoin de contrôler la sélection renvoie à un problème classique de la littérature sur les effets de mode. Il est souvent difficile de conclure quant à l'existence ou non d'effets de mode liés à la mesure car on ne sait pas toujours si les écarts constatés relèvent vraiment de la mesure ou plutôt de différences dans les populations qu'on cherche à comparer. Pour contrôler la sélection, plusieurs techniques peuvent être mobilisées pour essayer de rendre les populations les plus comparables possibles. Les jeux de pondérations utilisés (pour l'expérimentation comme pour l'ENL) sont ceux après correction de la non-réponse totale et calage (étapes décrites au chapitre 3 pour l'expérimentation).

Les variables de contrôle utilisées sont :
— la surface du logement ;
— la région ou le département pour l'IDF et le NPdC ;
— le taille d'unité urbaine ;
— le nombre d'habitants du logement ;
— le nombre de pièces du logement ;
— le type de propriétaire ;
— la date d'achèvement ;
— l'étage du logement ;
— la présence d'ascenseur ;
— la date d'emménagement ;
— l'entité à laquelle le loyer est versé ;
— le diplôme [1].

1. Le diplôme du répondant dans l'expérimentation, de la personne de référence dans l'ENL

5.2 Contrôle de la sélection pour estimer l'effet sur la moyenne

Plusieurs méthodes de contrôles de la sélection ont été mises en oeuvre : une régression linéaire, une repondération par inverse du score de propension et un calage sur marges. Ces différentes méthodes donnent des résultats plutôt similaires, présentés dans le tableau 5.1.

5.2.1 Contrôle par régression linéaire

La première méthode mobilisée est une régression linéaire du loyer redressé sur l'ensemble des explicatives citées plus haut et sur une indicatrice de source de l'observation (expérimentation ou ENL). Ainsi :

$$Y_i = \alpha + \delta \mathbb{1}_{[enquête=ENL]i} + \gamma X_i + \epsilon_i$$

Y_i : loyer redressé, X_i : variables listées plus haut, ϵ_i : terme d'erreur

Si on pose $\beta = (\alpha, \delta, \gamma)$ et que W est la matrice diagonale des poids w_i, l'estimateur est

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

Par ailleurs le même type de modélisation (hors indicatrice de source de l'observation) sera mobilisé pour savoir si les effets sur le loyer des différents déterminants sont du même ordre dans l'ENL et dans l'expérimentation. Deux régressions, l'une restreinte à l'ENL, l'autre à l'expérimentation sont effectuées et les coefficients des variables les plus significatives sont présentés en annexe D.

5.2.2 Contrôle par repondération par l'inverse du score de propension

L'idée générale de l'ajustement par le score de propension est de contrôler le biais lié aux différences de caractéristiques observables dans les deux populations à comparer (Givord, 2010; Hirano et Imbens, 2001; Hirano *et al.*, 2003). Ainsi, l'ajustement va sur-pondérer les unités "traitées" (la source est l'ENL) dont les caractéristiques sont comparables à celles des unités "non traitées" (la source est l'expérimentation). Inversement, sont également sur-pondérées les unités "non traitées" dont les caractéristiques sont comparables aux unitées "traitées"'.
Ici, le score de propension $P(X)$ est la "probabilité" d'être une observation dont la source est l'ENL sachant diverses caractéristiques X citées plus haut. Les facteurs de repondération sont $(P(X_i))^{-1}$ pour les observations issues de l'ENL et $(1-P(X_i))^{-1}$ pour les observations issues de l'expérimentation [2].

L'estimateur utilisé est le suivant :

$$\hat{\delta} = \frac{\sum_{i=1}^{n} \frac{\mathbb{1}_{[enquête=ENL]i}\, w_i Y_i}{\hat{P}(X_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}_{[enquête=ENL]i}\, w_i}{\hat{P}(X_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-\mathbb{1}_{[enquête=ENL]i})\, w_i Y_i}{1-\hat{P}(X_i)}}{\sum_{i=1}^{n} \frac{(1-\mathbb{1}_{[enquête=ENL]i})\, w_i}{1-\hat{P}(X_i)}}$$

2. Cela repose sur

$$E[Y_1] = E[\frac{P(X)Y_1}{P(X)}] = E[E[\frac{TY_1}{P(X)}|X]] = E[\frac{TY}{P(X)}]$$

et

$$E[Y_0] = E[\frac{(1-T)Y}{1-P(X)}]$$

avec les notations habituelles de la littérature sur l'évaluation des traitements, à savoir, Y la variable d'intérêt, Y_0 et Y_1 les outcomes potentiels, T la variable de traitement et $P(X) = P(T = 1|X)$.

Avec $\hat{P}(X_i)$ le score de propension estimé par régression logistique, où la variable à expliquer est $\mathbb{1}_{[enquête=ENL]}$, les explicatives sont les X_i et les pondérations utilisées sont les w_i.

5.2.3 Contrôle par calage sur marges

L'idée ici est proche de ce qui est développé dans un article de Hainmueller sur l'*entropy balancing* pour estimer un effet causal (Hainmueller, 2012). Pour rendre deux populations comparables, on va chercher à caler les totaux (ou les moments) d'une population sur les totaux (ou les moments) de l'autre population. Les totaux ou moments vont être rapprochés par la technique habituelle de calage sur marges (Sautory, 1993). L'effet causal pourra alors être estimé en utilisant le nouveau jeu de pondérations issu du calage.

Formellement, connaissant les totaux estimés $X_j(ENL)$ des différentes variables citées en 5.1 sur une population de référence, on cherche à caler l'autre population sur ces totaux de référence [3]. Les totaux doivent alors vérifier des équations de calage de type

$$\forall j = 1 \dots J \quad \sum_{i=1}^{n} (1 - \mathbb{1}_{[enquête=ENL]i}) w_k x_{ji} = X_j(ENL) \tag{5.1}$$

On va alors chercher un système de poids w_k solution de

$$\min_{w_i} \sum_{i=1}^{n} (1 - \mathbb{1}_{[enquête=ENL]k}) d_i \;\; G(\frac{w_i}{d_i})$$

où G est une fonction de distance, sous les contraintes des équations de calage (5.1) et les d_i sont les pondérations de l'expérimentation (après traitement de la non-réponse et calage sur les totaux de la base de sondage).

L'estimateur de l'effet sera alors :

$$\hat{\delta} = \frac{\sum_{i=1}^{n} \mathbb{1}_{[enquête=ENL]i} d_i Y_i}{\sum_{i=1}^{n} \mathbb{1}_{[enquête=ENL]i} d_i} - \frac{\sum_{i=1}^{n} (1 - \mathbb{1}_{[enquête=ENL]i}) w_i Y_i}{\sum_{i=1}^{n} (1 - \mathbb{1}_{[enquête=ENL]i}) w_i}$$

5.2.4 Synthèse des résultats

	Régression linéaire	Score de propension (repondération)	Calage
Logement collectif	-22	-21	-23
Ensemble des logements	-26	-25	-26

TABLE 5.1 – Écarts de loyer entre l'ENL et l'expérimentation après contrôle de la sélection

Le contrôle de la sélection réduit quelque peu les écarts : ainsi, dans le logement collectif, alors que la comparaison entre les estimateurs de moyennes calculés, après traitements, à partir des deux enquêtes, donne un écart de 33 euros, il est de 21 à 23 euros après contrôle de la sélection. Sans rentrer dans les détails, les caractéristiques du logement ou les variables issues de la base de sondage ne contribuent pas significativement à la réduction des écarts constatés, sans doute car ces différentes caractéristiques participent déjà aux différentes étapes de traitement post-collecte de l'expérimentation. En revanche, intégrer le diplôme dans le contrôle de la sélection semble pertinent pour rendre

3. Il s'agit donc ici de totaux estimés en utilisant les pondérations finales (après traitement de la non-réponse et calage) de l'ENL.

les populations comparables. Néanmoins, comme l'unité statistique de base pour ces enquêtes est le logement, il ne faut sans doute pas aller trop loin dans la recherche d'une parfaite comparabilité du point de vue sociodémographique et se limiter à ce qui semble essentiel pour caractériser le logement et son occupation.

5.3 Régressions quantiles

Alors que les précédentes méthodes de contrôles de la sélection sont bien adaptées pour estimer des effets sur les moyennes, la manière la plus pertinente pour estimer les écarts le long de la distribution des loyers est de mobiliser les régressions quantiles. Les régression quantiles tentent ainsi d'évaluer comment les quantiles conditionnels $q_\tau(Y|X)$ se modifient lorsque les déterminants X varient (Givord et D'Haultfoeuille, 2014).

On suppose que les quantiles de la distribution conditionnelle des loyers ont une forme linéaire :

$$q_\tau(Y|X) = X'\beta_\tau$$

où à chaque τ correspond un vecteur de coefficients $\beta_\tau = (\beta_{1\tau}, ..., \beta_{p\tau})$ correspondant aux p variables explicatives avec $X = (1, X_1, \ldots, X_p)$ l'ensemble des explicatives citées en 5.1 et Y le montant de loyer.

Les effets estimés pour chaque niveau de quantile sont les suivants pour le coefficient de la variable indicatrice de l'ENL :

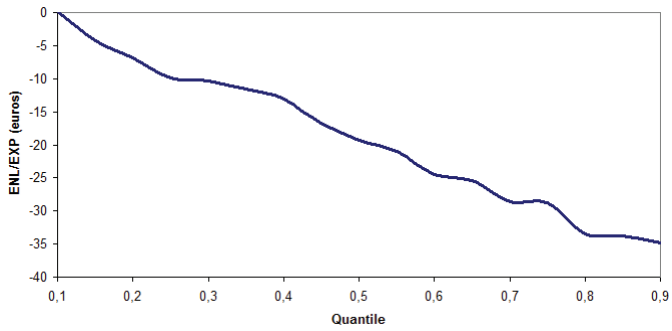

FIGURE 5.1 – Régression quantile sur l'ensemble des logements : effet ENL(/EXP) sur le loyer en euros

Le graphique 5.1 se lit ainsi : par exemple, le fait que l'abscisse 0,4 croise la courbe à -13 euros signifie que, conditionnellement aux autres caractéristiques observables, le quatrième décile de la distribution des loyers issus de l'ENL est inférieur de 13 euros au quatrième décile de la distribution des loyers issus de l'expérimentation. On remarque que ces écarts absolus sont globalement de plus en plus importants le long de la distribution : ils sont quasiment nuls en début de distribution, sont de -19 euros à la médiane et de -35 euros au neuvième décile. Le contrôle de la sélection n'est pas neutre et tend à réduire les écarts entre expérimentation et ENL : l'écart sur les médianes, après contrôle de la sélection, est plus faible que celui estimé directement en 4.1 qui était de -27 euros.

Mais ce constat peut se nuancer car les écarts ne grandissent pas énormément dans le haut de la distribution. On peut ainsi modéliser le logarithme du loyer au lieu du loyer, c'est-à-dire que l'on considère que ce sont les quantiles de la distribution conditionnelle des logarithmes des loyers qui ont une forme linéaire :

$$q_\tau(log(Y)|X) = X'\beta_\tau$$

La représentation graphique 5.2 des effets estimés peut alors s'interpréter en termes d'écarts relatifs. On observe alors qu'à partir du septième décile, les écarts relatifs se stabilisent aux alentours de 6 %.

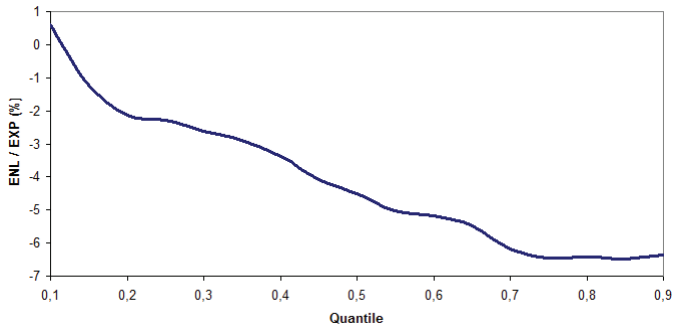

FIGURE 5.2 – Régression quantile sur l'ensemble des logements : effet ENL(/EXP) sur le loyer en %

Conclusion

Les objectifs de l'expérimentation Logement sont bien circonscrits. Il s'agit de savoir s'il est possible de recueillir de l'information de qualité via un mode auto-administré sur les montants de loyer et de comparer les estimations de ces montants avec ceux de l'enquête étalon qu'est l'enquête nationale Logement. Après traitements post-collecte, l'estimation du montant moyen de loyer de l'expérimentation est supérieure d'une trentaine d'euros à celle issue de l'ENL : cela correspond à un écart relatif d'environ 6 %. Après contrôle de la sélection, l'écart est plutôt de l'ordre de 25 euros. Difficile d'émettre un jugement sur l'ampleur de cet écart : est-il satisfaisant ou rédhibitoire ? D'un point de vue statistique, les différences sont significatives et les intervalles de confiance à 95 % calculés pour ces estimateurs ne se chevauchent pas. Mais il ne faut pas oublier que les coûts budgétaires et temporels ne sont pas du même ordre et qu'il est donc acceptable que la qualité de l'expérimentation auto-administrée soit moindre. Que faudrait-il néanmoins améliorer si l'on souhaite, dans le futur, estimer des loyers (que cela soit au niveau national ou à un échelon plus local) à partir d'un protocole auto-administré, à l'aune de cette expérimentation ?

Un premier axe d'amélioration consiste à réduire au maximum les biais liés à la non-réponse.
Cela consiste tout d'abord à tout faire pour augmenter les taux de réponse afin de ne pas avoir à émettre d'hypothèses trop fortes lors des étapes de correction. L'effort pourrait potentiellement porter sur toutes les différentes étapes que doit franchir l'enquêté avant d'être réellement engagé dans le questionnaire. Ainsi, les premiers contacts entre l'organisation enquêtrice et l'enquêté s'effectuent au travers de la lettre-avis et, même encore avant, de l'enveloppe qui la contient. Tout doit être fait pour que ces éléments dans leur forme comme dans leur contenu suscitent la participation de l'enquêté en insistant sur le caractère officiel du dispositif. Un autre point important est que l'effort à fournir pour se rendre sur la version en ligne du questionnaire (ou pour retourner la version papier du questionnaire) doit être minimal : il ne faut pas que cette étape soit susceptible de générer des abandons. Et comme les premières impressions peuvent être déterminantes, un soin particulier doit être apporté à la première page du site qui accueille le répondant.
En ce qui concerne le traitement de la non-réponse à proprement parler, il semble préférable de ne pas totalement faire l'impasse sur le recueil de quelques caractéristiques sociodémographiques même lorsque l'unité statistique est le logement. Une question sur le diplôme semble être d'un bon rapport qualité-prix et peut compléter les données qui proviennent directement de la base de sondage dans les modèles de correction de la non-réponse.

L'autre axe d'amélioration possible porte plus précisément sur la qualité de l'information collectée. Si l'exploitation de l'expérimentation n'apporte pas de conclusion définitive quant à l'ampleur des effets de mode liés à la mesure, qu'il est toujours difficile de distinguer des effets dus à la sélection non contrôlée, quelques constats et pistes d'amélioration peuvent être fournis. Ainsi, comme la littérature le soutient, cette expérimentation indique que la tendance au *satisficing* est plus forte en auto-administré que face à un enquêteur. En effet, lorsque l'effort de recherche et l'effort cognitif demandés pour reporter correctement l'information sont trop importants, l'enquêté peut être tenté

de ne pas tout faire pour donner la réponse optimale. Dans cette expérimentation cela se traduit par un moindre recours aux documents externes et par un comportement d'arrondi plus fréquent, ces deux mécanismes étant par ailleurs intimement liés.
Pour éviter le recours aux arrondis abusifs, une première possibilité serait d'insister fortement sur la nécessité d'obtenir un montant de loyer déclaré "exact", tel qu'il apparaît sur les quittances du locataire. Une autre manière de faire plus contraignante serait de mettre en place des contrôles qui se déclencheraient lorsque des valeurs arrondies seraient renseignées.
Cette expérimentation pose également la question de la charge cognitive du répondant. Les concepts ne sont pas évidents à comprendre et le répondant peut se sentir désemparé devant la difficulté en absence d'enquêteur pour le guider et le motiver. Peut-être qu'une piste à explorer serait de proposer une option de joindre électroniquement sa quittance au questionnaire : cela reporterait inévitablement une partie de la charge de l'enquêté vers l'organisation enquêtrice et engendrerait des coûts supplémentaires à arbitrer face aux éventuels gains en qualité.

Enfin, une enquête auto-administrée sur les loyers peut également s'envisager non pas comme une enquête totalement autonome mais plutôt comme un dispositif complémentaire qui se raccrocherait à une enquête étalon comme l'ENL d'une manière qui reste encore à imaginer ...

Bibliographie

AMIEL, M.-H. et DENOYELLE, T. (2012). Enquêtes en ligne : comparaison de modes de questionnement sur le thème du logement. *In Journées de Méthodologie Statistique*. http://jms.insee.fr/files/documents/2012/945_4-JMS2012_S26-3_AMIEL-ACTE.PDF.

de PERETTI, G. et RAZAFINDRANOVONA, T. (2014). Les enquêtes multimode : attention aux effets de mode. *Statistique et société*, 2.

DRECHSLER, J. et KIESL, H. (2015). Beat the heap : An imputation strategy for valid inferences from rounded income data. *Journal of Survey Statistics and Methodology*.

GIVORD, P. (2010). Méthodes économétriques pour l'évaluation de politiques publiques. *INSEE, Document de travail de la DESE, G2010-08*. http://www.insee.fr/fr/publications-et-services/docs_doc_travail/G2010-08.pdf.

GIVORD, P. et D'HAULTFOEUILLE, X. (2014). La régression quantile en pratique. *Économie et statistique*, (471):85–111.

GROVES, R. M. et LYBERG, L. (2010). Total survey error : Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879.

HAINMUELLER, J. (2012). Entropy balancing for causal effects : A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

HIRANO, K. et IMBENS, G. W. (2001). Estimation of causal effects using propensity score weighting : An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278.

HIRANO, K., IMBENS, G. W. et RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

HOLBROOK, A. L., ANAND, S., JOHNSON, T. P., CHO, Y. I., SHAVITT, S., CHÁVEZ, N. et WEINER, S. (2014). Response heaping in interviewer-administered surveys is it really a form of satisficing ? *Public Opinion Quarterly*, 78(3):591–633.

KENDALL, M. (1938). The conditions under which Sheppard's corrections are valid. *Journal of the Royal Statistical Society*, 101(3):592–605.

KROSNICK, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236.

LYBERG, L. (2012). La qualité des enquêtes. *Techniques d'enquête*, 38(2):115–142.

RAZAFINDRANOVONA, T. (2015). La collecte multimode et le paradigme de l'erreur d'enquête totale. *INSEE, Document de travail de la DMCSI, M2015-01*. http://www.insee.fr/fr/publications-et-services/docs_doc_travail/la%20collecte%20multimode-b.pdf.

SAUTORY, O. (1993). La macro CALMAR. *Redressement d'un échatillon par calage sur marges, Série des documents de travail de la Direction des Statistiques Démographiques et Sociales*, 55.

SHEPPARD, W. F. (1897). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, 1(1):353–380.

TOURANGEAU, R. (2000). *The psychology of survey response.* Cambridge University Press.

VOLAT, G. (2015). Le redressement des variables financières dans l'enquête logement 2013. *Mémoire de stage, Master Statistique Econométrie de l'Ensai.*

Annexe A

Déterminants de la réception des courriers et de la réponse à l'expérimentation

	Odds ratios	Significativité
Tranche d'âge		
Moins de 23 ans vs 40-49 ans	0,50	***
24-25 ans vs 40-49 ans	0,37	***
26-28 ans vs 40-49 ans	0,43	***
29-32 ans vs 40-49 ans	0,54	***
33-39 ans vs 40-49 ans	0,76	***
50-59 ans vs 40-49 ans	1,44	***
60-69 ans vs 40-49 ans	1,63	***
70 ans et plus vs 40-49 ans	1,28	***
Type de logement		
Maison vs Appartement	1,60	***
Surface du logement		
Moins de 20m2 vs Plus de 40m2	0,58	***
20 à 39m2 vs Plus de 40m2	0,78	***
Unité urbaine		
Agglomération de Paris vs France hors agglo Paris	1,74	***
Type de propriétaire		
HLM vs Personne physique	1,53	***
Autre personne morale qu'HLM vs Personne physique	1,23	***
Date d'achèvement		
2002 vs Avant 2002	0,79	***
Après 2002 vs Avant 2002	0,84	***

TABLE A.1 – Phase 1 : modélisation de la réception de la lettre-avis

Significativité : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

L'année 2002 est utilisée pour définir les catégories car il existe une incertitude dans les bases de sondage lorsque l'achèvement est renseigné à cette date là.

	Odds ratios	Significativité
Sexe		
Femme vs Homme	1,24	***
Age		
Moins de 29 ans vs 40-49 ans	1,17	***
30-39 ans vs 40-49 ans	1,04	
50-59 ans vs 40-49 ans	1,13	**
60-69 ans vs 40-49 ans	1,38	***
70 ans et plus vs 40-49 ans	1,32	***
Revenus		
Décile 1 vs Décile 6	0,63	***
Décile 2 vs Décile 6	0,74	***
Décile 3 vs Décile 6	0,72	***
Décile 4 vs Décile 6	0,78	***
Décile 5 vs Décile 6	0,89	.
Décile 7 vs Décile 6	1,16	*
Décile 8 vs Décile 6	1,16	**
Décile 9 vs Décile 6	1,19	***
Décile 10 vs Décile 6	1,26	***
Type de propriétaire		
HLM vs Personne physique	1,14	***
Autre personne morale qu'HLM vs Personne physique	1,06	
Unité urbaine		
UU 100 000 - 1 999 999 hab vs Rural et UU <100 000 hab	1,16	***
Agglo Paris (hors Paris) vs Rural et UU <100 000 hab	0,75	***
Paris vs Rural et UU <100 000 hab	1,33	***
Raccordements divers		
≥ 2 parmi (eau, égoût, chauffage) vs ≤ 1 parmi (eau, égoût, chauffage)	0,75	***

TABLE A.2 – Phase 2 : modélisation de la réponse à l'enquête

Significativité : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Annexe B

Comparaison des loyers au m^2

Les tableaux suivants fournissent le même type de résultats que ceux présentés dans le chapitre 4, mais cette fois, sur le rapport entre loyer et surface.
Deux types de résultats sont présentés : dans un premier temps, on présente des données sur les moyennes (et quelques autres indicateurs) du ratio loyer / surface, puis ensuite on s'intéresse aux ratios des totaux loyers et surfaces (pour lesquels sont fournis des intervalles de confiance à 95 %). Dans les deux cas, les conclusions sont du même type que pour le loyer : les indicateurs rapportant le loyer à la surface sont globablement plus élevés dans l'expérimentation que dans l'ENL.

B.1 Moyennes (et autres indicateurs) du ratio loyer / surface

	Moyenne	Q25	Médiane	Q75
Expérimentation	9,65	5,62	7,75	11,29
ENL	*8,71*	*5,19*	*7,06*	*10,38*

TABLE B.1 – Estimations de loyer au m^2 après redressements

	Moyenne	Q25	Médiane	Q75
EXP Individuel	7,41	5,43	6,67	8,37
ENL Individuel	6,85	5,04	6,27	8,00
EXP Collectif	10,20	5,81	8,33	12,19
ENL Collectif	9,25	5,25	7,48	11,27

TABLE B.2 – Comparaison Exp / ENL : nature du logement

	Moyenne	Q25	Médiane	Q75
EXP Parc privé	11,82	7,26	9,78	14,06
ENL Parc privé	10,72	6,72	9,14	12,64
EXP Parc social	6,94	4,63	5,95	7,76
ENL Parc social	6,04	4,40	5,40	6,74

TABLE B.3 – Comparaison Exp / ENL : parc privé ou social

	Moyenne	Q25	Médiane	Q75
EXP Rural	6,54	5,00	6,16	7,79
ENL Rural	6,04	4,52	5,64	7,14
EXP UU<100 000 hab	7,71	5,14	6,67	8,94
ENL UU<100 000 hab	6,87	4,72	6,25	8,09
EXP UU>100 000 hab	9,32	5,87	8,34	11,49
ENL UU>100 000 hab	8,65	5,42	7,81	10,91
EXP UU Paris	13,79	6,60	10,83	19,31
ENL UU Paris	12,74	6,26	9,98	17,58

TABLE B.4 – Comparaison Exp / ENL : taille d'unité urbaine

	Moyenne	Q25	Médiane	Q75
EXP Île-de-France	13,64	6,59	10,77	19,04
ENL Île-de-France	12,54	6,25	9,75	16,96
EXP Nord-Pas-de-Calais	8,27	5,18	6,69	9,64
ENL Nord-Pas-de-Calais	7,37	5,05	6,31	8,67

TABLE B.5 – Comparaison Exp / ENL : quelques estimations locales

	Moyenne	Q25	Médiane	Q75
EXP Avant 2010	8,59	5,05	6,80	9,94
ENL Avant 2010	7,80	4,69	6,27	9,11
EXP 2010 et après	10,89	6,36	8,96	12,81
ENL 2010 et après	9,73	5,93	8,08	11,67

TABLE B.6 – Comparaison Exp / ENL : date d'emménagement

B.2 Ratio des totaux

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
Expérimentation	8,03	8,13	8,23
ENL	*7,41*	*7,49*	*7,58*

TABLE B.7 – Ratio loyer / surface

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
EXP Individuel	6,71	6,89	7,07
ENL Individuel	6,32	6,48	6,63
EXP Collectif	8,50	8,62	8,74
ENL Collectif	7,86	7,96	8,07

TABLE B.8 – Comparaison Exp / ENL : nature du logement

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
EXP Parc privé	9,64	9,81	9,98
ENL Parc privé	8,86	8,99	9,13
EXP Parc social	6,10	6,22	6,33
ENL Parc social	5,54	5,62	5,70

TABLE B.9 – Comparaison Exp / ENL : privé ou social

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
EXP Rural	5,90	6,09	6,27
ENL Rural	5,35	5,54	5,74
EXP UU<100 000 hab	6,55	6,73	6,91
ENL UU<100 000 hab	6,17	6,31	6,46
EXP UU>100 000 hab	7,93	8,08	8,23
ENL UU>100 000 hab	7,42	7,56	7,69
EXP UU Paris	11,19	11,48	11,78
ENL UU Paris	10,55	10,69	10,82

TABLE B.10 – Comparaison Exp / ENL : taille d'unité urbaine

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
EXP Île-de-France	11,06	11,34	11,62
ENL Île-de-France	10,39	10,65	10,90
EXP Nord-Pas-de-Calais	6,65	6,94	7,22
ENL Nord-Pas-de-Calais	6,44	6,60	6,76

TABLE B.11 – Comparaison Exp / ENL : estimations locales

	Borne inf IC 95 %	Ratio	Borne sup IC 95 %
EXP Avant 2010	7,28	7,42	7,55
ENL Avant 2010	6,75	6,86	6,96
EXP 2010 et après	8,91	9,09	9,26
ENL 2010 et après	8,12	8,24	8,42

TABLE B.12 – Comparaison Exp / ENL : date d'emménagement

Annexe C

Effet du recours aux documents externes dans l'expérimentation

Dans l'expérimentation Logement, environ un locataire sur deux déclare avoir recours à sa quittance ou à son bail pour remplir le questionnaire. Une question qui se pose est celle de l'effet du non-recours sur le montant de loyer. S'il est possible que, dans certains cas, les personnes connaissent parfaitement leur montant de loyer, la qualité de l'information recueillie devrait être néanmoins globalement de moins bonne qualité lorsque les personnes ne prennent pas le temps de rechercher leurs documents pour répondre aux questions. Ainsi, les personnes qui n'ont pas recours aux documents arrondissent beaucoup plus fréquemment leur montant de loyer reporté que celles qui ont utilisé leur quittance ou leur bail (voir chapitre 2).

Cela peut-il avoir des effets sur les estimations des moyennes de montants de loyer ? Et, plus précisément, le non-recours aux documents a-t-il pour conséquence une estimation à la hausse ou à la baisse du montant de loyer ? Pour répondre à cette question, 3 régressions (pondérées) différentes du loyer sur un ensemble d'explicatives, dont le recours aux documents, sont réalisées : l'une sur l'ensemble des observations issues de l'expérimentation Logement, une autre sur les réponses web et une dernière sur les réponses papier.

$$Y_i = \alpha_1 + \delta \mathbb{1}_{[support=WEB]i} + \gamma_1 X_i + \phi_1 Z_i + \epsilon_{1i} \quad \text{(EXP - ensemble)}$$

$$Y_i = \alpha_2 + \gamma_2 X_i + \phi_2 Z_i + \epsilon_{2i} \quad \text{(EXP - web)}$$

$$Y_i = \alpha_3 + \gamma_3 X_i + \phi_3 Z_i + \epsilon_{3i} \quad \text{(EXP - papier)}$$

Y : loyer redressé, X_i : variables explicatives, Z_i : non-recours aux documents, ϵ_i : terme d'erreur

Au vu des résultats des trois régressions, le non-recours au document, s'il n'est pas le déterminant le plus significatif du montant de loyer [1], semble avoir un effet positif sur ce dernier. Les estimations de ϕ_1, ϕ_2 et ϕ_3 sont respectivement de +13, +16 et +12 euros.

Sans détailler les sorties des régressions, un point que l'on peut préciser est que le paramètre δ de la première régression, qui est associé à l'indicatrice de réponse par internet, est significativement non nul, et estimé à 22 euros. Ce résultat est plutôt contraire à la théorie qui affirme que les deux modes auto-administrés que sont internet et le papier ne devraient guère différer dans la mesure. Rappelons tout de même que les populations de répondants web et papier sont très différentes :

1. Les p-values sont pour les paramètres ϕ_1, ϕ_2 et ϕ_3 respectivement de 0,11, 0,06 et 0,38.

les écarts bruts sur le loyer médian sont de 75 euros (voir chapitre 2) et les écarts après traitements (non différenciés) sont de 85 euros (voir chapitre 4). Les variables de contrôle de la sélection jouent leur rôle en réduisant substantiellement les écarts mais il reste néanmoins un écart qu'il est difficile d'interpéter : s'agit-il d'un réel effet de mode lié à la mesure ou est-ce que des caractéristiques inobservables sont corrélées à ce paramètre ?

Annexe D

Effets des déterminants du loyer dans l'ENL et dans l'expérimentation

Une question que l'on peut se poser est de savoir si, en dépit des écarts constatés dans les niveaux de loyer (26 euros d'écart, après contrôle de la sélection, sur les moyenne) entre ENL et expérimentation, les déterminants du loyer ont des effets du même ordre.

Ainsi, nous effectuons deux régressions (pondérées) sur le périmètre du logement collectif : l'une sur les données de l'ENL, l'autre sur celles de l'expérimentation, avec le même jeu de variables X.

$$Y_i = \alpha_1 + \delta_1 X_i + \epsilon_{1i} \quad \text{(ENL)}$$

$$Y_i = \alpha_2 + \delta_2 X_i + \epsilon_{2i} \quad \text{(EXP)}$$

Y : loyer redressé, X_i : variables explicatives, ϵ_i : terme d'erreur

Les variables explicatives sont celles qui sont utilisées également dans le chapitre 5, à savoir :
— la surface du logement ;
— la région ou le département pour l'IDF et le NPdC ;
— le taille d'unité urbaine ;
— le nombre d'habitants du logement ;
— le nombre de pièces du logement ;
— le type de propriétaire ;
— la date d'achèvement ;
— l'étage du logement ;
— la présence d'ascenseur ;
— la date d'emménagement ;
— l'entité à laquelle le loyer est versé ;
— le diplôme [1].

Les coefficients estimés des variables les plus significatives (au sens de la p-value) sont les suivants dans les deux enquêtes [2] :

1. Le diplôme du répondant dans l'expérimentation, de la personne de référence dans l'ENL
2. Le coefficient de détermination R^2 est un peu plus élevé pour l'ENL (0,54) que pour l'expérimentation (0,45).

	EXP	EXP IC	ENL	ENL IC
Surface	7,68	[4,32 ;11,05]	5,12	[3,82 ;6,41]
75 vs (Rhône-Alpes)	381,14	[337,99 ;424,29]	338,99	[296,78 ;381,20]
77 vs (Rhône-Alpes)	100,53	[62,543 ;138,52]	77,48	[53,05 ;101,90]
78 vs (Rhône-Alpes)	116,56	[82,90 ;150,21]	121,49	[94,18 ;148,80]
91 vs (Rhône-Alpes)	83,404	[47,89 ;118,92]	94,66	[71,97 ;117,34]
92 vs (Rhône-Alpes)	221,31	[181,99 ;260,63]	181,57	[153,64 ;209,50]
93 vs (Rhône-Alpes)	105,23	[73,49 ;136,96]	123,82	[101,36 ;146,27]
94 vs (Rhône-Alpes)	123,79	[91,21 ;156,37]	138,04	[110,96 ;165,11]
95 vs (Rhône-Alpes)	90,93	[55,93 ;125,93]	85,44	[59,96 ;110,93]
PACA vs (Rhône-Alpes)	61,72	[37,03 ;86,41]	58,18	[40,32 ;76,04]
Rural (vs UU>200 000h)	-121,32	[-161,15 ;-81,49]	-60,39	[-86,44 ;-34,34]
2 000-4 999 h (vs UU>200 000h)	-92,77	[-124,98 ;-60,57]	-50,38	[-80,68 ;-20,08]
5 000-9 999 h (vs UU>200 000h)	-78,73	[-111,52 ;-45,94]	-62,32	[-86,10 ;-38,52]
20 000-49 999 h (vs UU>200 000h)	-68,79	[-96,93 ;-40,64]	-52,47	[-68,10 ;-36,84]
50 000-99 999 h (vs UU>200 000h)	-62,24	[-85,24 ;-39,24]	-28,42	[-40,66 ;- 16,17]
Parc privé (vs social)	195,52	[177,90 ;213,13]	212,49	[200,31 ;224,67]
Emménagement années 1980 et avant (vs 2013)	-135,73	[-167,18 ;-104,28]	-88,66	[-109,99 ;-67,33]
Emménagement années 1990 (vs 2013)	-77,81	[-105,3 ;-50,32]	-30,98	[-49,97 ;-11,98]
Emménagement années 2000 (vs 2013)	-39,77	[-57,76 ;-21,78]	-20,31	[-33,80 ;-6,81]
Versement à un professionnel (vs autres)	44,03	[26,59 ;61,48]	29,32	[16,92 ;41,71]
Aucun diplôme (vs > Bac +2)	-70,92	[-105,07 ;-36,78]	-93,98	[-108,96 ;-79,00]
CEP, Brevet (vs > Bac +2)	-92,89	[-118,67 ;-67,11]	-95,21	[-112,10 ;-78,32]
CAP BEP (vs > Bac +2)	-80,70	[-99,97 ;-61,43]	-94,59	[-109,53 ;-79,66]
Bac (vs > Bac +2)	-60,10	[-78,19 ;-42,00]	-70,97	[-87,13 ;-54,81]
Bac +2 (vs > Bac +2)	-49,25	[-69,27 ;-29,24]	-63,72	[-82,60 ;-44,83]

TABLE D.1 – Comparaison des déterminants du loyer

Globalement, les coefficients estimés sur les deux enquêtes sont du même ordre pour ces variables : tous les intervalles de confiance se chevauchent. Toutefois, les écarts s'ils ne sont pas significatifs au sens statistique du terme sont parfois conséquents. Par exemple, pour le rural (vs habiter dans une unité urbaine de + de 200 000 habitants), l'effet négatif sur le loyer passe de -60 euros pour l'ENL à -120 euros pour l'expérimentation.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : 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).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : 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) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS – K.MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS – K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR – R. LE SAOUT B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD – M. GUILLERM

M2016/06 : Économétrie spatiale : une introduction pratique
R. LE SAOUT – J-M. FLOCH

M2016/07 : La gestion de la confidentialité pour les données personnelles
M. BERGEAT