

Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité : application aux enquêtes annuelles de recensement

Document de travail

N° M2024-03 – Septembre 2024



M 2024/03

**Discuter l'existence d'un effet de sélection dans un
cadre multimode grâce à une analyse de sensibilité :
Application aux enquêtes annuelles de recensement**

**Loreline COURT
Simon QUANTIN**

Insee

SEPTEMBRE 2024

Remerciements : nous remercions Hélène Chaput, Patrick Sillard, Corinne Prost, Sébastien Hallépée, Christel Colin, Muriel Barlet, Julie Djiriguiian pour leur relecture attentive du document, ainsi que Jeanne Pages et Blandine Legendre pour leur contribution à ces travaux



Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques - Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : <http://www.insee.fr>

*Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.
Working papers do not reflect the position of INSEE but only their author's views.*

Résumé

Cette étude détaille l'approche par analyse de sensibilité qui permet de discuter de la pertinence des conclusions sur l'existence d'un effet du traitement sur une variable binaire quand on suppose qu'il subsiste un biais de composition caché après appariement entre unités traitées et non traitées. Cette méthode est appliquée pour expliciter l'évolution de l'écart entre les estimations d'emploi obtenues à partir des enquêtes annuelles de recensement (EAR) et celles issues des estimations annuelles d'emploi (EAE) suite à l'introduction, en 2015, de la collecte par internet dans le recensement. En effet, depuis cette date, une baisse de la participation est observée, ainsi qu'un écart accru entre les deux estimations d'emploi.

Nos résultats mettent en évidence un impact du mode de collecte sur le type de non-réponse : la collecte par internet accroîtrait la non-réponse totale aux questionnaires individuels des logements enquêtés alors qu'elle réduirait la non-réponse partielle à la question sur la situation principale vis-à-vis de l'emploi. Cet effet fragiliserait les pratiques de correction de la non-réponse se basant sur des imputations et telles qu'implémentées pour les EAR. En effet, la non-réponse totale aux questionnaires individuels est corrigée à partir des taux d'emploi des répondants de même sexe et de même tranche d'âge, mais sans tenir compte du mode de collecte. De plus, une telle correction de la non-réponse ignore aussi la possibilité que les non-répondants à un mode de collecte soient plus fréquemment en emploi que les répondants au même mode. Nos simulations suggèrent à l'inverse qu'une telle hypothèse serait à même d'expliquer l'accroissement de l'écart entre les estimations d'emploi à partir des EAR et celles issues des EAE, compte tenu notamment de l'accroissement progressif de la part des questionnaires du recensement collectés par internet.

Mots-clé : enquête multimode, évaluation, analyse de sensibilité

Classification JEL: C12, C14, C21, C83

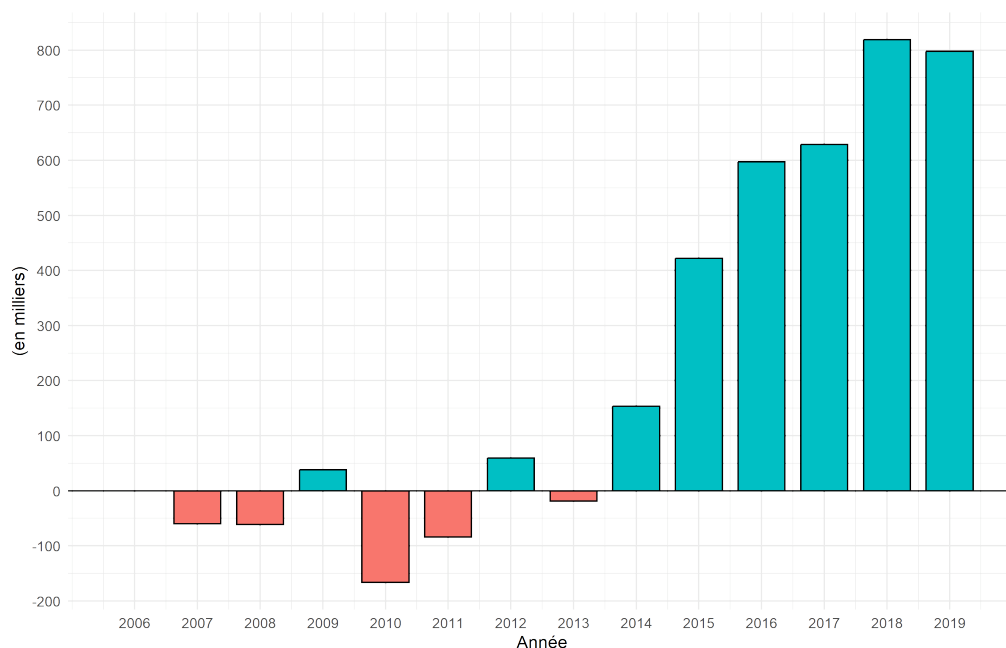
Table des matières

1	Introduction	2
2	L'impact global du multimode sur l'estimation du nombre de personnes en emploi dans les EAR	3
3	Collecte multimode et non-réponse dans l'enquête annuelle de recensement	5
3.1	Organisation du recensement, questionnaires logement et individuel	5
3.2	La question de l'emploi dans les enquêtes annuelles de recensement	5
3.3	Différents types de non-réponse partielle	7
3.4	La collecte par internet dans les enquêtes annuelles de recensement	7
4	Un impact du mode de collecte sur le type de non-réponse ?	8
4.1	Choix du mode de collecte et biais de composition caché	9
4.2	Le modèle d'analyse de sensibilité	10
4.3	Résultats	16
5	Un effet de sélection endogène lié à la situation vis-à-vis de l'emploi ?	20
6	Conclusion	22
A	Assignation à un mode de collecte avec des probabilités inconnues par strates sous l'hypothèse d'absence de biais caché	26
B	Résultats additionnels	27
C	Taux d'emploi des EAR par mode de collecte sans et avec imputation	29

1 Introduction

Les estimations d'emploi ont pour objectif de comptabiliser le nombre de personnes en emploi, tel que répertorié dans les sources administratives : les estimations annuelles d'emploi (EAE) issues de la synthèse de plusieurs sources administratives constituent ainsi la source de référence pour le suivi conjoncturel de l'emploi, tant en niveau qu'en évolution (et par échelon géographique national, régional, départemental et zone d'emploi). Elles s'appuient sur les déclarations sociales réalisées par les employeurs et les travailleurs indépendants. Tout emploi déclaré en France dans les données administratives y est donc comptabilisé¹. Les enquêtes annuelles de recensement (EAR) peuvent aussi être utilisées pour obtenir une estimation du nombre total de personnes en emploi puisque des questions relatives à la situation personnelle vis-à-vis de l'emploi sont posées à toute personne enquêtée de 14 ans ou plus. Le nombre d'emplois mesuré par le recensement de la population peut néanmoins différer de celui fourni par les EAE, du fait de différences de champ, de concept et de source². Ainsi, alors que les estimations d'emploi issues des sources administratives comptabilisent l'ensemble des emplois exercés sur le territoire français (concept d'emploi mesuré au lieu de travail), les enquêtes annuelles de recensement recensent les emplois des personnes résidant effectivement en France. Celles-ci travaillent éventuellement dans un pays limitrophe (concept d'emploi mesuré au lieu de résidence).

GRAPHIQUE 1 – Évolution de l'écart entre l'emploi issu des enquêtes annuelles de recensement (pondérées) et les estimations annuelles d'emploi



Note : l'écart constaté en 2006 est la référence.
Champ : France métropolitaine.

Les estimations de l'emploi fournies par les EAR sont plus faibles que les EAE. Entre 2006 et 2014, les EAR recensent ainsi, chaque année, 700 000 emplois de moins. Cependant, depuis 2015, cet écart, jusqu'alors relativement constant, n'a cessé d'augmenter (cf. graphique 1). L'augmentation de cet écart entre les deux sources coïncide avec le développement de la collecte par internet à partir de l'EAR 2015. Il est alors intéressant de s'attacher à évaluer l'impact de la modification des protocoles de collecte des EAR.

1. voir <https://www.insee.fr/fr/metadonnees/source/serie/s1283> pour la définition détaillée.

2. voir <https://www.insee.fr/fr/metadonnees/source/fichier/Emploi-quelle-source-pour-quel-usage.pdf> pour toutes les différences entre les sources.

Ce document de travail a pour objectif d'investiguer les liens entre l'augmentation de l'écart constaté entre les estimations d'emploi issues des enquêtes annuelles de recensement et celles des EAE, et l'introduction d'une collecte multimode. Pour cela, nous mettons en évidence, dans un premier temps, l'existence d'un effet causal de « l'introduction d'une collecte multimode », au sens large, sur les estimations d'emploi issues des EAR. Pour comprendre les mécanismes inhérents à cet effet global, nous analysons dans une deuxième partie l'impact du mode de collecte sur la non-réponse aux EAR. Pour cela, nous utilisons une méthode d'analyse de sensibilité de façon originale. Nos résultats suggèrent que la collecte par internet conduirait à accroître la non-réponse totale au questionnaire individuel associé aux EAR par rapport au questionnaire papier, mais réduirait la non-réponse partielle à la question sur la situation vis-à-vis de l'emploi. Ce constat différencié par type de non-réponse nous conduit à questionner, dans une troisième partie, la méthode de correction de la non-réponse par imputation actuellement pratiquée dans les enquêtes annuelles de recensement. Nos simulations de différentes méthodes de correction de la non-réponse par imputation révèlent que l'écart constaté entre les deux estimations d'emploi reposent essentiellement sur un effet de sélection sur inobservable. En particulier, la proportion accrue de non-répondants en emplois, notamment par internet, n'est pas prise en compte. La correction de la non-réponse totale au questionnaire individuel est indépendante du mode de collecte. Ainsi la variable d'emploi est corrigée en appliquant simplement les taux d'emplois par sexe et par âge des répondants sans tenir compte du fait que les répondants par internet sont en moyenne davantage en emploi.

2 L'impact global du multimode sur l'estimation du nombre de personnes en emploi dans les EAR

Entre 2006 et 2014, les estimations annuelles d'emploi diffèrent de manière constante de celles issues des enquêtes annuelles de recensement. Cependant, depuis 2015, année de l'introduction d'un processus de collecte multimode, l'écart entre les deux estimations n'a cessé de croître, tout comme le développement de la collecte par internet qui ne concernait que 33 % des logements enquêtés en 2015 pour atteindre 55 % en 2019. Peut-on conclure à un effet causal de l'introduction du processus de collecte multimode sur les estimations d'emploi issues des EAR ?

Pour répondre à cette question, nous adoptons une stratégie de différence de différences qui apparaît adéquate, sous l'hypothèse où les EAE et les estimations issues des EAR auraient conservé une évolution identique en l'absence d'un changement du mode de collecte (hypothèse dite de « tendance commune »). Il faut noter cependant que la période 2015-2019 est aussi une période de très forte croissance de l'emploi, après quelques années de relative stabilité. Il est possible que d'autres phénomènes économiques ayant des effets différenciés sur les concepts respectifs des EAR et des estimations d'emploi aient contribué à l'écart ; néanmoins, il faudrait une analyse plus précise des écarts de concepts qui n'est pas l'objet de cette étude.

TABLEAU 1 – Résultats de l'estimation par différences de différences (DiD)

	EAR	EAE	Différence
Emploi moyen entre 2006 et 2014	26 175 000 (77 000)	26 861 000 (87 000)	686 000 (116 000)
Emploi moyen entre 2015 et 2019	26 267 000 (168 000)	27 621 000 (104 000)	1 354 000 (197 000)
Différence	92 000 (130 000)	760 000 (189 000)	668 000 (221 000)

Note : erreurs standards entre parenthèses.

L'introduction d'une collecte multimode concurrentiel aurait conduit à réduire de 670 000 emplois (en moyenne entre 2015 et 2019) les estimations obtenues à partir des EAR, conduisant à accroître l'écart entre les deux estimations (cf. tableau 1). Il y aurait donc un effet de la collecte multimode, qui peut introduire plusieurs sources de biais qu'il faut analyser.

Comme toutes les enquêtes, les résultats des enquêtes annuelles de recensement peuvent être impactés par un effet de sélection dû à la non-réponse ; celle-ci peut aboutir à une composition de l'échantillon des répondants différente de la population totale, de telle sorte que cet échantillon n'est plus représentatif. Cette non-participation peut être liée à des caractéristiques observées (sexe, âge, catégorie socioprofessionnelle, etc.) mais aussi inobservées, de telle sorte qu'il convient de distinguer la sélection dite *sur observables* et celle *sur inobservables* (on parle aussi dans ce dernier cas de non-réponse non-ignorable, NMAR pour *Not Missing at Random*). Cet effet de sélection est commun à toutes les enquêtes, monomode ou multimode. Il est particulièrement problématique si ces variables inobservables expliquant la non réponse sont reliées aux paramètres que l'on cherche à estimer dans l'enquête.

Parallèlement à un effet sur la composition des répondants, un mode de collecte alternatif (ici la collecte internet par rapport à une collecte papier) peut aussi avoir un effet de mesure sur la réponse donnée par le répondant. Celui-ci se définit comme le fait qu'un même individu réponde différemment à une même question selon le mode de collecte. Cela peut refléter, par exemple, un effet de la présence de l'enquêteur dans les collectes intermédiées (face-à-face, téléphone), notamment lorsque la thématique de l'enquête est sensible (désirabilité sociale) ou encore sur internet et par téléphone quand le répondant a des difficultés face à une longue liste de réponses, par exemple.

Enfin, dans les estimations d'emploi issues des enquêtes annuelles de recensement, il existe aussi un autre effet lié à la correction de la non-réponse effectuée avant chaque estimation. Schématiquement, le redressement de la non-réponse dépend de la disponibilité ou non du questionnaire individuel. En cas de non-réponse à la question sur la situation vis-à-vis de l'emploi, celle-ci est éventuellement imputée au niveau individuel, à partir des réponses fournies à d'autres questions utilisées comme autant d'indices de travail (adresse de l'employeur, etc.). En leur absence ou dans le cas d'un questionnaire individuel non rempli, la situation en emploi est imputée par une méthode d'imputation *hot-deck*, c'est-à-dire à partir des réponses données par des enquêtés ayant répondu, en tenant compte de l'âge et du sexe de l'enquêté non-répondant mais pas du mode de collecte choisi³. On notera qu'une telle procédure suppose que l'effet de sélection est de moyenne nulle conditionnellement aux

3. L'imputation par *hot-deck* est une procédure qui consiste à remplacer les réponses manquantes par les réponses d'autres répondants présentant un certain nombre de caractéristiques communes.

caractéristiques retenues pour constituer les classes d'imputation. Dès lors, elle suppose que dans chaque classe, i.e. à caractéristiques observées retenues, les répondants sont similaires aux non-répondants vis-à-vis de leur situation en emploi. La correction par imputation ne retient donc pas une différence de composition éventuelle au regard de la situation en emploi entre répondants et non-répondants, i.e. la possibilité d'un effet de sélection sur inobservable y compris à mode de collecte donné.

L'effet causal de « l'introduction du multimode concurrentiel » mis en évidence précédemment englobe donc (i) l'impact d'un potentiel effet de sélection y compris sur inobservable, (ii) celui d'un effet de mesure de la collecte par internet sur la réponse à la question relative à la situation vis-à-vis de l'emploi et (iii) en lien avec les deux premiers effets, un effet de l'ajustement effectué en cas de non-réponse. Dans ce document de travail, nous considérons l'éventualité d'un effet de sélection sur inobservable et l'impact sur le redressement de la non-réponse. Pour cela, nous interrogerons dans un premier temps l'existence d'un effet du mode de collecte choisi sur la non-réponse totale et sur la non-réponse partielle. Puis, en simulant différentes méthodes de correction de la non-réponse, nous chercherons à mettre en évidence l'ampleur potentielle de la sélection endogène et son impact sur les écarts constatés entre les EAE et les EAR.

3 Collecte multimode et non-réponse dans l'enquête annuelle de recensement

3.1 Organisation du recensement, questionnaires logement et individuel

En France, depuis 2004, le recensement repose sur une collecte annuelle d'informations, couvrant successivement tous les territoires communaux sur une période de cinq ans. Les communes de moins de 10 000 habitants réalisent une enquête de recensement portant sur toute la population, à raison d'une commune sur cinq chaque année. Les communes de 10 000 habitants ou plus réalisent tous les ans une enquête par sondage auprès d'un échantillon d'adresses représentant 8 % de leurs logements. Ainsi, chaque année, une enquête annuelle de recensement (EAR) est réalisée auprès d'un *échantillon de logements*, auquel chaque habitant doit répondre. Grâce à ce schéma, la France est couverte sur un cycle quinquennal, exhaustivement pour les communes de moins de 10 000 habitants et avec un échantillon de 40 % de la population communale pour celles de plus de 10 000 habitants.

Pour chaque logement, l'enquête annuelle de recensement comporte au moins deux questionnaires à remplir, un questionnaire logement et un ou des questionnaires individuels. Le questionnaire logement permet de décrire les caractéristiques du logement (type, année d'achèvement, taille, statut d'occupation, logement social, etc.) et du ménage y résidant (composition, taille, liens familiaux, etc.). Le questionnaire individuel, lui, s'adresse à tous les habitants du logement et porte sur ses caractéristiques sociodémographiques : sexe, âge, état matrimonial, lieu de naissance, nationalité, lieu d'études, diplômes. Pour les personnes de 14 ans ou plus, des questions sur leur situation vis-à-vis de l'emploi, leur catégorie socio-professionnelle, etc. sont aussi posées⁴.

3.2 La question de l'emploi dans les enquêtes annuelles de recensement

Deux questions dans le questionnaire individuel des enquêtes annuelles de recensement permettent de déterminer si un enquêté est en emploi ou non. La première question porte sur la « situation principale vis-à-vis de l'emploi » et distingue sept modalités de réponse.

4. Les détails sur le traitement statistique du recensement français sont disponibles à l'adresse suivante <https://www.insee.fr/en/metadonnees/source/operation/s2091/processus-statistique>

Est considérée en emploi une personne se déclarant en emploi salarié, à son propre compte, en apprentissage ou en stage rémunéré. Cependant, si la personne enquêtée n'a pas répondu être en emploi ou en apprentissage à la question précédente, la question « Travaillez-vous actuellement ? » est posée. Une réponse positive à cette nouvelle question conduit aussi à considérer la personne enquêtée comme une personne en emploi.

Pour le questionnaire papier (cf. image 1) jusqu'à l'EAR 2024, les questions se présentent comme suit :

Image 1 – Questionnaire papier - questions liées à l'emploi

11 Quelle est votre situation principale ?
Ne cochez qu'une seule case.

- **Emploi (salarié ou à votre compte, y compris aide d'une personne dans son travail)**
☐ cochez puis passez en **18** ☐ 1
- **Apprentissage sous contrat ou stage rémunéré**
☐ cochez puis passez en **18** ☐ 2
- **Études (élève, étudiant) ou stage non rémunéré** ☐ 3
- **Chômage (inscrit ou non au pôle emploi)** ☐ 4
- **Retraite ou préretraite**
 (ancien salarié ou ancien indépendant) ☐ 5
- **Femme ou homme au foyer** ☐ 6
- **Autre situation** ☐ 7

12 Travaillez-vous actuellement ?
Si vous avez un emploi occasionnel ou de très courte durée, ou si vous êtes en apprentissage ou en stage rémunéré, cochez « Oui ».
Si vous êtes en congé maladie ou de maternité, cochez « Oui ».

- **Oui** ☐ cochez puis passez en **18** ☐ 1
- **Non** ☐ cochez puis passez en **13** ☐ 2

Pour le recensement en ligne, les intitulés de questions et modalités de réponses sont identiques, mais la visualisation est différente. Pour les questionnaires en ligne, usuellement, l'affichage d'une seule question par page est conseillé pour des raisons d'attention et de lecture [Koumarianos and Schreiber, 2021]. Cependant, de la même façon que les questionnaires papier, si le répondant coche une modalité autre que « Emploi » ou « Apprentissage », la question « Travaillez-vous actuellement ? » est posée. Enfin, il est aussi possible de ne pas répondre à une question et de passer à la question suivante.

Image 2 – Questionnaire web (version des EAR 2015 à 2023) - question sur la situation principale

Bulletin individuel : Pif T 55%

10 - Quelle est votre situation principale ?

Merci d'indiquer quelle est votre situation principale.

- ☐ **Emploi (salarié ou à votre compte, y compris aide d'une personne dans son travail)**
- ☐ **Apprentissage sous contrat ou stage rémunéré**
- ☐ **Études (élève, étudiant) ou stage non rémunéré**
- ☐ **Chômage (inscrit ou non à Pôle emploi)**
- ☐ **Retraite ou préretraite (ancien salarié ou ancien indépendant)**
- ☐ **Femme ou homme au foyer**
- ☐ **Autre situation**

[Compléter plus tard](#)

3.3 Différents types de non-réponse partielle

La concomitance de plusieurs questionnaires conduit à observer différents schémas de non-réponse. Tout d’abord, la non-réponse totale dans l’EAR se produit lorsque tous les habitants d’un logement échantillonné ne répondent pas. Dans ce cas, aucun formulaire n’est rempli et aucune information n’est disponible, à part éventuellement le nombre d’occupants du logement (si cette information a pu être recueillie auprès du voisinage). Cela ne se produit que si les occupants du logement sont absents de longue durée ou impossible à joindre durant toute la période de collecte, de sorte que la non-réponse totale sera considérée comme manquante au hasard (*Missing-At-Random*) et ignorée pour cette étude. Ce type de non-réponse concerne environ 4 % des logements. La non-réponse partielle, quant à elle, englobe deux raisons principales de non-observation. Premièrement, un habitant peut choisir de ne pas répondre à l’ensemble du questionnaire individuel (ci-après dénommé « non-réponse individuelle »)⁵) : dans ce cas, on dispose au moins de (certaines) caractéristiques du logement, mais aussi de l’âge et du sexe de chaque habitant, informations collectées grâce au questionnaire logement. Deuxièmement, un habitant répond au questionnaire individuel mais de manière incomplète : il manque au moins les réponses à certaines questions (ci-après dénommé « non-réponse partielle à une question »).

3.4 La collecte par internet dans les enquêtes annuelles de recensement

Depuis 2015, une collecte multimode concurrentiel a été mise en place : chaque logement échantillonné a le choix entre le mode papier et le mode internet. Cependant, dans chaque logement, *tous les habitants doivent répondre avec le même mode*. En pratique, celui-ci est donc choisi par la personne répondant au questionnaire logement et s’impose à tous les habitants par la suite. Comme l’introduction de la collecte de données en ligne est souvent associée à un taux de réponse plus faible par internet [Daikeler et al., 2020] ou à une non-réponse partielle plus importante que pour les enquêtes en face-à-face [Heerwegh and Loosveldt, 2008], plusieurs incitations à répondre ont été mises en place dans le formulaire individuel en ligne, bien que les questions posées restent identiques à celles du formulaire papier. Par exemple, lorsqu’un habitant souhaite passer une question sans répondre, un message rouge apparaît pour lui rappeler de répondre. À l’inverse, l’application internet développée permet, à la fin du questionnaire logement, d’envoyer sa réponse à l’enquête sans remplir (tous) les formulaires individuels associés (cf. image 3). Cette validation trop rapide n’est plus possible depuis la refonte de l’ergonomie du questionnaire internet apportée en 2024. Si le répondant envoie sa réponse sans avoir renseigné un des questionnaires individuels, un message explicite et très visible précisant le nom correspondant aux questionnaires qui n’ont pas été ouverts est affiché.

5. Nous choisissons de ne pas utiliser dans cet article le terme usuel de « non-réponse de l’unité enquêtée » car c’est le logement et non ses habitants qui est l’unité échantillonnée.

Image 3 – Questionnaire internet (version des EAR 2015 à 2023) - illustration de la page internet à la fin du questionnaire logement

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Recensement de la population

Bilan des réponses aux questionnaires Recensement de la population

PROGRESSION

Questions sur le logement

Questions sur :
- les habitants
- leurs liens
- leurs lieux d'habitation

Questions individuelles

Envoi du questionnaire

Bilan des réponses aux questionnaires

Attention vos questionnaires n'ont pas encore été envoyés. Merci de les compléter.

Merci de compléter les questionnaires des personnes concernées.

Rang	Nom	Prénom	Qui doit remplir un bulletin individuel
1	T	Pif	Finir de remplir le questionnaire de Pif T
2	C	Paf	Remplir le questionnaire de Paf C
3	G	Pof	Remplir le questionnaire de Pof G

Questions sur le logement

Votre questionnaire Logement a été entièrement renseigné. [Pour le modifier, cliquez ici.](#)

Questions sur le ménage

Votre questionnaire Ménage a été entièrement renseigné. [Pour le modifier, cliquez ici.](#)

[Compléter plus tard](#)

[Envoyer les questionnaires](#)

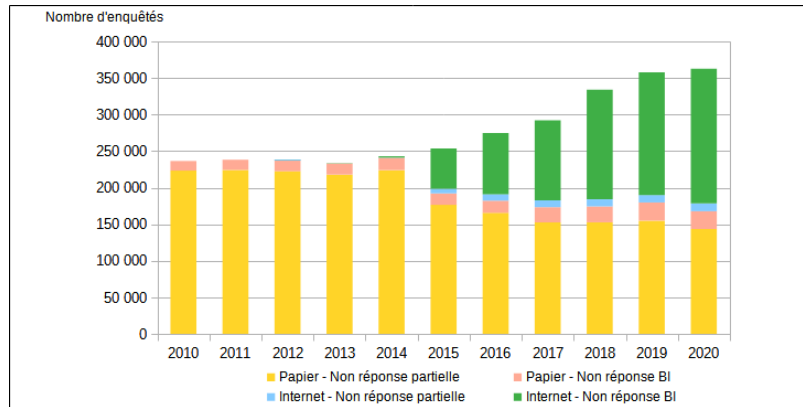
La participation au niveau individuel est donc potentiellement corrélée entre individus d'un même logement. À l'inverse, dans le cadre d'une collecte papier, les agents recenseurs vérifient que tous les formulaires papier ont été (au moins partiellement) complétés avant de les accepter. Au final par rapport à une collecte papier et en schématisant, répondre par internet pourrait réduire la non-réponse partielle mais induirait de la non-réponse totale.

4 Un impact du mode de collecte sur le type de non-réponse ?

Dans ce document de travail, nous nous concentrons sur la question relative à la situation principale vis-à-vis de l'emploi, qui est la principale question utilisée pour déterminer le nombre d'emplois à partir des enquêtes annuelles de recensement. L'absence de réponse à cette question peut refléter une non-réponse individuelle qui résulte de l'absence de questionnaire individuel, ou une non-réponse partielle qui résulte de l'existence d'un questionnaire individuel mais sans réponse à la question sur la situation principale vis-à-vis de l'emploi.

Dans les enquêtes annuelles de recensement, la non-réponse à cette question a augmenté depuis le début de la collecte multimode en 2015. Par ailleurs, le type de non-réponse associé à cette question est corrélé au mode de collecte choisi au sein du logement (graphique 2).

GRAPHIQUE 2 – Différents types de non-réponse à la question sur l'emploi



Ainsi, depuis l'introduction du multimode, le nombre d'enquêtés non-répondants à cette question est passé de 250 000 à plus de 350 000⁶, tandis que la proportion de logements répondant par internet passait de 33 % en 2015 à 53 % en 2019. Cela correspond à un taux de non-réponse passant de 3,5 % à 4,9 % entre 2015 et 2019. La non-réponse individuelle (tout le questionnaire individuel est vide) est presque le seul type de non-réponse des habitants enquêtés par internet. À l'inverse, la non-réponse partielle à la question sur l'emploi est presque le seul type de non-réponse des personnes enquêtées via le questionnaire papier.

4.1 Choix du mode de collecte et biais de composition caché

Deux hypothèses peuvent expliquer le lien entre le multimode et l'augmentation de la non-réponse. La première serait un impact du questionnaire internet sur la non-réponse : celui-ci impacterait la participation à l'enquête, par exemple car les incitations implémentées sur le site réduiraient la probabilité de non-réponse partielle des enquêtés. Cette première hypothèse pourrait expliquer pourquoi la non réponse partielle est surtout visible sur le questionnaire papier. Cette différence risque d'être fortement atténuée à partir de l'EAR 2025 puisque pour des raisons d'accessibilité et pour suivre les recommandations du mode de passation de questionnaire en ligne, ces relances ne seront plus formulées. La seconde hypothèse est celle d'un biais de composition : les enquêtés qui choisissent la collecte par internet diffèrent des enquêtés qui choisissent la collecte par papier, au regard de leur comportement de réponse, sans que la collecte par internet ne réduise ou accroisse leur probabilité de non-réponse. Autrement dit, les personnes qui choisissent de répondre par internet au questionnaire logement auraient, indépendamment du mode de collecte, une probabilité moindre de répondre - ensuite - aux questionnaires individuels que ceux qui choisissent le questionnaire papier, en raison par exemple d'un intérêt faible pour les thématiques abordées lors du recensement. Arbitrer entre ces deux hypothèses revient à tester l'hypothèse nulle (H_0) de l'absence d'effet causal du questionnaire internet sur la non-réponse.

La collecte par internet a, pour un enquêté, un effet sur sa participation si celle-ci diffère en présence d'un questionnaire web et d'un questionnaire papier. La participation éventuelle d'un même enquêté sous les deux modes de collecte disponibles pour le recensement à partir de 2015 n'est cependant pas connue, puisque seule la participation éventuelle à un des deux mode de collecte est observée. La mise en évidence éventuelle d'un effet causal repose dès lors sur la comparaison de la participation entre des enquêtés identiques mais

6. En utilisant les poids d'échantillonnage, les non-répondants représentent environ 1,5 million de non-réponses.

ayant choisi des modes de collecte différents.

Dans le cas du recensement, i.e. en l'absence d'une affectation aléatoire à un mode de collecte (puisque ce mode est choisi par l'enquête), les unités enquêtées par internet et sur papier peuvent ne pas être comparables au départ, avant de répondre à l'enquête. Dans les EARS, l'unité enquêtée est le logement et des différences préalables existent entre les logements dont les habitants ont choisi de répondre par internet et ceux dont les habitants ont choisi de répondre par papier (cf. tableau 2).

TABLEAU 2 – Mode de collecte choisi au sein des logements enquêtés selon quelques caractéristiques sociodémographiques

	Papier (%)	Internet (%)
Composition démographique du logement		
Avec au moins une personne de moins de 14 ans	19	36
Avec au moins une personne entre 14 et 24 ans	18	28
Nombre de personnes entre 25 et 59 ans		
Seulement une	28	27
Deux ou plus	29	54
Avec au moins une personne d'au moins 60 ans	49	21
Taille du logement		
Moins de 60 m ²	25	16
Entre 60 et 100 m ²	45	39
Plus de 100 m ²	30	45
Statut d'occupation		
Propriétaire	62	66
Locataire	38	34
HLM		
Oui	17	9
Non	83	91
Unité urbaine		
Rural	32	31
Unité urbaine de 2 000 à 9 999 hab.	17	17
Unité urbaine de 10 000 à 49 999 hab.	12	10
Unité urbaine de 50 000 à 99 999 hab.	13	10
Unité urbaine de 200 000 à 1 999 999 hab.	18	22
Agglomération de Paris	9	10

Champ : logements enquêtés, France métropolitaine

Source : EAR 2015

Ces différences préalables mesurées par des caractéristiques observées peuvent être éliminées par un appariement ou une approche par stratification. Cependant, les unités enquêtées appariées pourraient ne pas être comparables en présence d'hétérogénéité inobservée. En effet, une fois la première étape d'appariement ou de stratification effectuée, il est possible que certaines différences importantes n'aient pas été prises en compte, car inobservées, de sorte que des unités enquêtées qui semblent comparables ne le sont peut-être pas.

4.2 Le modèle d'analyse de sensibilité

Le modèle d'analyse de sensibilité proposé par Rosenbaum [Rosenbaum, 2002a,b] que nous implémentons permet de répondre en partie à cette éventualité. Il s'agit d'estimer de

quelle « importance » devrait être la caractéristique non mesurée pour arriver à modifier les conclusions émises lorsque l'on suppose l'absence de biais caché⁷, i.e. que les logements sont semblables une fois l'appariement ou la stratification effectuée. Par exemple, dans notre cas, le modèle d'analyse de sensibilité peut considérer qu'après un appariement exact ou une stratification, l'un des deux logements, pas nécessairement celui répondant par internet, a toujours deux fois plus de chance de recourir à internet : dans ce cas, quelles conclusions pourrions nous tirer sur l'effet du mode de collecte sur la non-réponse ? Plus précisément, l'approche implémentée teste l'hypothèse d'existence d'un effet du mode de collecte sur la non-réponse (individuelle et/ou partielle), et quantifie l'ampleur du biais de sélection inobservé qui devrait exister pour rejeter toute causalité dans la corrélation mise en évidence sous l'hypothèse d'absence d'hétérogénéité inobservée.

Schématiquement, l'approche s'appuie sur les données utilisées pour quantifier l'incertitude liée à une caractéristique inobservée sur nos estimations, comme un intervalle de confiance permet de tenir compte de l'incertitude liée à l'échantillonnage. À la critique habituelle « corrélation n'est pas causalité », soulevée à juste titre par l'existence possible d'une caractéristique inobservée, le modèle d'analyse de sensibilité propose une réponse que l'on pourrait formuler ainsi : « pour disqualifier toute causalité dans la corrélation mise en évidence sous l'hypothèse d'indépendance conditionnelle, le biais de sélection inobservé doit être de *telle* ampleur ».

Cette section est basée sur la présentation très complète de l'analyse de sensibilité de Rosenbaum [2002a] mais elle s'appuie aussi grandement sur Rosenbaum and Small [2017] et Hasegawa and Small [2017]. Cependant, nous ne présentons pas ici tous les détails nécessaires à son implémentation, mais nous fournissons seulement l'intuition nécessaire à sa compréhension.

Notation, participation potentielle et effet causal

Notre échantillon comporte S strates, chacune contenant n_s logements indexés par $i = 1, \dots, n_s$, où $Z_{si} = 1$ si le i^e logement de la s^e strate répond par internet (ci-après dénommé « traité ») et $Z_{si} = 0$ s'il répond au questionnaire papier (ci-après dénommé « contrôle »). Ainsi, il y a $m_s = \sum_{i=1}^{n_s} Z_{si}$ logements traités dans la strate s ⁸. Chaque strate est définie par un ensemble de covariables $\mathbf{x}$, de telle sorte que $\mathbf{x}_{si} = \mathbf{x}_s$ pour tout $i = 1, \dots, n_s$ (cf. section 4.1). Nous notons $\mathbf{Z} = {}^t(Z_{11}, Z_{12}, \dots, Z_{S, n_S})$, le vecteur de dimension $N = n_1 + \dots + n_S$ contenant les Z_{si} et $\mathbf{m} = {}^t(m_1, \dots, m_S)$.

Suivant les notations de Neyman [1923]-Rubin [1974], chaque logement a deux comportements de participation potentielle, celui r_{Tsi} qui serait observé en cas d'exposition au traitement (un refus de répondre à une question par internet correspond à $r_{Tsi} = 1$), et r_{Csi} en l'absence d'exposition au traitement, i.e. celui observé lorsque le logement est interrogé *via* le questionnaire papier. Nous nous intéressons à l'effet causal du mode de collecte par internet sur la non-réponse, δ_{si} qui compare donc r_{Tsi} et r_{Csi} , tel que $\delta_{si} = r_{Tsi} - r_{Csi}$. Les différentes valeurs possibles pour r_{Csi} , r_{Tsi} et δ_{si} sont détaillées dans le tableau 3, ci-après.

7. une hypothèse aussi appelée d'indépendance conditionnelle

8. Notez que S paires appariées sur des caractéristiques $\mathbf{x}$ est un cas spécial de celui décrit ici, avec $n_s = 2, m_s = 1$ pour $s = 1, \dots, S$.

TABLEAU 3 – Participation potentielle et effets causaux possibles

r_{Tsi}	r_{Csi}	δ_{si}	Interprétation dans notre cas
1	1	0	non-réponse quel que soit le mode de collecte
1	0	1	la collecte internet cause la non-réponse
0	1	-1	la collecte internet prévient la non-réponse
0	0	0	participation quel que soit le mode de collecte

Il convient de comprendre que les résultats potentiels sont fixes dans la population finie de N logements. Les (r_{Tsi}, r_{Csi}) sont juste $2N$ nombres *fixes* décrivant la population finie. Leur observation dépend du mode choisi, mais pas leur valeur. Par conséquent, chaque effet causal, δ_{si} est également fixe, mais non observé. Le résultat observé R_{si} qui dénote la présence d’une non-réponse à la question sur l’emploi est lié aux résultats potentiels comme $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$. Contrairement à l’effet causal, δ_{si} , qui est une caractéristique fixe mais inobservable, Z_{si} est une variable aléatoire observable. Le résultat observé R_{si} est également une variable aléatoire puisqu’il dépend de l’affectation aléatoire Z_{si} .

Étant donné qu’un seul des résultats potentiels est observé pour chaque logement, il est nécessaire d’inférer comment chaque logement aurait répondu selon le mode qu’il n’a pas choisi. Est-il plausible que le fait de répondre par internet n’ait aucun effet sur la participation ? Répondre à cette question revient à tester l’hypothèse nulle d’absence d’effet (“ *sharp null hypothesis* ”) de Fisher (1935) : $H_0 : r_{Tsi} = r_{Csi}$ pour tous (s, i) , ou $\mathbf{r}_T = \mathbf{r}_C$ avec $\mathbf{r}_T = {}^t(r_{T11}, \dots, r_{TS, n_S})$ et $\mathbf{r}_C = {}^t(r_{C11}, \dots, r_{CS, n_S})$.

Inférence aléatoire en l’absence de biais de composition caché

Comme indiqué précédemment, chaque logement a des covariables observées $\mathbf{x}_{si}$, mais il peut également présenter une covariable non observée u_{si} susceptible de provoquer une non-réponse et d’influer sur le mode choisi pour répondre à l’enquête, ce qui est généralement appelé « biais caché » ou « hétérogénéité non observée ». Les strates contrôlent les covariables observées dans le sens où $\mathbf{x}_{si} = \mathbf{x}_s$ pour tous les logements $i = 1, \dots, n_s$ dans la même strate s . Mais la strate ne contrôle généralement pas u_{si} , de sorte que deux logements, i, j dans la même strate s peuvent différer comme $u_{si} \neq u_{sj}$.

Nous notons $\mathcal{F} = \{(r_{Tsi}, r_{Csi}, \mathbf{x}_{si}, u_{si}), i = 1, \dots, n_s, s = 1, \dots, S\}$.

Soit $R_{s+} = \sum_{i=1}^{n_s} R_{si}$ le nombre total de logements non-répondants dans la strate s , et $T_s = \sum_{i=1}^{n_s} Z_{si}R_{si}$, le nombre de logements non-répondants parmi ceux traités. La notation (pour une strate) est résumée dans le tableau 4.

TABLEAU 4 – Notation pour une strate s

	Non-réponse $R_{si} = 1$	Pas de non-réponse $R_{si} = 0$	Total
Internet $Z_{si} = 1$	$T_s = \sum_{i=1}^{n_s} Z_{si}R_{si}$	—	m_s
Papier $Z_{si} = 0$	—	—	$n_s - m_s$
Total	R_{s+}	$n_s - R_{s+}$	n_s

Considérons, tout d’abord que dans chaque strate s , m_s logements parmi n_s ont été sélectionnés aléatoirement pour répondre par internet, comme cela aurait été le cas dans une expérience aléatoire stratifiée. On note $\mathcal{Z}$ l’ensemble des valeurs possible $\mathbf{z} = (z_{11}, z_{12}, \dots, z_{S, n_S})$

de $\mathbf{Z}$ dans cette expérience, de telle sorte $z_{si} = 0$ ou $z_{si} = 1$ avec $m_s = \sum_{i=1}^{n_s} z_{si}$ et $|\mathcal{Z}| = \prod_{s=1}^S \binom{n_s}{m_s}$. Ainsi, comme nous considérons une expérience aléatoire stratifiée, chaque assignation possible au traitement $\mathbf{z}$ de $\mathbf{Z}$ a la même probabilité, à savoir $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{|\mathcal{Z}|}$ pour tout $\mathbf{z}$ ⁹.

Nous avons précédemment expliqué que les seules grandeurs aléatoires sont $\mathbf{Z}$ (i.e. le mode choisi) et celles dépendantes de $\mathbf{Z}$, comme la participation observée $\mathbf{R}$. Cependant, sous l'hypothèse nulle H_0 d'absence d'effet du mode de collecte sur la participation, la participation d'un logement à l'enquête $\mathbf{R}$ est aussi fixée au sens où elle est inchangée quelque soit le mode de collecte auquel le logement est assigné : sous H_0 la participation observée d'un logement $R_{si} = r_{Csi} = r_{Tsi} = r_{si}$ pour tout $s = 1, \dots, S$ et $i = 1, \dots, n_s$, soit $\mathbf{R} = \mathbf{r} = \mathbf{r}_C$.

Afin de tester l'hypothèse nulle H_0 , on utilise une statistique de test. Une statistique de test est une fonction $t(\cdot)$ de l'assignation au traitement $\mathbf{Z}$ et des participations observées $\mathbf{R}$. L'objectif, quelle que soit la statistique de test $t(\mathbf{Z}, \mathbf{R})$, est de déterminer un niveau de significativité qui rejette l'hypothèse nulle H_0 quand $t(\mathbf{Z}, \mathbf{R})$ est élevée. Plus précisément, on recherche donc la probabilité d'obtenir une valeur de la statistique de test $t(\mathbf{Z}, \mathbf{r})$ supérieure ou égale à celle observée, k , sous H_0 , c'est-à-dire la p -value associée à la statistique de test. Comme la participation d'un logement $\mathbf{R}$ n'est pas aléatoire sous H_0 , la distribution de la statistique de test $t(\mathbf{Z}, \mathbf{R})$ sous H_0 ne dépend que de la distribution de $\mathbf{Z}$ qui est par ailleurs connue dans le cadre d'un échantillonnage aléatoire (stratifié).

En effet, cette p -value est simplement la somme des probabilités des assignations au mode de collecte $\mathbf{z} \in \mathcal{Z}$ qui conduisent à obtenir des valeurs de $t(\mathbf{z}, \mathbf{r})$ supérieures ou égales à la valeur observée k ¹⁰ :

$$\mathbb{P}(t(\mathbf{Z}, \mathbf{r}) \geq k \mid \mathcal{F}, \mathcal{Z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{1}(t(\mathbf{z}, \mathbf{r}) \geq k) \mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) \quad (1)$$

Comme $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z})$ est connu, la distribution de la statistique de test sous H_0 est aussi connue.

Dans notre expérience aléatoire stratifiée - et dans tout le reste de l'étude - nous considérons comme statistique de test le nombre de logements non-répondants par internet, $t(\mathbf{Z}, \mathbf{r}) = T = \sum_{s=1}^S T_s$. Dans chaque strate, du fait du processus d'assignation aléatoire, sous H_0 , le nombre T_s de logements non-répondants par internet suit une distribution hypergéométrique dont les paramètres sont connus et fixes (R_{s+}, m_s, n_s) (cf. tableau 4). Ainsi, (1) correspond à la distribution de la somme de S variables indépendantes hypergéométriques. Ce test est le test de [Mantel and Haenszel \(1959\)](#) pour $2 \times 2 \times S$ tableaux de contingences sur S strates. Il est analogue au test exact de Fisher mais dans le cas où il y a au moins deux strates ($S \geq 2$). Il est aussi analogue au test de [McNemar \(1947\)](#) en considérant S paires, avec $n_s = 2$ et $m_s = 1$.

L'analyse précédente était justifiée par l'assignation aléatoire issue de l'expérience aléatoire stratifiée considérée. Cependant, l'enquête annuelle de recensement n'assigne pas aléatoirement un mode de collecte aux logements d'une même strate. Chaque logement a

9. Spécifiquement, dans une expérience aléatoire, l'information sur $(r_{Tsi}, r_{Csi}, \mathbf{x}_{si}, u_{si})$ (ou $\mathcal{F}$) n'est d'aucune utilité dans la prédiction de Z_{si} . Mais le *design* d'une expérience aléatoire stratifiée impose à $\mathbf{z}$ d'appartenir à $\mathcal{Z}$, ce que nous notons de manière concise en disant que l'évènement $\mathcal{Z}$ survient.

10. Comme $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{|\mathcal{Z}|}$, (1) a une expression plus simple :

$$\mathbb{P}(t(\mathbf{Z}, \mathbf{r}) \geq k \mid \mathcal{F}, \mathcal{Z}) = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}) \geq k\}|}{|\mathcal{Z}|}$$

une probabilité inconnue, $\pi_{si} = \mathbb{P}(Z_{si} = 1 \mid \mathcal{F})$, de répondre par internet, ce qui implique différentes probabilités pour différents logements.

Supposons que les probabilités de répondre par internet pour des logements distincts soient indépendantes et que pour chaque logement i , dans chaque strate s , $0 < \pi_{si} < 1$. Nous discuterons dans la partie 4.2 ces deux assertions.

Que se passe-t-il si, de plus, la probabilité de répondre par internet d'un logement π_{si} ne dépende pas de i , mais qu'elle est identique pour tous les logements i dans une strate s ? Par exemple parce que cette probabilité ne dépend que des caractéristiques observées $\mathbf{x}_{si} = \mathbf{x}_s$ mais pas de $(r_{Tsi}, r_{Csi}, u_{si})$, une hypothèse communément appelée d'« indépendance conditionnelle ».

Dans ce cas de figure, les probabilités $\pi_{si} = \pi_s$ des logements au sein d'une même strate sont toujours inconnues, mais on peut montrer que la distribution conditionnelle de $\mathbf{Z}$ sachant $\mathbf{Z} \in \mathcal{Z}$ ¹¹ est la même pour tout $\mathbf{z}$ (une démonstration et des détails sont donnés dans l'annexe A). Plus précisément, $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{|\mathcal{Z}|}$ pour tout $\mathbf{z} \in \mathcal{Z}$. L'échantillon peut alors être analysé en utilisant les mêmes méthodes et procédures que celles associées à une expérience aléatoire stratifiée, alors même que π_{si} est inconnu pour tout couple (s, i) . En résumé, même si l'assignation à un mode de collecte n'a pas été effectuée de manière aléatoire, la distribution de Z est connue *si l'analyse est sans biais de composition inobservé et que l'on stratifie exactement sur $\mathbf{x}$* .

Le modèle d'analyse de sensibilité

Dans notre étude, cependant, des logements peuvent être similaires du point de vue des caractéristiques observées, mais différents sur des caractéristiques inobservées. Formellement, rien ne permet de croire que $\pi_{si} = \mathbb{P}(Z_{si} = 1 \mid \mathcal{F}, \mathcal{Z})$ soit la même pour tous les logements i d'une même strate s , notamment car l'attribution du mode de collecte peut dépendre d'aspects non observés de $\mathcal{F}$ tels que u_{si} .

Une analyse de sensibilité vise alors à évaluer la robustesse des conclusions obtenues sous l'hypothèse d'indépendance conditionnelle, à des violations de cette hypothèse. Plus précisément, elle cherche à déterminer l'ampleur de l'écart, par rapport à une hypothèse d'affectation aléatoire au traitement, qui devrait être présent pour modifier les conclusions faites sous cette hypothèse. Si les conclusions ne changent pratiquement pas en cas d'écart important par rapport à l'hypothèse d'une affectation aléatoire, ces conclusions seront considérées comme insensibles à l'existence d'un biais caché d'une *telle* ampleur.

Un modèle d'analyse de sensibilité a été proposé par Rosenbaum [2002a]¹². Ce modèle pose que les affectations d'un mode de collecte à chaque unité enquêtée sont initialement indépendantes les unes des autres¹³ et que les unités d'une même strate s peuvent différer dans leurs rapports de chances de répondre par internet par un facteur d'au plus $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{si}/(1 - \pi_{si})}{\pi_{sj}/(1 - \pi_{sj})} \leq \Gamma \quad (2)$$

pour $1 \leq i, j \leq n_s, s = 1, \dots, S$.

Notons que si $\Gamma = 1$ dans (2), alors $\pi_{si} = \pi_{sj}$; c'est-à-dire que $\Gamma = 1$ correspond à l'expérience aléatoire (stratifiée) mentionnée précédemment. Par conséquent, Γ est une

11. (ou de manière équivalente sachant $\mathbf{m}$)

12. Voir chapitre 4.

13. De nouveau, nous précisons les conséquences d'une telle hypothèse dans la section 4.2 sur le choix de l'unité enquêtée retenue dans cette étude.

mesure du degré d'écart par rapport à une étude sans présence de biais caché. L'analyse de sensibilité proprement dite consistera donc à considérer plusieurs valeurs de Γ et à étudier comment les inférences changent. Ainsi les conclusions seront considérées comme sensibles à l'existence d'un biais caché si les valeurs de Γ proches de 1 conduisent à des inférences très différentes de celles obtenues en supposant l'étude exempte de biais caché. Dans le cas contraire, elles seront considérées comme d'autant plus insensibles que des valeurs élevées de Γ sont nécessaires pour modifier ces conclusions.

Lorsque les strates sont des paires (i.e. $n_s = 2$ et $m_s = 1$), le modèle d'analyse de sensibilité peut s'écrire différemment. Il permet, en effet, d'encadrer la probabilité θ_s d'observer, au sein d'une paire, le logement 1 (par exemple) répondre par internet ¹⁴ :

$$\frac{1}{1+\Gamma} \leq \theta_s \leq \frac{\Gamma}{1+\Gamma} \quad (3)$$

L'analyse de sensibilité en présence d'un biais de composition caché

Pour tester l'existence d'un effet du traitement, il est nécessaire de connaître la distribution de la statistique de test T sous l'hypothèse nulle H_0 . En présence d'un biais de composition caché, cette distribution n'est pas connue puisque la distribution de $\mathbf{Z}$, qui dépend de la covariable non observée u , n'est pas connue. Cependant, comme le montre Rosenbaum [2002a], le modèle d'analyse de sensibilité permet de déduire, à Γ fixé, un encadrement de cette distribution inconnue de la statistique de test T , par les distributions de deux statistiques, notées ci-dessous T^+ et T^- , dont les paramètres sont connus et dépendent de Γ .

Proposition 1

$$\mathbb{P}(T^+ \geq k) \geq \mathbb{P}(T \geq k \mid \mathcal{F}, \mathcal{Z}) \geq \mathbb{P}(T^- \geq k), \forall k$$

Formellement, comme le montre Rosenbaum ¹⁵, la distribution de T^+ correspond, dans le cadre général d'un appariement par stratification ¹⁶, à celle de la somme de S variables aléatoires indépendantes suivant chacune une loi hypergéométrique étendue de paramètres $(R_{s+}, m_s, n_s, \Gamma)$ (voir tableau 4 pour un rappel de la définition de R_{s+}, m_s et n_s associés à chaque strate). Celle de T^- suit une distribution analogue mais dont les paramètres des lois hypergéométriques étendues sont $(R_{s+}, m_s, n_s, \frac{1}{\Gamma})$.

Cet encadrement de la distribution, sous l'hypothèse nulle, de la statistique de test T permet d'obtenir des bornes à la p -value associée à une valeur k de la statistique de test, à Γ donné. Celles-ci définissent donc un intervalle de valeurs possibles pour la p -value ¹⁷ qui reflète l'incertitude sur nos conclusions liée à un possible biais de composition caché, dont l'ampleur est caractérisée par Γ .

En faisant varier Γ , l'intervalle des valeurs possibles des p -values grandit. Il existe donc une valeur Γ_{max} de l'ampleur du biais de composition caché, au-delà de laquelle l'hypothèse d'absence d'effet de mode ne peut être rejetée. Discuter la possibilité, après appariement

14. En effet, θ_s correspond à $\mathbb{P}(Z_{s1} = 1, Z_{s2} = 0 \mid \mathcal{F}, \mathcal{Z}, Z_{s1} + Z_{s2} = 1)$.

Or $\mathbb{P}(Z_{s1} = 1, Z_{s2} = 0 \mid \mathcal{F}, \mathcal{Z}, Z_{s1} + Z_{s2} = 1) = \pi_{s1}(1 - \pi_{s2})$, et $\mathbb{P}(Z_{s1} + Z_{s2} = 1 \mid \mathcal{F}, \mathcal{Z}) = \pi_{s1}(1 - \pi_{s2}) + \pi_{s2}(1 - \pi_{s1})$.

D'après (2), $\pi_{s2}(1 - \pi_{s1}) \leq \Gamma \pi_{s1}(1 - \pi_{s2})$, il en découle que $\pi_{s2}(1 - \pi_{s1}) + \pi_{s1}(1 - \pi_{s2}) \leq (\Gamma + 1)\pi_{s1}(1 - \pi_{s2})$, soit $\frac{1}{1+\Gamma} \leq \theta_s$. Une démarche analogue pour la majoration nous permet au final d'obtenir l'encadrement (3).

15. cf. chapitre 4

16. dans la partie consacrée aux résultats, nous expliciterons plus précisément ces distributions dans le cas d'un appariement par paires

17. Lorsque $\Gamma = 1$ (qui caractérise l'absence de biais de composition caché), les deux statistiques ont une distribution identique et l'intervalle se réduit à une valeur.

ou stratification, d’une telle ampleur Γ_{max} du biais permet d’argumenter sur la pertinence ou non de l’existence d’un effet du traitement.

Retour sur deux hypothèses importantes

Dans les parties précédentes, nous avons considéré deux résultats potentiels concernant la non-réponse, r_{Tsi} qui serait observé si l’unité enquêtée répond par internet, et r_{Csi} dans le cas contraire. Cette modélisation est cependant inadaptée lorsque, par exemple, la réponse de l’unité i au traitement dépend du mode de collecte de l’unité j , comme cela peut précisément se produire pour les habitants d’un même logement. En effet, il est possible pour un habitant d’envoyer l’ensemble du questionnaire sans avoir rempli les questionnaires individuels associés (voir la section 3.3). Ainsi, l’hypothèse faite ici qu’il existe une valeur unique, disons r_{Tsi} , correspondant à l’unité i répondant par internet, hypothèse dite de SUTVA¹⁸ [Rubin, 1980], n’est valable que si nous considérons le logement comme l’unité enquêtée¹⁹.

Comme le montre clairement la définition de Γ donnée en (2), une deuxième hypothèse est faite dans le cadre de l’analyse causale utilisé dans notre étude. Nous supposons, en effet, que tous les logements ont une chance de répondre par internet, c’est-à-dire $0 < \pi_{si} = P(Z_{si=1} | \mathcal{F}, \mathcal{Z})$ pour tous les $1 \leq i \leq n_s, s = 1, \dots, S$. Dans notre cas, cependant, cette hypothèse peut ne pas être appropriée, car les habitants de certains logements peuvent ne pas avoir accès à internet²⁰. En France, en 2019, 83 % des ménages possèdent un ordinateur et 96 % un téléphone portable [Gleizes et al., 2021]. Même si les personnes âgées adoptent pour partie aujourd’hui ces nouveaux outils et services, 18 % des 60-74 ans n’ont toujours pas d’ordinateur. Afin de tenir compte de cette possibilité, nous avons vérifié la robustesse de nos conclusions en nous concentrant sur les logements comptant au moins un habitant âgé de 14 à 59 ans (soit 69 % des logements répondants), une tranche d’âge que nous supposons apte à répondre via internet ou au moins apte à savoir comment y accéder, ce qui rend l’hypothèse $0 < \pi_{si} = P(Z_{si} = 1 | \mathcal{F}, \mathcal{Z})$ plus plausible dans ce cas. Les conclusions obtenues sur ce sous-échantillon sont identiques à celles présentées dans les parties suivantes.

4.3 Résultats

Dans cette partie, nous présentons les résultats de l’analyse de sensibilité sur l’effet de la collecte internet en 2015²¹ sur plusieurs mesures de la non-réponse au questionnaire individuel ou à la question sur la situation principale de l’enquêté vis-à-vis de l’emploi. L’analyse est effectuée au niveau logement qui est la seule unité enquêtée au niveau de laquelle la non-réponse est indépendante de celle des autres unités enquêtées, comme nous l’avons expliqué précédemment²².

Plus précisément, nous nous intéressons aux différents types de non-réponse suivants :

- non-réponse individuelle totale : aucun habitant du logement n’a rempli son questionnaire individuel ;
- non-réponse individuelle : au moins un habitant du logement n’a pas rempli son questionnaire individuel ;
- non-réponse partielle totale : tous les habitants du logement n’ont pas répondu à

18. pour *Stable Unit Treatment Value*

19. De plus, deux habitants du même logement ne peuvent pas répondre avec deux modes différents, de sorte que leur probabilité de répondre par internet ne sont pas indépendantes.

20. Dans ce cas, le seul Γ plausible que nous pouvons considérer est $+\infty$.

21. Des résultats similaires sur 2018 sont disponibles en Annexe.

22. Pour autant, notre analyse ne questionne que la participation des habitants de plus de 14 ans, c’est-à-dire la participation éventuelle de ceux tenus de répondre à un questionnaire individuel.

- la question sur l'emploi (mais ont tous répondu à une partie de leur questionnaire individuel) ;
- non-réponse partielle : au moins un habitant du logement n'a pas répondu à la question sur l'emploi (mais a répondu à une partie de son questionnaire individuel).

Préalablement à l'analyse, un appariement exact par paires de logements répondant par internet et par papier est effectué à partir des (seules) caractéristiques disponibles à l'issue du questionnaire logement. Plus précisément, les logements de chaque paire ont le même nombre de personnes selon trois tranches d'âge (15-24 ans, 25-59 ans, 60 ans et plus). De même, deux logements ayant des enfants sont regroupés dans une même paire, tout comme ceux déclarés en HLM ou dont le statut d'occupation (propriétaire ou locataire) est identique.

Le faible nombre de caractéristiques observables retenu pour effectuer l'appariement ne permet pas d'exclure la possibilité d'un biais de composition caché, ce qui rend nécessaire l'analyse de sensibilité. Dans ce qui suit, celle-ci est menée pour chaque type de non-réponse à partir des résultats de l'EAR 2015. Des résultats aux conclusions similaires, obtenus pour cette même enquête, mais sans appariement ou après stratification sont disponibles en Annexe B, tout comme ceux obtenus en utilisant l'EAR 2018. Contrairement à un appariement exact par paires, l'approche par stratification permet d'avoir dans chaque strate au moins un répondant par internet et un répondant par papier, mais sans se limiter dans leur nombre maximal respectif.

Lorsque l'on considère des paires de logements, la statistique de test utilisée est la statistique associée au test de McNemar. Nous en explicitons tout d'abord le calcul avec un exemple, avant de commenter les résultats de l'analyse de sensibilité correspondante. L'appariement exact sur l'ensemble des caractéristiques évoquées nous permet de constituer 592 797 paires de logements. Parmi elles, il existe 26 952 paires où *exactement* un seul des deux logements n'a complété aucun questionnaire individuel (non-réponse individuelle totale). Parmi ces paires discordantes, il y a 25 538 paires où le logement sans questionnaire individuel est celui ayant répondu par internet.

En l'absence de biais de composition caché, le test de McNemar compare les $T = 25\,538$ paires à une distribution binomiale²³ avec 26 952 essais et une probabilité de succès de $\frac{1}{2}$, puisqu'au sein de chaque paire, après appariement, les logements sont (supposés) identiques et qu'ainsi la probabilité θ_s d'observer tel logement répondre par internet est de $\frac{1}{2}$. La valeur critique associée au test unilatéral d'absence d'effet du mode du collecte $H_0 : \delta = 0$ v.s. $H_1 : \delta > 0$ au seuil de significativité de 0.1 % de cette distribution binomiale est 13 285, bien en-deçà de la valeur de la statistique 25 538. La p -value associée à ce même test est inférieure à 0.001. En l'absence de biais de composition caché, il y a donc de fortes raisons de penser que le mode de collecte par internet engendre de la non-réponse individuelle totale. Avec un raisonnement analogue pour les autres types de non-réponse il est possible de conclure qu'en l'absence de biais de composition caché, au total, le mode de collecte par internet engendrerait d'une part de la non-réponse individuelle et d'autre part réduirait la non-réponse partielle.

En présence de biais caché, comme nous l'avons expliqué, il est possible d'encadrer la p -value associée à une valeur k pour la statistique de test T par celles associées à la même valeur k pour deux statistiques de test T^+ et T^- dont les distributions sont connues (cf.

23. Dans la partie méthodologique, il est fait mention de la loi hypergéométrique et non de la loi binomiale. La loi hypergéométrique permet de traiter dans un cadre théorique unifié l'analyse après un appariement par stratification ou dans le cas particulier de paires appariées. Si l'on considère un appariement par paires, il y a une correspondance entre la loi hypergéométrique et la loi binomiale : cette dernière distribution étant plus familière, il nous a semblé plus opportun de la mentionner explicitement.

proposition 1). Plus précisément, dans le cas de paires appariées, T^+ et T^- suivent des lois binomiales, dont les paramètres sont dans notre exemple, 26 952 essais avec des probabilités de succès $p^+ = \frac{\Gamma}{1+\Gamma}$ et $p^- = \frac{1}{1+\Gamma}$, respectivement. Formellement la proposition 1 s'écrit alors :

Proposition 2

$$\sum_{a=k}^S \binom{S}{a} (p^+)^a (1-p^+)^{S-a} \geq \mathbb{P}(T \geq k) \geq \sum_{a=k}^S \binom{S}{a} (p^-)^a (1-p^-)^{S-a}$$

avec $k = 25\,538$, la valeur de la statistique de test et $S = 26\,952$, le nombre de paires discordantes.

Afin de mener une analyse de sensibilité, on examine donc au tableau 5 l'ampleur du biais caché, représenté ici par Γ et croissant avec ce paramètre, nécessaire pour modifier ces conclusions. Dans chaque colonne, nous reportons la borne supérieure de la p -value de la statistique de test associée à différentes valeurs de Γ .

TABLEAU 5 – p -value maximale associée au test unilatéral $H_0 : \delta = 0$ v.s. $H_1 : \delta > 0$ en fonction de Γ

Γ	Intervalle des valeurs possibles de θ_s		Non-réponse			
			(i)		(ii)	
			Individuelle totale	Individuelle	Partielle totale	Partielle
1	0.50	0.50	< 0.001	< 0.001	< 0.001	< 0.001
3	0.25	0.75	< 0.001	< 0.001	< 0.001	< 0.001
5	0.17	0.83	< 0.001	< 0.001	< 0.001	< 0.001
7	0.13	0.88	< 0.001	< 0.001	< 0.001	< 0.001
9	0.10	0.90	< 0.001	< 0.001	0.981	< 0.001
11	0.08	0.92	< 0.001	0.301		< 0.001
13	0.07	0.93	< 0.001			< 0.001
15	0.06	0.94	< 0.001			
17	0.06	0.94	0.013			
19	0.05	0.95	0.968			
21	0.05	0.95				
23	0.04	0.96				

Notes : p -value associée à la statistique de test de McNemar. Une valeur manquante indique une P -value > 0.999 . La non-réponse individuelle dite totale correspond à l'absence de questionnaires individuels pour tous les habitants du logement. La non-réponse individuelle caractérise l'absence d'au moins un questionnaire individuel. De même la non-réponse partielle dite totale correspond à l'absence de réponse à la question sur l'emploi pour tous les habitants dans les questionnaires retournés. La non-réponse partielle caractérise l'absence d'au moins une réponse à la question sur l'emploi dans les questionnaires retournés.

Pour $\Gamma = 3$, on suppose qu'après appariement, au sein de chaque paire, la probabilité de répondre par internet de deux logements appariés peut ne pas être égale : leurs rapports de chance de répondre par internet peuvent différer d'un facteur au plus égal à 3 en raison d'une covariable non observée. Ainsi, malgré l'appariement, la probabilité au sein d'une paire d'observer que le logement 1 réponde par internet n'est pas nécessairement égale à 50 %, et peut n'être que de 25 % (mais pas en dessous). De même celle d'observer plutôt le logement 2 répondre par internet peut être de 75 % (mais pas plus). Dans ce cas, le niveau de significativité maximum est toujours inférieur à 0.001. Un biais de composition caché correspondant à $\Gamma = 3$ est donc insuffisant pour expliquer la différence de participation observée entre les logements répondant par internet et les logements répondant par papier : l'hypothèse nulle H_0 d'une absence d'effet de la collecte internet sur la non-réponse individuelle n'est donc pas plausible.

Plus Γ augmente, plus l'intervalle des valeurs possibles des p -values grandit, de sorte qu'il existe une valeur Γ_{max} au-delà de laquelle il n'est plus possible de rejeter l'hypothèse H_0 . L'analyse de sensibilité montre que cette hypothèse n'est plus rejetée (au seuil de 0,1 %) dès lors que $\Gamma \geq 11$. En d'autres termes, pour attribuer la part plus élevée de non-réponse individuelle sur internet à une covariable non observée plutôt qu'à un effet du mode, il faudrait que cette covariable non observée multiplie par onze les « chances » de répondre par internet de l'un des logements par rapport à celui auquel il est apparié. Pour la non-réponse individuelle totale, il faudrait supposer un Γ égal à 17 pour que l'hypothèse d'absence d'effet ne soit pas rejetée. De même, concernant l'effet du mode internet sur la réduction de la non-réponse partielle, il faudrait qu'une covariable inobservée multiplie par 13 les « chances » de répondre sur papier pour expliquer la part plus élevée de ce type de non-réponse sur ce mode.

Est-il possible d'envisager qu'une caractéristique inobservée puisse impliquer de considérer une valeur aussi élevée de Γ ? Parce qu'elle est inobservée, l'impact d'une caractéristique non retenue dans l'appariement sur la probabilité de choisir la collecte par internet ne peut être déterminé, mais il peut néanmoins être utile d'offrir au lecteur des éléments éclairants sur ce sujet à partir des quelques caractéristiques observées. Pour cela, nous avons estimé l'ampleur des rapports de chance des différentes caractéristiques observables retenues pour l'appariement. En effet, on rappelle que Γ se définit précisément en considérant le rapport de chance :

$$\frac{1}{\Gamma} \leq \frac{\pi_{si}/(1 - \pi_{si})}{\pi_{sj}/(1 - \pi_{sj})} \leq \Gamma$$

où π_{si} et π_{sj} désignent respectivement les probabilités de choisir la collecte par internet des logements i et j appariés au sein de la paire s . Plus précisément, nous proposons au tableau 6 des estimations de rapports de chance fondées sur une régression logistique du choix d'une collecte par internet sur des caractéristiques du logement. Nos résultats soulignent qu'aucun rapport de chance associé à l'une des caractéristiques utilisées pour notre appariement n'excède une valeur de 10. Cette valeur extrême est plus faible que toutes les valeurs de Γ qui devraient être atteintes pour rejeter l'existence d'un effet de mode sur la non-réponse (et l'associer à une variable cachée).

TABLEAU 6 – Rapports de chance de la probabilité de choisir la collecte par internet

	HLM		non-HLM	
	Rural	Urbain	Rural	Urbain
Logement				
Pas que des retraités (60 ans et plus)	4.4	4.6	3.6	3.2
avec enfants (< 14 ans)	1.4	1.8	1.5	1.9
avec étudiants (14-24 ans)	1.1	1.2	1.2	1.2
Statut d'occupation				
Locataire	3.4	9.8		
Propriétaire			1.4	1.3

Champ : logements enquêtés, France métropolitaine.
Source : EAR 2015

Au final, ces résultats suggèrent que, malgré l'existence d'un biais de composition caché, il est plausible de conclure que (1) les incitations à répondre, dans le formulaire internet, réduisent la non-réponse partielle, mais que (2) la possibilité de ne pas remplir tous les questionnaires individuels une fois que le questionnaire sur le logement a été renseigné peut également entraîner de la non-réponse totale. La robustesse de ces conclusions ne doit pas surprendre : elle reflète simplement comme le suggérait déjà le graphique 2 - mais sans contrôler de différences éventuelles entre enquêtés sous l'un ou l'autre mode - le caractère quasi-déterministe du type de non-réponse en fonction du mode de collecte.

5 Un effet de sélection endogène lié à la situation vis-à-vis de l'emploi ?

Répondre par internet engendre donc de la non-réponse individuelle pour certains ménages et réduit la non-réponse partielle pour d'autres. Le mode de collecte impacte ainsi le type de non-réponse. Ceci étant, comme pour toutes les enquêtes, une correction de la non-réponse est réalisée, de telle sorte que l'existence d'un écart croissant entre les estimations de l'emploi issues des EAE et celles issues des EAR depuis l'introduction de la collecte multimode questionne le mécanisme actuel de correction de la non-réponse par imputation, et notamment sa pertinence compte tenu de l'impact du mode de collecte sur le type de non-réponse.

Dans la chaîne de production des EAR, la correction de la non-réponse partielle diffère de celle de la non-réponse individuelle. La *non-réponse partielle* à la question sur l'emploi est imputée en utilisant seulement les réponses de l'enquêté à d'autres questions. Par exemple, si un individu ne répond pas à cette question mais donne le nom de son entreprise dans la suite du questionnaire, il sera corrigé en emploi par la suite. La *non-réponse individuelle* est, elle, imputée à partir des réponses données par les autres enquêtés. En pratique il s'agit d'une imputation *hotdeck* par sexe et âge, et qui ne tient donc pas compte du mode de collecte. Une telle correction de la non-réponse suppose de manière implicite que les répondants sont représentatifs des non-répondants, à sexe et âge donné, du point de vue de leur situation vis à vis de l'emploi. Elle omet donc l'existence possible d'un effet de sélection endogène quant à la non-réponse individuelle, i.e. la possibilité que les non-répondants, principalement par internet pour ce type de non-réponse, soient plus souvent en emploi que les répondants et notamment que les répondants par papier.

Dans cette partie, nous questionnons donc l'hypothèse d'un effet de sélection endogène pour expliquer les écarts d'estimations d'emploi constatés entre les EAE et les EAR. Pour cela, nous simulons des méthodes différentes de correction de la non-réponse par imputation, en distinguant ou non le mode de collecte.

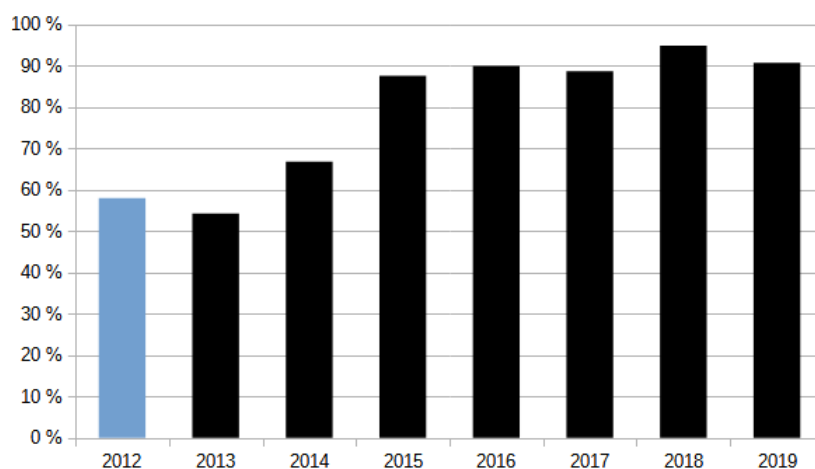
Notre première simulation s'appuie sur le constat d'un écart constant entre les estimations d'emploi EAE et celles des EAR avant l'entrée en vigueur de la collecte multimode (cf. graphique 1 en introduction). Si, comme dans notre estimation par différence de différences de la partie 2, nous supposons que cet écart serait resté constant sans l'introduction de la collecte multimode, quel devrait être le taux d'emploi des non-répondants²⁴ ? Plus précisément, nous calculons le taux d'emploi des non-répondants sur les années 2013-2019 pour que l'écart entre les EAE et les EAR se maintienne à son niveau de 2012. Les résultats sont présentés dans le graphique 3²⁵. Nos simulations suggèrent que le taux d'emploi des non-répondants devrait être, après introduction de la collecte multimode, presque deux fois plus élevé que celui de référence (environ 90 % entre 2015 et 2019, contre 58 % en 2012) pour observer un écart identique à celui de 2012 et constant sur la période considérée entre les deux estimations d'emploi. Le taux de 90 % est très élevé mais inférieur à 100 % ; cette première simulation suggère donc que l'hypothèse d'un effet de sélection endogène, concomitant ou lié à l'introduction de la collecte multimode, ne peut être écarté pour expliquer une grande partie de l'évolution des écarts d'estimation d'emplois entre les EAE

24. Cette hypothèse exclut donc toute source d'écart autre que l'introduction d'une collecte multimode, notamment, par exemple, tout effet lié aux différences de champ entre ces deux estimations, comme une évolution de l'emploi transfrontalier, ou toute autre évolution ayant une signature tendancielle. On peut néanmoins rappeler que le développement de l'apprentissage et le cumul emploi retraite pourrait avoir une influence sur la nature de l'écart.

25. On trouvera en Annexe C le taux d'emploi, entre 2012 et 2019, des répondants par mode de collecte sans et avec imputation par les services responsables de l'enquête, ainsi que la proportion de répondants par internet parmi les répondants.

et celles issues des EAR.

GRAPHIQUE 3 – Taux d’emploi des non-répondants (tous modes confondus) nécessaire pour un écart constant en volume



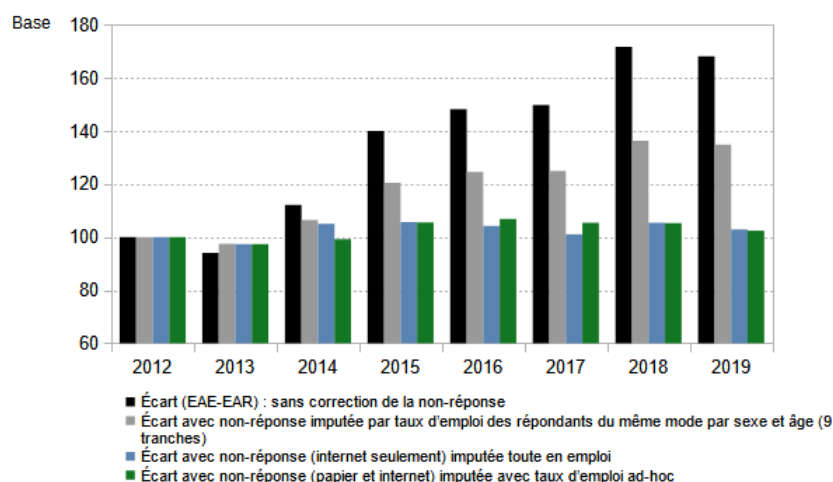
Lecture : En 2015, le taux d’emploi des non-répondants doit être de 90 % pour observer un écart en nombre d’emplois entre les EAE et les EAR identique à celui de 2012.

Notre seconde approche consiste à effectuer plusieurs exercices de simulation de façon à obtenir des ordres de grandeur. Elle vise à simuler plusieurs méthodes de correction de la non-réponse par imputation tenant compte du mode de collecte choisi. La première correction applique le taux d’emploi, par sexe et tranches d’âge, des répondants du même mode : elle diffère donc sur ce dernier point de la correction appliquée actuellement dans la chaîne de production des EAR qui ignore le mode de collecte choisi. La deuxième correction corrige la non-réponse individuelle des seuls non-répondants par internet en supposant qu’ils sont tous en emploi. Cette correction n’est pas crédible mais permet de mesurer le nombre maximal de personnes en emploi que l’on peut attendre en jouant seulement sur la correction des individus répondant par internet. Enfin, la dernière correction applique un taux d’emploi *ad-hoc* aux non-répondants par internet de 80 % à partir de 2015 et un taux d’emploi des non-répondants par papier de 30 % en 2012 qui augmente pour atteindre 60 % en 2019. Cette dernière correction fait (i) l’hypothèse que la collecte par internet entre 2015 et 2019 engendre une proportion importante d’enquêtés en emploi parmi les non-répondants par internet mais aussi (ii) l’hypothèse d’une augmentation plus générale du non renseignement par des enquêtés en emploi, du module emploi pour répondre plus rapidement au questionnaire ; un comportement qui affecterait donc aussi la collecte papier du recensement et aurait été d’autant plus fréquent que la collecte par internet se généraliserait.

Le graphique 4 représente, pour chaque simulation, l’ampleur et l’évolution de l’écart (en base 100 en 2012²⁶) ainsi que celles en l’absence de correction de la non-réponse.

26. c’est-à-dire, plus précisément, par rapport à l’écart constaté en 2012 avec la correction de la non-réponse envisagée pour chaque cas.

GRAPHIQUE 4 – Impact de différentes simulations de correction de la non-réponse sur l'écart des estimations d'emploi par EAE et EAR (en relatif, base 100 : 2012)



La correction de la non-réponse en imputant aux non-répondants le taux d'emploi observé par sexe, par âge et par mode de collecte des répondants correspondants (en gris) conduit à obtenir une estimation de l'emploi à partir des EAR plus faible que celle des EAE après l'entrée en vigueur de la collecte multimode ; l'écart, stable, avant l'entrée en vigueur de la collecte multimode, s'accroît au fur et à mesure que la collecte par internet augmente mais dans une ampleur moindre à celle observée en l'absence de correction de la non-réponse (en noir). Si cette correction de la non-réponse tient compte du mode de collecte choisi pour imputer la non-réponse, i.e. qu'elle tient compte de différences de taux d'emploi entre les répondants des deux modes de collecte, elle ne permet pas d'expliquer l'ensemble de l'écart observé.

La deuxième correction de la non-réponse impute à l'ensemble des non-répondants par internet une situation en emploi (en bleu). Avec cette correction, l'écart entre les deux estimations d'emploi (EAE et EAR) reste constant et similaire à celui observé antérieurement à l'entrée en vigueur de la collecte multimode, et ce malgré l'importance accrue de la collecte par internet. Considérer que les non-répondants par internet sont tous en emploi permettrait d'expliquer l'évolution de l'ampleur de l'écart observé²⁷. Si cette hypothèse ne peut pas être validée, elle suggère néanmoins qu'un effet de sélection inobservable existe, i.e. que les non-répondants qui choisissent de répondre par internet présentent certainement un taux d'emploi supérieur à celui des non-répondants, y compris du même mode. La troisième correction de la non-réponse avec des taux d'emplois *ad-hoc* plus cohérents (en vert) montre que l'écart reste stable dans le temps avant et après l'introduction puis l'élargissement de la collecte par internet.

6 Conclusion

Notre étude détaille l'approche par analyse de sensibilité qui permet de discuter de la pertinence des conclusions sur l'existence d'un effet du traitement sur une variable binaire, quand on suppose qu'il subsiste un biais de composition caché après appariement entre unités traitées et non traitées. Cette méthode est appliquée ici à l'étude de l'impact du mode de collecte sur le type de non-réponse. Elle pourrait aussi être implémentée pour mettre en évidence un effet de mesure - notamment en l'absence de protocole dédié - en

27. On rappelle que l'écart est en base 100 en 2012, i.e. que son évolution représentée ici dépend de la correction de la non-réponse choisie et ne reflète pas nécessairement son niveau.

comparant les réponses de répondants à des modes de collecte différents. Il faut souligner cependant que la pertinence des conclusions d'une analyse de sensibilité dépend de la qualité de l'appariement. Dans le cas du recensement, comme nous l'avons souligné, les caractéristiques observées sont peu nombreuses. De ce point de vue, il était donc important de tester pour des valeurs élevées de l'insensibilité des conclusions à la présence d'un biais conséquent de composition cachée.

Au-delà, cette étude visait à expliciter l'évolution de l'écart entre les estimations d'emploi obtenues à partir des EAR et celles issues des EAE suite à l'introduction de la collecte par internet. Nos résultats ont mis en évidence un impact du mode de collecte sur le type de non-réponse : la collecte par internet accroîtrait la non-réponse totale aux questionnaires individuels des logements enquêtés alors qu'elle réduirait la non-réponse partielle à la question sur la situation principale vis-à-vis de l'emploi. Cette conclusion est renforcée par l'analyse de la mise en œuvre du questionnaire internet qui « permet » à la fin du questionnaire logement d'envoyer sa réponse à l'enquête sans remplir (tous) les formulaires individuels associés aux habitants. Cette facilité offerte sur internet est impossible avec les formulaires papiers qui sont (au moins partiellement) contrôlés par l'agent recenseur. À l'inverse, la présence d'un rappel en cas de non-réponse à une question posée lors du remplissage du questionnaire individuel par internet pourrait expliquer la plus faible non-réponse partielle associée à la collecte internet.

Cette mécanique rend fragile les pratiques de correction de la non-réponse se basant sur des imputations et telles qu'implémentées pour les EAR. En effet, la non-réponse totale aux questionnaires individuels est corrigée à partir des taux d'emploi des répondants de même sexe et de même tranche d'âge, mais sans tenir compte du mode de collecte. Ainsi, une telle méthode suppose implicitement que les non-répondants par internet auraient des taux d'emploi similaires aux répondants, y compris à ceux ayant répondu aux formulaires papier. Pourtant tenir compte du mode de collecte dans la procédure de redressement permettrait de réduire de moitié l'écart sur l'estimation du volume d'emploi. De plus, une telle correction de la non-réponse ignore aussi la possibilité que les non-répondants à un mode de collecte soient plus fréquemment en emploi que les répondants au même mode. Nos simulations suggèrent à l'inverse qu'une telle hypothèse serait à même d'expliquer l'accroissement de l'écart entre les estimations d'emploi à partir des EAR et celles issues des EAE, compte tenu notamment de l'accroissement progressif de la part des questionnaires du recensement collectés par internet.

Cette étude a permis de nourrir la réflexion du projet *Recens'smart*. Ce projet visait à développer un questionnaire internet du recensement de la population (RP) accessible tous publics, compatible tous supports, dans la filière d'enquête de l'Insee. Cela devait permettre, entre autre, de résoudre le problème lié à la non-réponse totale individuelle observée sur internet. De fait, avec le nouveau site, et la première collecte (2024), ce type de non-réponse a fortement baissé : pour les répondants internet, elle passe de 5,6% en 2023 à 1,4% en 2024. Cette évolution laisse penser que la différence de comportement de réponse entre les deux modes de collecte était en partie lié à un problème d'ergonomie du questionnaire. Le projet *Recens'Smart* a ainsi permis de résoudre de la manière la plus satisfaisante la non réponse. En jouant sur l'amélioration du protocole de collecte, la baisse de la non réponse individuelle permet en effet de la corriger en amont avec une données et non pas en aval avec une données reposant sur l'hypothèse du modèle de correction. Concernant la correction de la non-réponse, il est envisagé de modifier la chaîne de redressement à l'aval de l'enquête afin de répondre à plusieurs besoins. Le processus de redressement doit en effet être mis à jour pour l'adapter à la rénovation du questionnaire individuel prévue pour l'EAR 2025. À cette occasion, il serait possible d'intégrer le mode de collecte ou, plus généralement, de nouvelles variables provenant du nouveau questionnaire du recensement.

Ces changements sont encore à l'étude et ne pourront être testés avant l'enquête annuelle du recensement de 2026.

Références

- J. Daikeler, M. Bošnjak, and K. L. Manfredaand. Web Versus Other Survey Modes : An Updated and Extended Meta-Analysis Comparing Response Rates. *Journal of Survey Statistics and Methodology*, 08 :513–539, 2020. URL <https://doi.org/10.1093/jssam/smz008>.
- R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.
- F. Gleizes, S. Legleye, and A. Pla. Ordinateur et accès à internet : les inégalités d'équipement persistent selon le niveau de vie. *Insee Focus*, 226, 2021. URL <https://www.insee.fr/fr/statistiques/5057474>.
- Raiden Hasegawa and Dylan Small. Sensitivity analysis for matched pair analysis of binary data : From worst case to average case analysis. *Biometrics*, 73(4) :1424–1432, 2017. doi : <https://doi.org/10.1111/biom.12688>.
- Dirk Heerwegh and Geert Loosveldt. Face-to-Face versus Web Surveying in a High-Internet-Coverage population : Differences in Response Quality. *Public Opinion Quarterly*, 72 : 836–846, 2008. URL <https://doi.org/10.1093/poq/nfn045>.
- H. Koumarianos and A. Schreiber. Conception de questionnaires auto-administrés. *Document de travail - Insee*, 02, 2021.
- N. Mantel and W. Haenszel. Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, 22 :719–748, 1959.
- Q. McNemar. Note on the sampling error of the differences between correlated proportions or percentage. *Psychometrika*, 12 :153–157, 1947.
- Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczych*, Tom X :1–51, 1923. Réimprimé en anglais dans *Statistical Science*, 1990, 5, 463–480.
- Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics, 2002a.
- Paul R. Rosenbaum. Attributing Effects to Treatment in Matched Observational Studies. *Journal of the American Statistical Association*, 97(457) :183–192, march 2002b. URL <https://doi.org/10.1198/016214502753479329>.
- Paul R. Rosenbaum and Dylan S. Small. An Adaptive Mantel-Haenszel Test for Sensitivity Analysis in Observational Studies. *Biometrics*, 73(2) :422–430, june 2017. URL <https://doi.org/10.1111/biom.12591>.
- D.B. Rubin. Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology*, 66(5) :688–701, 1974. URL <https://doi.org/10.1037/h0037350>.
- D.B. Rubin. Randomization Analysis of Experimental Data : The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371) :591–593, 1980. URL <https://doi.org/10.2307/2287653>.

Annexe A Assignation à un mode de collecte avec des probabilités inconnues par strates sous l'hypothèse d'absence de biais caché

On note $\pi_k = \mathbb{P}(Z_k = 1 \mid \mathcal{F})$ la probabilité inconnue (et indépendante de celle des autres unités) d'une unité enquêtée k de répondre par internet ($0 < \pi_k < 1$).

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}) = \mathbb{P}(Z_1 = z_1, \dots, Z_N = z_N) = \prod_{k=1}^N \pi_k^{z_k} (1 - \pi_k)^{1-z_k}$$

En outre, on suppose ces probabilités π , bien qu'inconnues, de ne dépendre que des caractéristiques observées $\mathbf{x}$. Il existe donc une fonction (inconnue) $\lambda(\cdot)$ telle que $\pi_k = \lambda(x_k)$ pour tous $k = 1, \dots, N$. Ainsi, dans le cas d'une stratification exacte sur $\mathbf{x}$, toutes les unités enquêtées i dans une même strate s ont la même probabilité de choisir de répondre par internet, $\lambda(\mathbf{x}_{si}) = \lambda_s$, et

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}) = \prod_{s=1}^S \prod_{i=1}^{n_s} \lambda_s^{z_{si}} (1 - \lambda_s)^{1-z_{si}} \quad (4)$$

Dans (4), la distribution de la variable d'assignation au mode de collecte, $\mathbb{P}(\mathbf{Z} = \mathbf{z})$ est toujours inconnue car les fonctions λ_s sont inconnues et $m_s = \sum Z_{si}$ sont des variables aléatoires. Cependant, chaque assignation $\mathbf{z}$ telle que $m_s = \sum_i z_{si}$ a la même probabilité :

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod_{s=1}^S \lambda_s^{m_s} (1 - \lambda_s)^{n_s - m_s}$$

Il s'ensuit que pour toute affectation $\mathbf{z}$, la probabilité conditionnelle sachant $\mathcal{Z}$ est *constante*. Ainsi $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathbf{m}) = \frac{1}{|\mathcal{Z}|}$ avec $|\mathcal{Z}| = \prod_{s=1}^S \binom{n_s}{m_s}$. C'est pourquoi, si l'analyse est *sans biais inobservé et que l'on stratifie exactement sur $\mathbf{x}$* , alors l'échantillon peut être analysée en utilisant les méthodes et les mêmes procédures que celles associées à une expérience aléatoire, alors même que π_{si} est inconnu pour tous (s, i) .

Annexe B Résultats additionnels

Le tableau B.1 présente les résultats de l'analyse de sensibilité de l'effet du mode de collecte par internet sur différents types de non-réponse en 2015 en s'appuyant sur la statistique de test de Fisher. Plus précisément, les résultats présentés sont obtenus en ne considérant aucun appariement par paires ni par stratification.

TABLEAU B.1 – p -value maximale associée au test unilatéral $H_0 : \delta = 0$ v.s. $H_1 : \delta > 0$ en fonction de Γ en l'absence d'appariement

Γ	Avec non-réponse			
	Individuelle totale	Individuelle	Partielle totale	Partielle
1	< 0.001	< 0.001	< 0.001	< 0.001
3	< 0.001	< 0.001	< 0.001	< 0.001
5	< 0.001	< 0.001	< 0.001	< 0.001
7	< 0.001	< 0.001	< 0.001	< 0.001
9	< 0.001	< 0.001	< 0.001	< 0.001
11	< 0.001		< 0.001	< 0.001
13	< 0.001		0.996	
15	0.075			
17				
19				
21				
23				

Note : Le tableau reporte la p -value associée à la statistique de test de **Fisher**. Une valeur manquante indique une p -value > 0.999. La non-réponse individuelle dite totale correspond à l'absence de questionnaires individuels pour tous les habitants du logement. La non-réponse individuelle caractérise l'absence d'au moins un questionnaire individuel. De même la non-réponse partielle dite totale correspond à l'absence de réponse à la question sur l'emploi pour tous les habitants dans les questionnaires retournés. La non-réponse partielle caractérise l'absence d'au moins une réponse à la question sur l'emploi dans les questionnaires retournés.

Le tableau B.2 présente les résultats de l'analyse de sensibilité de l'effet du mode de collecte par internet sur différents types de non-réponse en 2015 en s'appuyant sur la statistique de test de Cochran-Mantel-Haenszel-Birsch, c'est-à-dire en considérant un appariement par stratification.

TABLEAU B.2 – p -value maximale associée au test unilatéral $H_0 : \delta = 0$ v.s. $H_1 : \delta > 0$ en fonction de Γ après stratification - EAR 2015

Γ	Individuelle totale	Avec non-réponse		
		Individuelle	Partielle totale	Partielle
1	< 0.001	< 0.001	< 0.001	< 0.001
3	< 0.001	< 0.001	< 0.001	< 0.001
5	< 0.001	< 0.001	< 0.001	< 0.001
7	< 0.001	< 0.001	< 0.001	< 0.001
9	< 0.001	< 0.001	0.752	< 0.001
11	< 0.001	< 0.001		< 0.001
13	< 0.001	0.700		< 0.001
15	< 0.001			
17	< 0.001			
19	< 0.001			
21	< 0.001			
23	0.884			

Note : Le tableau reporte la p -value associée à la statistique de test de **Cochran-Mantel-Haenszel-Birsch**. Une valeur manquante indique une p -value > 0.999. La non-réponse individuelle dite totale correspond à l'absence de questionnaires individuels pour tous les habitants du logement. La non-réponse individuelle caractérise l'absence d'au moins un questionnaire individuel. De même la non-réponse partielle dite totale correspond à l'absence de réponse à la question sur l'emploi pour tous les habitants dans les questionnaires retournés. La non-réponse partielle caractérise l'absence d'au moins une réponse à la question sur l'emploi dans les questionnaires retournés.

Enfin, le tableau B.3 présente, lui-aussi, les résultats de l'analyse de sensibilité de l'effet du mode de collecte par internet sur différents types de non-réponse après appariement par stratification, mais pour l'année 2018.

TABLEAU B.3 – p -value maximale associée au test unilatéral $H_0 : \delta = 0$ v.s. $H_1 : \delta > 0$ en fonction de Γ après stratification - EAR 2018

Γ	Avec non-réponse			
	Individuelle totale	Individuelle	Partielle totale	Partielle
1	< 0.001	< 0.001	< 0.001	< 0.001
3	< 0.001	< 0.001	< 0.001	< 0.001
5	< 0.001	< 0.001	< 0.001	< 0.001
7	< 0.001	< 0.001	< 0.001	< 0.001
9	< 0.001	< 0.001		< 0.001
11	< 0.001	< 0.001		< 0.001
13	< 0.001	< 0.001		< 0.001
15	< 0.001			0.127
17	< 0.001			
19	< 0.001			
21	< 0.001			
23	< 0.001			

Note : Le tableau reporte la p -value associée à la statistique de test de **Cochran-Mantel-Haenszel-Birsch**. Une valeur manquante indique une p -value > 0.999. La non-réponse individuelle dite totale correspond à l'absence de questionnaires individuels pour tous les habitants du logement. La non-réponse individuelle caractérise l'absence d'au moins un questionnaire individuel. De même la non-réponse partielle dite totale correspond à l'absence de réponse à la question sur l'emploi pour tous les habitants dans les questionnaires retournés. La non-réponse partielle caractérise l'absence d'au moins une réponse à la question sur l'emploi dans les questionnaires retournés.

Annexe C Taux d'emploi des EAR par mode de collecte sans et avec imputation

TABLEAU C.4 – Taux d'emploi dans les EAR sans et avec imputation des répondants

	Tous modes		Papier		Internet		% répondants par internet
	sans	avec	sans	avec	sans	avec	
2012	47	52	45	50	57	64	<0.5 %
2013	47	52	45	50	54	60	<0.5 %
2014	46	52	45	50	56	62	2
2015	46	51	37	42	56	62	35
2016	45	51	35	40	55	61	43
2017	46	51	33	39	53	59	51
2018	46	52	33	38	53	59	53
2019	46	52	32	38	52	59	56

Note : De manière résiduelle, il existe des répondants par internet entre 2012 et 2014.

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 multidimensionnelle.
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'enquête Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par pondération 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'HOURL, R. LE SAOUT, B. ROUPPERT

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

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD, R. LE SAOUT

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

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01 : Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01 : Estimation avec le score de propension sous 
S. QUANTIN

M2018/02 : Modèles semi-paramétriques de survie en temps continu sous 

S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK, E. COUDIN, S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOURL

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK, A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL, P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS, A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE, E. COUDIN

M2022/02 : Le zonage en unités urbaines 2020
V. COSTEMALLE, S. OUJIA, C. GUILLO, A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la

statistique publique : méthodes et cas d'usages
D. BABET, Q. DELTOUR, T. FARIA, S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L. CASTELL, C. FAVRE-MARTINOZ, N. PALIOD, P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L. MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL, M. CLERC, D. CROZE, S. LEGLEYE, A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A. QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautille - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT, S. QUANTIN