Estimation *a priori* de l'impact du nombre de contacts sur les taux de réponse

Steve Jakoubovitch[1]

[1] *Insee,*
Département des méthodes statistiques
18 boulevard Adolphe Pinard, timbre L120,
75675 Paris cedex 14
steve.jakoubovitch@insee.fr

Résumé. La qualité des enquêtes ménages de l'Insee repose en grande partie sur de bons taux de réponse et donc sur les efforts déployés par les enquêteurs pour contacter les ménages à interroger et les convaincre de répondre à l'enquête. Jusqu'à fin 2012, il n'existait pas de consignes précises concernant le nombre de contacts à effectuer avant de renoncer à un questionnaire. De nouvelles conditions d'emploi des enquêteurs, en place depuis 2013, amènent cependant l'Insee à planifier plus en amont l'organisation des collectes. Dans ce contexte, l'ajout d'une consigne limitant le nombre de contacts à effectuer pour obtenir une réponse des ménages a été envisagé, et son impact sur les taux de réponse et sur la qualité des estimateurs a été évalué. L'absence de consignes précises sur le nombre de contacts ne permettait pas de procéder par comparaison entre deux enquêtes ménages. L'étude profite de l'enquête Difficultés de Collecte 2011 qui permet de connaître le nombre de tentatives de contacts effectuées pour chaque questionnaire et donc de simuler l'application de consignes limitant le nombre de tentatives de contacts. On détaillera la méthode employée et les résultats obtenus. L'impact du nombre de tentatives de contacts sur le taux de réponse et sur la position des ménages répondants sera analysé. Enfin nous commenterons les difficultés rencontrées pour quantifier l'impact du nombre de contacts sur la qualité des estimateurs en utilisant nos simulations.

Mots-clés. Enquête ménage, SRCV, Difficultés de Collecte, Méthodes de Collecte, Simulation, Repondération

1 Contexte de l'étude

Au 1er janvier 2013, l'Insee adoptait les Nouvelles Conditions d'Emploi des Enquêteurs (NCEE). L'institut abandonnait sa logique de payement en fonction du nombre de Fiches Adresse (FA) réalisées par les enquêteurs pour leur verser un salaire mensuel fixe. Cette évolution organisationnelle nécessite une planification des collectes plus en amont et posait à l'époque de l'étude, fin 2012, des questions sur les consignes qu'il était raisonnable de donner aux enquêteurs en ayant en tête la qualité des enquêtes.

L'étude s'arrête en particulier sur l'impact de l'effort déployé par les enquêteurs pour obtenir un entretien avec les enquêtés sur la qualité des données finales. Il s'agissait de quantifier les effets d'une consigne potentielle consistant à limiter le nombre de déplacements effectués par les enquêteurs, consigne qui n'avait jamais été donnée lors des enquêtes des années précédentes et ne pouvait donc pas être évaluée directement. L'approche choisie consiste à utiliser des données préexistantes, celles de l'enquête Statistiques sur les Ressources et Conditions de Vie (SRCV) 2011, pour simuler une diminution du nombre de tentatives de contact.

2 La phase de contact dans les enquêtes ménages

L'unité de base des enquêtes ménages Insee est la FA qui correspond à un logement à interroger. Avant la collecte, ces FA sont tirées dans l'Échantillon Maître du recensement et affectées aux enquêteurs. Pour chacune de ses FA, l'enquêteur peut potentiellement passer par 3 étapes emboitées :

-l'étape du repérage : l'enquêteur va chercher à identifier le logement et à vérifier le nom de ses habitants

-l'étape de contact (seulement si l'étape de repérage est un succès) : l'enquêteur va chercher à contacter un des membres du ménage et à lui faire accepter de répondre à l'enquête. Ces contacts peuvent être des contacts « sur le terrain », l'enquêteur se rendant au domicile des enquêtés, ou téléphoniques. Ces contacts incluent à la fois des tentatives de parler avec les enquêtés potentiels et des tentatives de les convaincre de répondre. Cette étape se termine par l'acceptation/le refus des enquêtés de répondre à l'enquête ou par le renoncement de l'enquêteur

-la passation du questionnaire (si l'étape de contact est un succès)

Dans les enquêtes ménages antérieures à 2013, il n'existait pas de consignes portant sur le nombre de contacts à effectuer avant de renoncer. C'est pourtant ce type de consigne (par exemple : « renoncer à la FA au bout de 3 tentatives de contacts infructueuses ») dont on cherche à estimer l'impact.

Dans la présentation, on se limitera à évaluer les consignes invitant à renoncer aux FA après un certain nombre de contacts en face à face (3 ou 4).

On travaillera avec l'hypothèse[1] qu'en l'absence de consigne précise, les enquêteurs ne renonçaient pas à leurs FA avant d'avoir obtenu un refus explicite de la part des habitants du logement enquêté.

3 Présentation des données

On travaille sur l'enquête SRCV. L'enquête est la partie française de l'enquête communautaire EU-SILC (European union-Statistics on income and living conditions), elle porte sur les revenus, la situation financière et les conditions de vie des ménages français.

Il s'agit d'un panel par échantillons rotatifs renouvelés chaque année d'un neuvième (l'enquête ayant débuté en 2004, les premiers entrants n'ayant pas renoncé à l'enquête étaient toujours présents en 2011).

On choisit d'utiliser l'édition 2011 de l'enquête et de se restreindre aux FA nouvellement entrantes dans l'enquête. Ce choix est justifié par la présence de métadonnées supplémentaires pour ces FA : en effet en parallèle de l'enquête SRCV 2011 était menée l'enquête Difficultés de collecte 2011 auprès des enquêteurs. Ces derniers devaient, pour chaque FA nouvellement entrante, remplir un questionnaire indiquant notamment le nombre de tentatives de contact (en séparant les contacts en face à face et téléphoniques).

Le nombre de FA répondantes concernées par notre champ est de 2 015 (taux de réponse de 71 %).

4 Simuler une limitation du nombre de contacts : le travail sur les données

Pour estimer leur impact, on cherche à exprimer les difficultés de collecte en termes de population non-répondante supplémentaire. On supprime alors cette population (on parlera de la population « censurée ») des données initiales de l'enquête, on la traite comme une population de non-répondants, avant de repondérer la population de répondants restants[2] et de calculer de nouvelles

[1] L'hypothèse étant un peu forte, on peut l'alléger en se contentant de supposer que les renoncements d'enquêteurs n'intervenaient qu'après un nombre de contacts supérieur à celui de la consigne testée.

[2] Notre repondération se fait en deux étapes :

estimations.
La baisse de qualité de l'enquête est alors évaluée en comparant les estimations obtenues après censure avec celles qui utilisent l'ensemble des données récoltées.

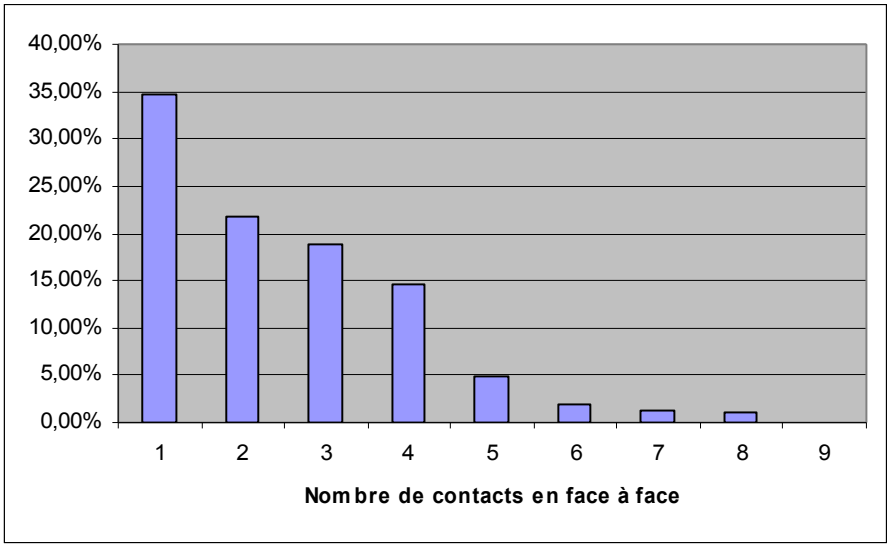
4-1 Une nécessaire imputation

Les variables concernant le nombre de contacts ne peuvent cependant pas être utilisées directement. L'enquête Difficultés de collecte 2011 contenant de la non-réponse de la part des enquêteurs (l'information est manquante pour 20% des FA où les ménages ont répondu). La non-réponse est importante dans trois régions : Île-de-France, Auvergne et Picardie (taux de réponse respectifs de 55 %, 64 % et 68 % des enquêteurs).
Nous avons donc imputé pour chaque FA non renseignée un nombre de contacts (en imputant séparément la variable sur le nombre de contacts en face à face et celle sur le nombre total de contacts, la méthode utilisée est détaillée dans l'Annexe). Les résultats présentés dans le cadre de cette étude s'appuient sur des bases contenant des données imputées.

4-2 Le nombre de contacts dans l'enquête SRCV 2011

En moyenne, sur une nouvelle FA, un enquêteur réalisera 2,6 contacts en face à face et 3,8 contacts au total. Dans le cas des FA se terminant par une réponse du ménage, la moyenne est de 2,3 contacts en face à face et 3,3 contacts totaux. Dans le cas des FA où l'entretien n'aboutit pas, la moyenne est de 3,3 contacts en face à face et 5,0 contacts au total.

Répartition des fiches adresse par nombre de contacts en face à face

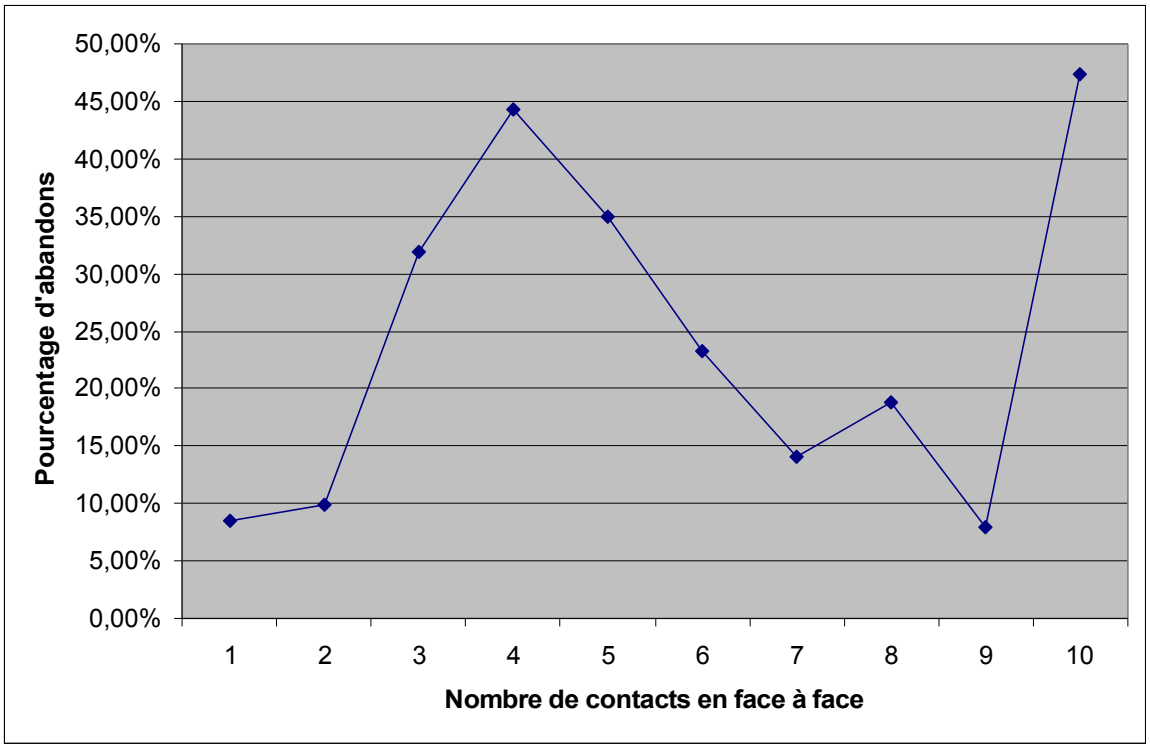


Lecture : 35 % des fiches adresse n'ont donné lieu qu'à un contact en face à face.

Pour les 3/4 des FA, l'enquêteur effectue 3 contacts en face à face ou moins avant d'obtenir les réponses des enquêtés ou d'abandonner. Seuls 10 % des fiches adresse donnent lieu à 5 contacts ou plus.

Pourcentage de renoncements en fonction du nombre de contacts en face à face

- une modélisation et une correction de la non-réponse ;
- un calage des données à partir d'informations issues de l'enquête Emploi 2011.



Lecture : En cas de quatrième contact en face à face ne débouchant pas sur la passation du questionnaire, les enquêteurs renoncent définitivement à la FA concernée dans 45 % des cas.

4-3 Les populations censurées

On construit alors deux tables :

- une table considérant que les répondants ayant nécessité 5 contacts en face à face ou plus n'auraient finalement pas répondu. Elle correspond à la consigne « Renoncer à une FA après 4 contacts en « face à face » infructueux ». Il reste 1 905 répondants en utilisant un tel critère. Le taux de réponse passe alors à 65 % ;
- une table considérant que les répondants ayant nécessité 4 contacts en face à face ou plus n'auraient finalement pas répondu. Elle correspond à la consigne « Renoncer à une FA après 3 contacts en « face à face » infructueux ». Il reste 1 249 répondants en utilisant un tel critère. Le taux de réponse passe alors à 42 %.

5- Étude des populations « censurées »

La stratégie mise en œuvre pour étudier l'impact des renoncements aux entretiens nous amène à construire des populations dont nous supposons que les individus qui les composent n'auraient pas répondu dans des conditions de collecte plus contraintes. L'objectif final de nos études est d'analyser l'impact de ces non-répondants supplémentaires sur les estimations des variables d'intérêt de l'enquête. Dans un premier temps nous pouvons cependant chercher à caractériser ces populations dans l'échantillon (on ne se préoccupe pas ici de questions de repondération).

Composition des populations censurées

	Ont répondu après 4 contacts en face à face	Ont répondu après 3 contacts en face à face	Tous les répondants
Composition du ménage			

Seul	26%	24%	17%
Famille monoparentale	22%	25%	31%
Couple sans enfant	11%	10%	8%
Couple avec enfant(s)	8%	12%	15%
Autre	30%	27%	29%
Situation professionnelle			
En emploi	55%	47%	51%
Retraité	15%	20%	28%
Étudiant	13%	9%	7%
Chômeur	9%	8%	9%
Au foyer	6%	6%	8%
Sexe			
Homme	45%	46%	48%
Femme	55%	54%	52%
Âge moyen (en années)	42	45	49
Difficultés à concilier vie familiale et vie professionnelle	40%	36%	29%
Difficultés à équilibrer son budget	27%	21%	19%
Déclare des problèmes de violences ou de vandalisme dans les environs	18%	24%	18%

Lecture : Les individus retirés de l'analyse car ils ont nécessité plus de 4 contacts en face à face pour répondre sont dans 55 % des femmes contre 52 % pour l'ensemble des répondants.

Les personnes vivant seules, les étudiants et les personnes ayant du mal à concilier vie familiale et vie professionnelle sont plus difficiles à contacter et plus nombreuses à avoir nécessité un nombre important de contacts avant d'accepter de répondre à l'enquête.

6 Stratégie de repondération

La stratégie de « censure » utilisée dans notre simulation nous amène à considérer certaines FA comme non répondantes. On cherche alors à comparer les estimations obtenues avec celles issues de la table de données sans censure. On ne peut cependant pas se contenter de comparer les données en utilisant les poids de sondage initiaux, les estimations calculées à partir des enquêtes ménages Insee cherchent à prendre en compte les processus de non-réponse. On cherchera à adopter la même méthode de repondération pour les deux jeux de données.
Si la plupart des études basées sur SRCV utilisent l'aspect longitudinal de l'enquête et doivent donc en tenir compte au moment de repondérer les données, nos études se concentrent sur une vague d'entrants et nous pouvons donc penser SRCV comme une enquête transversale. Le processus de repondération est donc simplifié.

On fera le choix d'une repondération en deux étapes :

- une modélisation de la non-réponse au niveau des fiches adresse permettant de

constituer des Groupes de Réponse Homogène (GRH) et une correction au niveau de chacun de ces GRH ;
- un calage au niveau ménage ou individu.

6.1-Modélisation de la non-réponse

La méthode retenue est celle des groupes de réponses homogènes construits à partir de croisements des modalités des variables explicatives de la non-réponse. Ces groupes sont déterminés à l'aide de régressions logistiques impliquant différentes variables de niveau logement ou ménage disponibles pour l'ensemble des logements de l'échantillon. Seules sont retenues les variables significatives, et les groupes de réponses homogènes sont obtenus par croisement des modalités de ces variables.

6.2- Calage au niveau ménage

Le calage final mis en œuvre se base sur des données ménages mais fait l'hypothèse d'une équivalence logement/ménage. Il utilise la Macro Calmar et la méthode de calage par raking ratio.

Les marges de l'enquête Emploi (nous utilisons la dernière édition disponible à l'heure de notre analyse : la version 2011) nous permettent de caler en utilisant à la fois des informations de niveau individus et des informations de niveau ménage.

Sont utilisées pour le calage les variables suivantes :
- la composition et le type de ménage ;
- le diplôme et l'âge de la personne de référence (contrairement aux concepteurs de SRCV, nous n'utilisons pas les informations sur la CS) ;
- la strate de tirage du logement.

7 Exploitation des variables décomposant le revenu

Notre exploitation porte sur les postes de décomposition du revenu final disponible :
- les revenus d'activité (pour les individus salariés ou indépendants) ;
- les revenus des retraités du secteur privé (pour les individus bénéficiant d'une retraite du secteur privé) ;
- les allocations chômage (pour les individus bénéficiant d'allocations chômage) ;
- les revenus disponibles du ménage.

On exploite aussi la variable du montant des impôts versés par les ménages.

On utilise 2 jeux de pondérations différents, correspondant aux différents choix de « répondants ». Pour chacun des ces jeux de pondérations 4 statistiques sont calculées :
- la moyenne
- la médiane
- le 1er décile
- le 9ème décile

Effet de la non-prise en compte des répondants n'ayant répondu qu'après au moins 4 ou 5 contacts en

face à face

Salaire	Données non censurées (2111 FAs)	Sans les contactés+ de 4 fois en face à face (1945 FAs)	Sans les contactés plus de 3 fois en face à face (1722 FAs)
Moyenne	19821	1,1%	0,5%
Médiane	17688	0,7%	0,5%
1er décile	2868	0,6%	1,3%
9ème décile	35954	0,8%	-1,1%

Lecture : En supprimant les enquêtés ayant nécessité plus de 4 contacts en face à face pour répondre, l'estimateur de la moyenne des salaires est plus élevé de 1,1 %
En ne prenant en compte que les répondants contactés 4 fois ou moins, il n'en reste que 1 945 pour lesquels l'information est renseignée.

Chômage	Données non censurées (351)	Sans les contactés+ de 4 fois en face à face (322 FAs)	Sans les contactés plus de 3 fois en face à face (271 FAs)
Moyenne	6127	-3,6%	-6,1%
Médiane	4522	-3,9%	-6,6%
1er décile	822	-0,7%	-9,9%
9ème décile	13354	-1,5%	-1,7%

Lecture : En supprimant les enquêtés ayant nécessité plus de 4 contacts en face à face pour répondre, l'estimateur de la moyenne des allocations chômage est moins élevé de 3,6 %
En ne prenant en compte que les répondants contactés 4 fois ou moins, il n'en reste que 322 pour lesquels l'information est renseignée.

Retraites	Données non censurées (1143 FAs)	Sans les contactés+ de 4 fois en face à face (1095 FAs)	Sans les contactés plus de 3 fois en face à face (997 FAs)
Moyenne	17251	0,4%	-0,5%
Médiane	14953	0,9%	0,0%
1er décile	4965	1,7%	0,0%
9ème décile	30165	0,0%	2,0%

Lecture : En supprimant les enquêtés ayant nécessité plus de 4 contacts en face à face pour répondre, l'estimateur de la moyenne des retraites du secteur privé est plus élevé de 0,4 %
En ne prenant en compte que les répondants contactés 4 fois ou moins, il n'en reste que 1 095 pour lesquels l'information est renseignée.

Revenus disponibles	Données non censurées (2041 FAs)	Sans les contactés+ de 4 fois en face à face (1887 FAs)	Sans les contactés plus de 3 fois en face à face (1662 FAs)
Moyenne	36871	4,3%	2,7%
Médiane	30530	3,3%	2,2%
1er décile	13447	5,6%	3,7%
9ème décile	65425	0,6%	-2,1%

Lecture : En supprimant les enquêtés ayant nécessité plus de 4 contacts en face à face pour répondre, l'estimateur de la moyenne des revenus disponibles est plus élevé de 4,3 %
En ne prenant en compte que les répondants contactés 4 fois ou moins, il n'en reste que 1 887 pour lesquels l'information est renseignée.

Impôts	Données non censurées (1453 FAs)	Sans les contactés+ de 4 fois en face à face (1351 FAs)	Sans les contactés plus de 3 fois en face à face (1182 FAs)
Moyenne	3150	1,5%	1,6%
Médiane	1532	1,4%	0,3%
1er décile	214	0,4%	0,3%
9ème décile	6072	0,5%	0,0%

Lecture : En supprimant les enquêtés ayant nécessité plus de 4 contacts en face à face pour répondre, l'estimateur de la moyenne des impôts est plus élevé de 1,5 %
En ne prenant en compte que les répondants contactés 4 fois ou moins, il n'en reste que 1 351 pour lesquels l'information est renseignée.

La comparaison entre les estimateurs sur données réelles et les estimateurs sur données « censurées » trouve cependant ici sa limite. L'enquête SRCV est conçue avec en tête une logique d'exploitation transversale sur 9 vagues. Elle n'est pas pensée dans le but d'exploiter les données en se restreignant à une vague unique.
L'enquête Difficultés de Collecte ne prenant pour champ que la seule vague d'entrants en 2011, les estimateurs calculés sont associés à des intervalles de confiance (à 95 %) bien plus étendus que les variations crées par la simulation[3]. On se heurte ici à un problème de taille de l'échantillon, celui-ci est très petit pour permettre d'obtenir des données d'une précision suffisante à l'analyse.

Conclusion

La logique de simulation par « censure » permet *a priori* de mener une analyse en deux étapes :

1) L'analyse de l'impact du nombre de contacts sur les taux de non-réponse et la population de répondants
2) L'analyse de l'impact du nombre de contacts sur la qualité des estimateurs

Les données disponibles ne nous permettent cependant de traiter correctement que la première analyse. Elle révèle un effet élevé de l'effort de contact sur les taux de réponse et la composition des ménages répondants. Ces résultats ont conduit à ne pas ajouter de contraintes sur le nombre de contacts.
Les données ne sont cependant pas suffisamment nombreuses pour permettre de mener la deuxième analyse. La méthodologie ne pourra être appliquée de façon satisfaisante que sur des bases contenant une information plus complète sur le nombre de contacts effectués par les enquêteurs. La nouvelle interface des postes de collecte (qui deviendront des tablettes) qui sera utilisée à partir de 2016 (Projet Capi 3G) permettra de récupérer de façon systématique l'information sur le nombre de contacts.

Annexe : l'imputation des valeurs manquantes sur le nombre de contacts

Cette imputation utilise des données relatives aux enquêteurs provenant du logiciel Saige qui gère la paye des enquêteurs, ainsi que des données relatives aux fiches adresse. Pour chaque enquêteur, nous utilisons les données concernant son âge, son sexe, son nombre d'années d'expérience à l'Insee et la taille de sa région (au sens du nombre d'enquêteurs SRCV de la région). Les données utilisées sont le type de voisinage et le type de zone urbaine.
Pour imputer, on cherche d'abord à modéliser les variables au moyen d'un modèle de régression. Les variables ressortant significativement de nos modélisations sont le nombre d'enquêteurs SRCV de la région et le type d'habitation de la fiche adresse, ce sont ces variables que nous conservons

[3] Le plan d'échantillonnage des enquêtes ménages étant complexe, il est délicat de calculer les intervalles de confiance des estimateurs obtenus. Des analyse menées sur les enquêtes CVS et PIAAC montre que l'effet du plan de sondage sur les intervalles de confiance est de l'ordre de 1,5 à 1,9 (le plan d'échantillonnage des enquêtes produit des intervalles de confiance de 1,5 à 1,9 fois plus étendus qui si on utilisait un plan de sondage aléatoire simple).
On estime l'amplitude des intervalles de confiance des données en appliquant un multiplicateur de 1,5 à celle que l'on obtiendrait si les vagues d'entrants dans SRCV étaient tirées suivant un plan aléatoire simple.

dans nos imputations finales.
On préfère utiliser une modélisation aléatoire que déterministe, en effet dans le deuxième cas le nombre de contacts imputés n'est jamais suffisamment élevé pour impacter la suite de notre analyse (imputer de façon déterministe revient à ne rien faire et à considérer que toutes les fiches adresse pour lesquelles nous n'avons pas les données issues de Difficultés de collecte sont des fiches adresse où l'approche des ménages était facile et rapide). On ajoute donc des résidus simulés pour éviter cet effet.