L'UTILISATION DES HISTORIQUES D'APPELS POUR REDRESSER UNE ENQUÊTE TÉLÉPHONIQUE : UNE ÉTUDE PAR SIMULATION À PARTIR DE L'ENQUÊTE FECOND

Stéphane LEGLEYE[1] (*), Nirintsoa RAZAKAMANANA[2] (*), Géraldine CHARRANCE[3] (*) et Hélène JUILLARD[4] (*)

(*) Ined

Résumé

La non-réponse totale, malgré les efforts mis en place lors de la collecte, peut créer du biais et réduire la précision des estimateurs. Pour limiter le biais, il est courant de corriger l'échantillon par des méthodes de repondération (post-stratification ou calage). Ces traitements peuvent mobiliser des variables auxiliaires de la base de sondage, renseignées pour les répondants et non répondants ou bien des données externes concernant la population cible, dans le cas d'un calage sur marges. Lors de la conduite d'enquêtes téléphoniques aléatoires, il n'existe pas de base de sondage et donc pas de variables auxiliaires, en dehors des données décrivant le processus de collecte (paradonnées). Celles-ci sont constituées par les historiques d'appels des numéros mis en production (date et heure de l'appel, issue de l'appel, rang de l'appel). Elles sont complexes à traiter et c'est pourquoi seul un calage sur marges est en général effectué. Toutefois, elles contiennent de l'information prédictive de la participation (par définition), et sont potentiellement très liées aux réponses à l'enquête.

Nous proposons de montrer l'utilité des paradonnées en comparant quatre méthodes de traitement post-collecte d'un échantillon d'une enquête téléphonique : le calage direct et la correction de la non-réponse totale par groupes homogènes mobilisant les paradonnées suivant trois méthodes : la modélisation logistique, l'analyse en correspondance principale suivie d'une classification, enfin l'analyse harmonique, auxquelles un calage est ensuite appliqué. Ces quatre méthodes sont comparées sur des données simulées extraites d'une enquête réelle, l'enquête Fecond, réalisée par téléphone en 2010 par l'INSERM et l'INED où 8639 individus âgés de 15-49 ans ont rempli un questionnaire. Cet échantillon de répondants est considéré comme étant la population cible. La simulation repose sur cinq mécanismes de non-réponse totale avec un taux d'environ 50% : un mécanisme aléatoire pur (MCAR), trois mécanismes ignorables ou missing at random (MAR) (suivant les variables sociodémographiques utilisées pour le calage ; un jeu de paradata ; un mix des deux) et un mécanisme dépendant des variables d'intérêt (non missing at random –NMAR). Pour chacun, 1000 tirages aléatoires sans remise sont effectués. Biais et variance empiriques sont mesurés par rapport à la population cible.

Les résultats montrent que même en présence d'une faible corrélation entre paradonnées et variables d'intérêt, paradonnées et participation, il est préférable de recourir à une correction de la non-réponse utilisant les paradonnées avant calage. En revanche, la constitution des groupes homogènes par régression logistique est supérieure à celle opérée par classification. Une macro SAS paramétrableestdisponible pour répliquerl'analyse.

[1] stephane.legleye@ined.fr
[2] nirintsoa.razakamanana@ined.fr
[3] geraldine.charrance@ined.fr
[4] helene.juillard@ined.fr

Abstract

In telephone surveys, the classical way of weighting is a direct calibration on sociodemographic variables of the target population, aiming at reducing non-response bias and variance. However, paradata may be helpful in a preliminary phase of correction of total non-response if they are associated with the key variables of interest and with the participation. We test if this is the case within the telephone survey FECOND held in 2010 in France, focusing on sexual and reproductive health (SRH). Five methods using three sets of telephone paradata (using days and hour of call and issues of call: refusals, non-contact, appointments, beginning of questionnaire) are compared in the weighting process of a random survey, all based on homogenous response groups to correct for total non-response: logit modeling, clustering (based on the total calls or by call days), harmonic qualitative analysis.Five non-response mechanisms are defined, based on: sociodemographics (MAR 1), paradata (MAR 2), mix of paradata and sociodemographics (MAR 3), variable of interest (NMAR) and completely at random (MCAR). 1000 replications are generated, with an approximate participation rate of 50%. The variables used for the calibration procedure are the same the one used in MAR1 and MAR3. Results show that the direct calibration is never the optimal weighting procedure. The gain in mean square of error is low in the MAR1 mechanism but big in the MAR3 mechanism while it is negligible in the NMAR mechanism. The choice of the type of paradata is of some importance, while the use of clustering methods to build the homogenous response groups is not superior to the classical logit modeling. Results and limitations are discussed. The parametrizable SAS macro that can handle a large variety of non-response mechanisms into account is available on demand.

Mots-clés

Estimation avec non-réponse, historiques d'appels, mécanismes de non-réponse, paradonnées, simulations

1. Introduction

1.1 Contexte

Le téléphone est théoriquement un moyen très efficace de joindre la population et donc de mener des enquêtes en population générale. En 2011, plus de 74% des individus âgés de 12 ans et plus possédaient à la fois un téléphone filaire et un GSM, tandis que 15% n'avaient qu'un filaire et 10% qu'un GSM: ainsi, moins d'un pourcent de la population était injoignable par téléphone (Bigot and Croutte 2011). Toutefois, le paysage de la téléphonie est devenu très complexe ces dernières décennies : si le nombre de téléphones mobiles s'est considérablement accru, une proportion croissante de ménages (58%) est également connectée à Internet via des installations téléphoniques filaires. La couverture de la population est donc croissante, mais joindre des individus dans des conditions satisfaisantes pour mener des enquêtes aléatoires est devenu plus complexe : la multiplication des opérateurs, la disparition de l'annuaire unique des abonnés du téléphone filaire de l'opérateur historique, l'absence d'annuaire d'abonnés de mobiles, la multiplication des numéros pour joindre un individu unique, la part croissante de numéros dégroupés (dont le numéro ne renseigne pas sur la localisation géographique de l'abonné) et la variété des types d'équipement en sont les principales causes (Gautier, Beck et al. 2006; Brick, Cervantes et al. 2011). De plus, les enquêtes voient leur taux de participation global baisser (Lan 2009; Schouten, Cobben et al. 2009). Ceci peut être dû à la sollicitation marketing importante à laquelle sont soumis les abonnés téléphoniques sur filaire comme sur mobile, un désintérêt pour les enquêtes, l'absence de temps et de disponibilité et la

possibilité de filtrage des appels entrants à partir de l'affichage du numéro. Cette non-participation génère des coûts supplémentaires (multiplication des tentatives d'appels avant d'obtenir une acceptation, allongement de la durée de la collecte) et est potentiellement la cause de biais importants. Ces biais résultent de deux facteurs : le taux de non-participation et la différence de comportement entre les participants et les non-participants (Groves 2006).
L'estimation des biais de non-réponse nécessite donc la connaissance des comportements d'intérêt de l'enquête déclarés par les non-répondants, ce qui est par définition impossible. Comme il n'existe pas de base de sondage téléphonique en raison de l'absence d'annuaire universel, il n'existe pas de données auxiliaires de type géographique ou socio-démographique disponibles pour les non-répondants (sauf pour les numéros présents dans l'annuaire, soit environ 45%). Les seules variables auxiliaires disponibles sont les paradonnées, c'est-à-dire les informations produites lors de la collecte : les dates et heures d'appel, le type de numéro (mobile ou fixe, trouvé dans l'annuaire ou pas), le rang d'appel et l'issue d'appel. Ces paradonnées sont produites par la collecte mais ne sont pas a priori destinées au traitement statistique de l'échantillon des répondants etdes non-répondants : ce sont plutôt des données de gestion du terrain téléphonique. Elles sont abondantes et assez complexes à manipuler et peu étudiées dans ce but.
La correction des biais de non-réponse dans les enquêtes téléphoniquesrepose donc toujours (à notre connaissance), sur une procédure simple de repondération post-collecte, fondée sur l'usage de données auxiliaires (quasi-exclusivement des variables sociodémographiques) dont les marges dans la population cible sont connues : calage sur marges ou post-stratification.
Les paradonnées téléphoniques recèlent pourtant potentiellement une information utile au redressement. D'abord, elles sont, par construction, très liées à la non-réponse, puisqu'elles sont en quelque sorte la description du processus de contact avecchaque individu pouvant être joint à l'aide d'un numéro de téléphone. Mais elles renseignent aussi sur la disponibilité d'un ménage ou d'un individu (le téléphone décroche seulement certains jours ou à certaines heures), sur son accessibilité téléphonique (il faut un plus ou moins grand nombre de tentatives d'appels pour un décroché) et sa bienveillance à l'égard de l'enquête (présence systématique d'un répondeur, filtrage du numéro donc non-contacts répétés, refus initial, refus sans sélection de la personne au sein du foyer, rendez-vous, virulence du refus, etc.). Ces caractéristiques reflètent certains aspects du mode de vie des individus : mode de vie allocentré, horaires de travail décalés, bienveillance à l'égard des enquêtes, etc. Par cet intermédiaire, elles peuvent être liées aux variables d'intérêt de l'enquête. Par ailleurs, ce sont des caractéristiques qui ne sont a priori pas fortement déterminées par les variables socio-démographiques classiquement utilisées en repondération (post-stratification, calage) car elles reflètent potentiellement des attitudes irréductibles à des caractéristiques comme l'âge, le sexe ou la composition du foyer.

1.2 Objectifs

Nous proposons ici une première évaluation de l'utilité de l'usage des paradonnées téléphoniques dans une étude de simulation tirée d'une enquête réelle en complément d'un calage sur marges. Nous faisons l'hypothèse que toutes les paradonnées ne sont pas équivalentes en raison de l'information qu'elles contiennent, et que certaines méthodes statistiques de correction de la non-réponse sont préférables à d'autres sur un même jeu de paradonnées.
La plupart des travaux portant sur l'usage de variables auxiliaires ou de paradonnées pour la correction de la non-réponse reposent sur une simulation complète des données (Little and Vartivarian 2005; Haziza and Beaumont 2007). Elles ne prennent pas en compte l'étape de calage qui est très souvent proposé en deuxième étape du redressement ni la multiplicité des variables de l'enquête. L'originalité de notre travail est de se placer dans le cadre particulier d'une enquête précise, afin de fournir quelques éléments permettant de décider s'il est judicieux de procéder à une correction de la non-réponse-totale avec les paradonnées disponibles. A terme, nous ambitionnons de fournir un outil de simulation permettant d'évaluer les gains potentiels dans des enquêtes réelles, en tenant compte des relations réelles entre paradonnées, non-réponse, variables d'intérêt et variables sociodémographiques.

1.3 Cadre de travail

L'échantillon des répondants de l'enquête initialeFecondest utilisé comme population cible. Nous nous plaçons dans un cadre sans sondage : une enquête exhaustive a lieu et la non-participation à cette enquête est provoquée par certains mécanismes de non-réponse. Nous avons choisi cette approche

pour deux raisons : nous ne disposons pas d'échantillon de taille suffisante pour simuler une première phase de sondage suivie d'une phase de non-réponse ; nous préférons également travailler dans un cadre aussi proche du réel que possible afin de restituer la complexité des relations entre variables d'intérêt, paradonnées et variables sociodémographiques.

1.4 Plan d'analyse

Dans un premier temps, nous présenterons l'enquête, les variables d'intérêt, les variables sociodémographiques utilisées pour le calage sur marges et les paradonnées ainsi que les liens existant entre ces trois types de données. Dans un deuxième temps, nous présenterons les méthodes de traitement de la non-réponse totale mobilisées. Dans un troisième temps, nous exposerons les mécanismes de non-réponse étudiés. Enfin, les méthodes seront comparées dans une simulation de non-réponse à partir de l'enquête. La comparaison portera sur le calcul de pseudo-variances, de pseudo-biais et d'écarts quadratiques moyens. Les limites de l'exercice et la portée des résultats, ainsi que les recommandations pour les études à venir, seront ensuite discutées.

2. Présentation de l'enquête, des variables sociodémographiques, d'intérêt et des paradonnées

2.1 L'enquête

Les données proviennent de l'enquête téléphonique Fecond, portant sur la fécondité, la santé sexuelle et reproductive, menée en 2010 par l'Unité mixte Inserm-Ined(Bajos, Bohet et al. 2012). Cette enquête reprend la suite des enquêtes déjà menées depuis 1992 sur le sujet (Bajos, Spira et al. 1992; Bajos and Bozon 2009). Elle repose sur un sondage aléatoire à deux degrés (ménage puis individu) stratifié (téléphones filaires et mobiles). La population éligible est francophone et sans handicap (suffisamment apte pour la passation d'un questionnaire téléphonique) et âgée de 15 à 49 ans. Le protocole imposait un minimum de 20 appels sans contact humain (i.e. sans décroché ou avec répondeur) ou avec rendez-vous systématique avant abandon et les refus étaient rappelés deux fois maximum afin de tenter de les convertir en acceptation. Si un contact avait lieu avant 20 appels, 20 autres étaient possibles. Le nombre maximal de tentatives d'appels n'était donc pas limité (le maximum atteint 373, mais concerne un cas unique et le nombre total maximald'appelsnécessaire à un questionnaire est 158. Dans les analyses suivantes, le nombre d'appels est tronqué à 50 (la dernière issue étant reportée).

Pour constituer cet échantillon, des numéros ont été générés au hasard et filtrés sur les racines ouvertes à l'exploitation commerciale par l'ARCEP (Autorité de régulation des communications électroniques et des postes) : 92516 numéros furent ainsi exploités. Le taux de participation final s'élève à 44.8%, le taux de refus/abandon à 20.2%, le taux de non-contact à 29.9% et le taux d'impossibles à enquêter à 5.1%. Le taux de non-contact est très élevé, car une partie des numéros utilisés correspondencore à des artisans très difficiles à joindre ou à des numéros de box, tous deux hors cible mais difficilement identifiables comme tels : le taux final de participation est donc sans nul doute sous-estimé. La méthodologie a été présentée en détail dans (Legleye, Charrance et al. 2013).

L'échantillon final comprend 8645 répondants : 6232 entretiens ont été passés avant 20 appels et sans refus (72.1%) ; 1557 ont été passés après un refus mais avec moins de 20 appels (18.0%) ; 772 ont été passés sans refus après 20 appels (8.9%) et 84 ont été passés après un refus et plus de 20 appels (0.9%). Au total, 27.9% des entretiens ont donc été passés après un refus ou plus de 20 appels, et plus spécifiquement 9.9% après 20 appels. La taille de l'échantillon de répondants analysé dans cette étude s'élève à 8639 (sur les 8645 initiaux) : les questionnaires exclus comprennent des valeurs manquantes sur quelques paradonnées.

Dans tout ce qui suit, nous n'utilisons aucune pondération initiale, considérant que notre étude porte sur un recensement.

2.2 Les variables sociodémographiques

L'enquête a donné lieu à un calage sur marges sur la population cible du recensement de la population (2008) mobilisant les 8 variables sociodémographiques suivantes : sexe, âge (7 tranches d'âge),

diplôme le plus élevé (5 niveaux, d'inférieur au BEPC jusqu'à Bac +4 et au-delà), zone géographique de résidence (Île-de-France et Bassin parisien, Nord-Est-Ouest, autre), lieu de naissance (en France, à l'étranger), situation professionnelle (emploi ou non), vie de couple (en couple ou non), taille du ménage (1 personne, 2, 3-4, 5 et plus).
Dans le calage, la variable vie de couple a une troisième modalité (en couple, avec ses parents, autre). Pour la génération du mécanisme de non-réponse associé aux variables sociodémographiques, la zone géographique de résidence a été omise, et la variable vie de couple comporte deux modalités uniquement.

2.3 Les paradonnées

Les paradonnées recueillies comprennent la date et l'heure d'appel et l'issue de l'appel. Les dates et heures ont été converties en plage horaire, définie par le croisement du jour de la semaine et de l'horaire d'appel. Après examen des dates et horaires d'appels passés sur les répondants, trois types de jours et trois types d'horaires ont été retenus : lundi, mardi et jeudi ; mercredi et vendredi ; samedi ; 9h-16h ; 16h-19h, 19h-21h. Cette combinaison définit ainsi 9 plages horaires. Les issues d'appels sont initialement au nombre de 54 ; après examen d'une analyse factorielle multiple opérée sur plusieurs typologies concurrentes plus ou moins détaillées, 4 issues ont été conservées pour cette étude : refus, rendez-vous, non-contact, passation (partielle ou totale). Trois jeux de paradonnées sont proposés dans ce document :

1. P1 : le premier jeu de paradonnées défini pour cette étude est une synthèse anhistorique des issues d'appels : il se compose de 4 variables totalisant le nombre d'appels pour chacune des quatre issues retenues, refus, rendez-vous, non-contact, passation (partielle ou totale).
2. P2 : le deuxième jeu détaille le premier en ventilant le nombre d'issues par plage horaire : il comprend donc 36 variables et est également anhistorique.
3. P3 : le troisième jeu de paradonnées repose sur une reconstitution simplifiée de l'historique d'appels, tenant compte de la succession des issues d'appels. Tous les numéros n'ayant pas été appelés 50 fois, les séquences courtes sont complétées jusqu'à 50 par des appels fictifs dont l'issue également fictive est « non-concerné ». Le rang d'appel varie alors de 1 à 50 pour tous les numéros exploités et représente le compteur de temps ; cette durée est découpée en blocs. Ces blocs sont définis après description de la distribution des appels : 1-3, 4-9, 10-19, 20-29, 30-50. Pour chaque bloc, on définit la proportion de temps passé dans chacune des 5 issues (les 4 réelles plus l'issue fictive « non-concerné ») : il y a donc 25 variables. Ce troisième jeu de paradonnées permet, au moyen d'une analyse factorielle des correspondances, d'établir une analyse harmonique qualitative (AHQ) des trajectoires (ou historiques) d'appels(Deville and Saporta 1979).

Les jeux de paradonnées 1, 2 et 3 seront analysés par régression (modélisation logistique de la participation) et par classification (au moyen de K-means opérées sur les premières composantes d'une analyse en composantes principales).

2.4 Les variables d'intérêt

Les variables d'intérêt sont au nombre de 9 : Elles sont toutes binaires.
Les valeurs manquantes sont très peu nombreuses (moins d'1% pour chaque variable) : elles ont été recodées à 0.

3. Liens entre paradonnées, variables sociodémographiques et d'intérêt

La non-réponse totale, malgré les efforts mis en place lors de la collecte, peut créer du biais et réduire la précision des estimateurs. Pour limiter le biais, il est courant de corriger l'échantillon par des méthodes de repondération (post-stratification ou calage). Certains utilisent les informations contenues dans la base de sondage, disponibles pour les répondants et les non-répondants, afin de modéliser la participation et de corriger la non-réponse totale. Un autre traitement peut consister à caler l'échantillon de répondants sur la structure d'une population cible (par exemple une enquête de

référence ou le recensement de la population) ; il peut être opéré directement ou bien après le premier traitement, s'il mobilise des variables différentes. Les deux techniques consistent à repondérer l'échantillon de répondants afin que sa structure et celle de la population d'inférence soient identiques.

Dans les enquêtes téléphoniques par génération aléatoire de numéros, la base de sondage n'existe pas. Ainsi, ordinairement, seul le calage direct est utilisé. Pour redresser un échantillon de façon efficace, sans augmenter la variance, il faut disposer d'une information auxiliaire possédant quatre propriétés (Little and Vartivarian 2005; Sarndal and Lundstrom 2005) :

1. Etre disponible sans valeur manquante sur les répondants et les non-répondants
2. Etre mesurée sans erreur
3. Etre associée aux comportements d'intérêts mesurés dans l'enquête
4. Etre un prédicteur de la participation à l'enquête

Dans une enquête téléphonique à génération aléatoire de numéros de téléphone, donc sans base de sondage, les historiques d'appels, comprenant pour chaque numéro la succession des issues des appels (décrites suivant une certaine typologie), vérifient le point 1, et assez largement le point 2, car une large partie de l'information est codée sans erreur par l'automate d'appel (ligne occupée ou non, rang de l'appel), et l'enquêteur (refus, sans réponse, demande de rendez-vous, début de passation). Ils vérifient sans peine le point 4, presque par définition. Qu'en est-il du point 3 ? Dans les enquêtes téléphoniques, les ménages et les personnes qui refusent l'enquête sont souvent rappelés une fois, dans l'espoir de les joindre à un moment ultérieur plus opportun, ou de réussir à joindre une autre personne du ménage afin de la convaincre de participer (ou le cas échéant de nous aider à établir l'éligibilité de son ménage). Des analyses ont permis de montrer que le refus initial ou le nombre d'appels nécessaires à la passation d'un questionnaire sont associés à des profils sociodémographiques particuliers, mais aussi à des profils de réponse particuliers à des variables d'intérêt (Beck, Legleye et al. 2005; Durrant and Steele 2009; Legleye, Charrance et al. 2013).

Plus précisément dans notre étude, une analyse en composante principale (ACP) a été opérée sur les jeux de paradonnées 1 et2, les variables d'intérêt et les variables sociodémographiques (en ôtant la zone de résidence) et l'AHQ a été faite sur le jeu de paradonnées 3.

- Pour les variables sociodémographiques (SD), les trois premières composantes totalisent 60.3% (28.1%, 17.9%, 14.4%) de la variance totale ;
- pour les variables d'intérêt (VI) on obtient 53.9% (22.2%, 16.9%, 14.8%) ;
- pour le jeu de paradonnées 1 (P1) on obtient 84.2% (34.0%, 27.3%, 22.9%) ;
- pour le jeu de paradonnées 2 (P2) on obtient 29.9% (15.3%, 11.0%, 3.5%) ;
- pour le jeu de paradonnées3 (P3), on obtient 56.3% (31.8%, 14.2%, 10.2%).

Afin de présenter une synthèse lisible, les trois premières composantes ont été conservées et la matrice des corrélations a été calculée (tableau 1). L'examen de la matrice de corrélation montre que les premiers axes des paradonnées sont très corrélés entre eux, ce qui signifie que les trois jeux de paradonnées présentent une première dimension commune. Les corrélations entre les seconds axes sont plus modestes, sauf entre les plages horaires et l'historique (paradonnées 2 et3).
Les liens sont relativement importants entre les variables d'intérêts (VI) et les variables sociodémographiques (SD) : 8 corrélations sur 9 sont significatives, le coefficient variant entre 0.038 et 0.347, 5 sur 8 étant supérieurs à 0.1.
Les liens entre paradonnées et SD sont moins nombreux et plus modérés : 12 corrélations sur 27 sont significatives, celles-ci étant comprises entre 0.021 et 0.177 (3 étant supérieures à 0.1). Les liens sont surtout concentrés sur l'axe 2 des SD (8 corrélations, dont les 3 plus élevées).
Enfin, les liens entreles VI et les paradonnées sont également modérés : seules 11 corrélations sur 27 sont significatives, celles-ci étant comprises entre 0.018 et 0.058 en valeur absolue (aucune ne dépasse 0.1). Le jeu P3 est le plus lié (5 corrélations significatives sur 9), devant le jeu P1 (4 corrélations significatives sur 9) et le jeu P2 (3 corrélations significatives sur 9).

Pour compléter cette analyse, nous nous sommes demandés s'il existait une liaison propre entre les jeux de paradonnées et les variables d'intérêt lorsque les variables sociodémographiques sont contrôlées. Une réponse positive pourrait signer l'utilité potentielle d'utiliser les paradonnées en sus des variables sociodémographiques pour le redressement de l'échantillon, autrement dit préalablement au calage.

Le tableau 2 présente la synthèse des régressions linéaires des trois premiers axes de l'ACP des variables d'intérêt sur les trois premiers axes des jeux de paradonnéesP1 à P3 ajustés sur les trois premiers axes des données sociodémographiques.

Tableau 1 : Matrice des corrélations des premiers axes factoriels des jeux de paradonnées, de variables sociodémographiques et d'intérêt : rho de Pearson et valeur-p du test de nullité

	P2_1	P2_2	P2_3	P3_1	P3_2	P3_3	SD_1	SD_2	SD_3	VI_1	VI_2	VI_3
P1_1	**0.773**	**0.353**	**0.141**	**0.665**	**0.454**	**0.297**	0.014	**0.062**	**0.028**	**0.053**	-0.021	0.002
	<.0001	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*0.182*	*<.0001*	*0.009*	*<.0001*	*0.0515*	*0.886*
P1_2	**0.403**	**-0.635**	**0.326**	**0.493**	**-0.521**	**-0.072**	**-0.032**	**-0.071**	-0.021	-0.009	**0.030**	0.000
	<.0001	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*0.0025*	*<.0001*	*0.0543*	*0.4287*	*0.0053*	*0.9975*
P1_3	**-0.458**	**0.235**	**0.480**	**-0.347**	**0.306**	**-0.350**	-0.007	**0.149**	**-0.028**	**-0.054**	**-0.039**	-0.017
	<.0001	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*<.0001*	*0.4921*	*<.0001*	*0.009*	*<.0001*	*0.0003*	*0.1159*
P2_1	1	0	0	**0.860**	**0.071**	**0.405**	-0.005	**-0.035**	0.021	**0.057**	0.012	0.004
		1	*1*	*<.0001*	*<.0001*	*<.0001*	*0.6223*	*0.0010*	*0.0525*	*<.0001*	*0.2664*	*0.6935*
P2_2		1	0	**-0.189**	**0.860**	**0.243**	-0.005	**0.135**	-0.003	-0.007	**-0.036**	-0.019
			1	*<.0001*	*<.0001*	*<.0001*	*0.6107*	*<.0001*	*0.9761*	*0.5076*	*0.0009*	*0.0778*
P2_3			1	**0.117**	**0.052**	**-0.177**	-0.005	**0.021**	0.001	**-0.029**	-0.006	-0.001
				<.0001	*<.0001*	*<.0001*	*0.9562*	*0.0497*	*0.8989*	*0.007*	*0.5791*	*0.8921*
P3_1				1	0	0	-0.000	**-0.058**	0.018	**0.058**	0.013	0.006
					1	*1*	*0.9674*	*<.0001*	*0.0902*	*<.0001*	*0.2346*	*0.5833*
P3_2					1	0	-0.001	**0.177**	-0.016	-0.009	**-0.047**	**-0.022**
						1	*0.9201*	*<.0001*	*0.1281*	*0.3906*	*<.0001*	*0.0381*
P3_3						1	-0.014	-0.021	**0.029**	**0.029**	0.017	0.006
							0.1937	*0.0548*	*0.0073*	*0.0073*	*0.1239*	*0.589*
SD_1							1	0	0	**0.347**	**-0.102**	**0.121**
								1	*1*	*<.0001*	*<.0001*	*<.0001*
SD_2								1	0	**-0.180**	**-0.102**	**-0.068**
									1	*<.0001*	*<.0001*	*<.0001*
SD_3									1	**-0.038**	-0.011	**-0.049**
										0.0004	*0.3281*	*<.0001*

Les corrélations linéaires significatives au seuil 0.05 sont en gras ; celles dépassant 0.1 en valeur absolue sont dans des cases grisées.
P1_n, P2_n, P3_n, SD_n, VI_n : axe principal numéro n des ACP (ou de l'AHQ pour le jeu de Paradonnées 3) des jeux de Paradonnées 1, 2, 3, des variables sociodémographiques et des variables d'intérêt.

Tableau 2 : Régression linéaire des trois premiers axes factoriels des variables d'intérêt (VI_1, VI_2, VI_3) sur les trois premiers axes de chaque jeu de paradonnées, (Pi_1, Pi_2, Pi_3, *i*=1, 2, 3), ajustés sur les trois axes des variables sociodémographiques (SD_1, SD_2, SD_3)

	VI_1			VI_2			VI_3		
	Coeff.	Sd	Valeur-p	Coeff.	Sd	Valeur-p	Coeff.	Sd	Valeur-p
Intercept	0.000	0.014	1	0.000	0.013	1	0.000	0.012	1
P1_1	**0.067**	0.012	<.0001	-0.015	0.011	0.197	0.006	0.011	0.5643
P1_2	-0.007	0.013	0.5837	0.024	0.013	0.0576	-0.003	0.012	0.7835
P1_3	**-0.049**	0.015	0.0008	**-0.033**	0.014	0.0166	-0.003	0.013	0.8238
SD_1	**0.391**	0.010	<.0001	**-0.083**	0.009	<.0001	**0.109**	0.009	<.0001
SD_2	**-0.156**	0.013	<.0001	**-0.102**	0.012	<.0001	**-0.098**	0.011	<.0001
SD_3	**-0.046**	0.014	0.0009	-0.004	0.013	0.788	**-0.025**	0.012	0.0408
Intercept	0.000	0.014	1	0.000	0.013	1	0.000	0.012	1
P2_1	**0.033**	0.006	<.0001	0.004	0.006	0.4439	0.001	0.005	0.8448
P2_2	0.008	0.007	0.2352	**-0.015**	0.007	0.0291	-0.003	0.006	0.6046
P2_3	**-0.033**	0.012	0.0077	-0.004	0.012	0.7065	0.001	0.011	0.9525
SD_1	**0.393**	0.010	<.0001	**-0.084**	0.009	<.0001	**0.109**	0.009	<.0001
SD_2	**-0.157**	0.013	<.0001	**-0.105**	0.012	<.0001	**-0.097**	0.011	<.0001
SD_3	**-0.044**	0.014	0.0015	-0.004	0.013	0.7698	**-0.025**	0.012	0.043
Intercept	0.000	0.014	1	0.000	0.013	1	0.000	0.012	1
P3_1	**0.099**	0.019	<.0001	0.012	0.018	0.488	0.001	0.017	0.9341
P3_2	0.036	0.029	0.21	**-0.077**	0.027	0.0045	-0.014	0.025	0.5783
P3_3	**0.102**	0.033	0.002	0.043	0.031	0.167	0.016	0.029	0.5844
SD_1	**0.393**	0.010	<.0001	**-0.084**	0.009	<.0001	**0.109**	0.009	<.0001
SD_2	**-0.156**	0.013	<.0001	**-0.102**	0.012	<.0001	**-0.097**	0.011	<.0001
SD_3	**-0.044**	0.014	0.0014	-0.005	0.013	0.7152	**-0.025**	0.012	0.0413
R^2	0.1717			0.020			0.0271		
	0.1713			0.019			0.0271		
	0.1712			0.020			0.0271		

En gras, les coefficients significatifs au seuil 0.05. Un modèle a été effectué par jeu de paradonnées.

Il subsiste des liens significatifs avec chaque jeu de paradonnées (3 liens sur 9 possibles pour chaque modèle). Ils sont les plus forts pour le jeu 3 (historique d'appels traité en AHQ), intermédiaires pour le jeu P1 (bilan anhistorique) et les plus faibles pour le jeu 2 (bilan anhistorique par plage horaire). C'est aussi avec l'axe 1 des variables d'intérêt que les liens sont les plus forts et nombreux (2 sur 3). Mais les 3 premiers axes de paradonnées sont significatifs pour chaque modèle : axes 1 et 3 pour le premier axe des variables d'intérêt, axe 2 pour l'axe 2 des variables d'intérêt.
Les axes des variables sociodémographiques sont, par contraste, nettement plus souvent significatifs et leurs effets nettement plus forts.

La part de variance expliquée par le jeu P1 qui n'est pas expliquée par les SD s'élève à 2.4% pour l'axe VI_1, 6.0% pour l'axe VI_2et à 0.4% pour VI_3. Soit une part de variance pondérée (par les valeurs propres) expliquée par le jeu P1 pour les 3 premiers axes des variables d'intérêt approximativement égale à 3.0%.
La part de variance expliquée par le jeu P2 non expliquée par les SD s'élève à 2.2% pour VI_1, 3.1% pour VI_2 et 0.4% pour VI_3. Soit une part de variance pondérée expliquée de 2.0%
Les chiffres correspondants pour le jeu P3 sont : 2.2%, 6.0% et 0.4%, soit une part totale de 2.9%.

Le bilan anhistorique (jeu P1) et les historiques d'appels sont donc a priori les données qui sont les plus utiles pour une correction de la non-réponse totale en amont d'un calage sur les variables

sociodémographiques utilisées dans cette étude. Le point 3 est donc bien vérifié, bien que l'intensité de la liaison ne soit pas très forte.
Ainsi, une post-stratification ou un calage sur les variables sociodémographiques omettent très vraisemblablement une information importante : la difficulté à joindre et à convaincre les individus. Utiliser les paradonnées peut donc être utile pour une correction de la non-réponse totale, éventuellement préalable à un calage.

4. Techniques de correction de la non-réponse totale

Notre population cible est celle des répondants (n=8639). Il n'y a pas de pondération. La méthode classique de redressement et de correction de la non-réponse totale dans les enquêtes téléphoniques est le calage sur marges direct (méthode 0). Cette méthode ne mobilise pas les paradonnées. Plusieurs méthodes lui sont comparées : tout d'abord cinq méthodes de correction de la non-réponse totale par Groupes de Réponses Homogènes (GRH), puis ces cinq mêmes méthodes suivies d'un calage.
La première méthode de correction de la non-réponse totale est celle des GRH utilisant la modélisation logistique de la participation à partir des paradonnées. L'autre, appartenant aussi à la famille des GRH, repose sur l'élaboration d'une typologie de répondants et de non-répondants définie à partir de leurs paradonnées. Cette classification est opérée sur toutes les composantes d'une analyse en composantes principales (ACP) normée. L'intérêt potentiel est de tenir compte de toutes les interactions entre les composantes retenues, ce qui n'est pas fait dans un modèle logistique simple et, au final, de produire des groupes de réponse homogènes plus homogènes que ceux obtenus par modélisation logistique[5].La classification est ensuite opérée par K-means, de façon à ce que 7 groupes d'au moins 400 individus soient définis[6]. Nous faisons l'hypothèse que cette méthode estsupérieure à la régression logistique simple[7]. On décline ces deux méthodes en faisant varier le jeu de paradonnées (P1 et P2). Une cinquième méthode est proposée, utilisant l'analyse harmonique sur le troisième jeu de paradonnées : comme précédemment, les GRH sont obtenus par K-means sur les composantes de l'AHQ. Les cinq méthodes se présentent donc comme suit :

- Méthode 1, acP_P1 : ACP sur groupe de paradonnéesP1 (bilan anhistorique)
- Méthode 2, log_P1 : régression logistique sur groupe de paradonnéesP1 (bilan anhistorique)
- Méthode 3, acP_P2 : ACP sur groupe de paradonnéesP2 (plages horaires)
- Méthode 4, log_P2 : régression logistique sur groupe de paradonnéesP2 (plages horaires)
- Méthode 5, ahq_P3 : analyse harmonique sur groupe de paradonnéesP3 (trajectoires d'appel).

Ces cinq méthodes sont suivies d'un calage sur les données sociodémographiques (même jeu de variables sociodémographiques) utilisant la méthode du raking ratio.
Deux autres méthodes témoin sont également utilisées : d'abord une méthode 00, qui est l'absence de tout redressement ; ensuite, la méthode 0 qui est la méthode de référence dans les enquêtes téléphoniques et qui consiste en un calage direct sur les données sociodémographiques sans correction de la non-réponse. Les méthodes 1 bis à 5 bis seront comparées aux méthodes 0 et 00.

- Méthode 1 bis, calage_acp_P1 : Méthode 1 suivie d'un calage
- Méthode 2 bis, calage_log_P1 : Méthode 2 suivie d'un calage
- Méthode 3 bis, calage_acp_P2 : Méthode 3 suivie d'un calage
- Méthode 4 bis, calage_log_P2 : Méthode 4 suivie d'un calage
- Méthode 5 bis, calage_ahq_P3 : Méthode 5 suivi d'un calage

[5]Ce type de méthode a été utilisé dans une étude de simulation complète de données et s'est révélé un peu moins performante que la méthode des scores(Haziza, D. and J.-F. Beaumont (2007). "On the Construction of Imputation Classes in Surveys." International statistical review**75**(1): 25-43.) ; toutefois, dans nos analyse préalable sur nos données, nous avions plutôt trouvé des arguments en sa faveur.
[6] Ce nombre est arbitraire mais assure que tous les GRH contiennent des répondants et des non répondants en nombres suffisants.
[7] Un modèle logistique avec interaction soigneusement construit pour maximiser le pouvoir explicatif peut se montrer supérieur à cette méthode, mais son établissement prendrait trop de temps dans le cadre de cette simulation et l'on ne serait pas à l'abri d'une mauvaise spécification du modèle –multicolinéarité par exemple-. L'intérêt de la classification opérée sur ACP est qu'elle ne nécessite aucune condition de validité.

- Méthode 0, calage_direct : calage direct, méthode témoin
- Méthode 00, aucun redressement, méthode témoin

5. Génération des mécanismesde non-réponse, mise en œuvre et estimation des performances

5.1 Génération de la non-réponse

Nous utilisons les données des 8639 répondants, que nous considérons comme la population statistique, c'est-à-dire les données du questionnaire et les paradonnées, pour obtenir un échantillon de répondants. Pour simuler la non-réponse au sein de cette population, nous avons programmé cinq mécanismes : un mécanisme complètement au hasard (MissingCompletely at Random, MCAR), trois mécanismes aléatoires ou ignorables (Missing At Random, MAR) et un mécanisme de non-réponse non aléatoire ou non ignorable (Not Missing At Random, NMAR).

- Pour le mécanisme **MCAR**, on suppose des probabilités uniformes.
- Les mécanismes MAR, sont fondés sur l'utilisation du premier axe d'une ACP définie sur un set de variables à partir duquel on retient les quintiles, qui partagent notre population en cinq strates.
    - On note **MAR1** le mécanisme utilisant le jeu de 7 variables sociodémographiques présentées précédemment : sexe, classe d'âge –7 modalités-, situation professionnelle –actif occupé, chômeur/inactif-, diplôme –5 modalités-, vie de couple –en couple, ou non-, taille du ménage –4 classes-, lieu de naissance –France, étranger.
    - On note **MAR2**, le mécanisme utilisant le jeu de paradonnéesP1 : somme des non-contacts, refus, rendez-vous, passation.
    - On note **MAR3**, le mécanisme utilisant à la fois les jeux de paradonnéesP1 et P2 (la somme des deux premiers axes est ici utilisée pour définir les quintiles)[8]. Il est une sorte de compromis entre MAR1 et MAR2.
- Pour le mécanisme **NMAR**, une ACP de la population est également effectuée avec découpage au quintile, mais cette fois-ci à partir des 8 variables d'intérêt de l'enquête : rapport sexuel au cours de la vie avec un partenaire du sexe opposé, avec un partenaire du même sexe ; rapport sexuel au cours des douze derniers mois ; cinq partenaires sexuels du même sexe, du sexe opposé (au cours de la vie) ; dernier rapport sexuel avec le partenaire régulier ; même partenaire pour toutes les grossesses ; relation sexuelle imposée au cours des douze derniers mois ; avoir causé ou avoir vécu une interruption volontaire de grossesse.

Pour chaque mécanisme, chaque individu k de la population se voit attribuer une probabilité p_k de réponse. Relativement à un mécanisme donné, les indicatrices de réponse sont générées par un tirage de Bernoulli de paramètre p_k dans chaque strate. Pour chacun des cinq mécanismes, les probabilités p_k ont été choisies telles que le taux de non-réponse soit égal (en moyenne) à 50 % : 0.35, 0.40, 0.5, 0.6, 0.75 pour les quintiles des premiers axes.

[8]Le premier axe se définit essentiellement par les variables socio-démographiques (15.9% d'inertie) et l'axe 2 par les paradonnées (14.0% d'inertie) et les plages de variation sont presque identiques ; ce relatif équilibre d'interprétation et d'inertie permet d'assurer que la somme des deux axes représente un bon compromis des deux mécanismes MAR1 et MAR2.

5.2 Mise en œuvre logicielle

Tous les traitements ont été effectués avec le logiciel SAS, sur le serveur LINUX de calcul de l'Ined. Les simulations de non-réponse ont été opérées au sein d'une étape DATA par des tirages de nombres aléatoires suivant une loi uniforme. La correction de la non-réponse totale par modélisation logistique a été effectuée avec la procédure *logistic* ; celle utilisant la typologie avec les procédures *princomp* et *fastclus*pour définir 7 classes avec la contrainte d'au moins 400 individus par classe. MAR3 a été obtenu par la procédure*corresp* pour l'AHQ suivi de la procédure*fastclus* avec le paramétrage précédent.
Une macro SAS documentée est disponible auprès des auteurs, autorisant le paramétrage des mécanismes de non-réponse (variables en entrée, nombres d'axes factoriels et nombre de classes à retenir), puis les traitements de correction de la non-réponse (nombre de groupes homogènes) et de calage (variables et marges à considérer), puis le calcul des biais et variances empiriques.

5.3 Mesure des performances

Pour chacun des cinq mécanismes de non-réponse, la sélection de l'échantillon est répétée 1000 fois. Pour chaque échantillon et chaque variable d'intérêt, on calcule les différents estimateurs de la moyenne correspondant à chacune des méthodes d'ajustement présentées en section 4. Nous avons donc 5000 échantillons d'environ 4319 individus correspondant à cinq mécanismes de non-réponse avec un taux moyen de non-réponse de 50%.
Pour chaque mécanisme, chaque méthode et chaque variable d'intérêt, l'efficacité est mesurée par le biais et la variance empiriques de l'estimateur du paramètre, la valeur de comparaison considérée étant celle de l'échantillon initial de répondants pris comme population statistique.

Les variances et biais empiriques pour les moyennes estimées $\hat{\bar{y}}$ de différentes variables d'intérêt de l'enquête sont calculés à partir des échantillons simulés, par comparaison aux valeurs $\bar{y}$ de l'échantillon initial considéré comme la population statistique. Pour chacun des b=1, …, 1000 échantillons, le total estimé $\hat{\bar{y}}^b$ est calculé ; le biais relatif BR_{Emp}, la variance Var_{Emp} et le MSE_{Emp} empiriques sont donnés par

$$BR_{Emp}(\hat{\bar{y}}) = \frac{1}{1000}\sum_{b=1}^{1000}\frac{\hat{\bar{y}}^b - \bar{y}}{\bar{y}},$$

$$Var_{Emp}(\hat{\bar{y}}) = \frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{\bar{y}}^b - \frac{1}{1000}\sum_{c=1}^{1000}\hat{\bar{y}}^c\right)^2,$$

$$MSE_{Emp}(\hat{\bar{y}}) = \left[\frac{1}{1000}\sum_{b=1}^{1000}\left[\left(\hat{\bar{y}}^b - \bar{y}\right)\right]^2\right] + Var_{Emp}(\hat{\bar{y}}).$$

5.4 Résultats

Liens entre non-réponse, variables sociodémographiques, paradonnées et variables d'intérêt
Par construction, dans MAR1, la corrélation linéaire entre non-réponse et variables sociodémographiques est de l'ordre de 0.10 ; dans MAR2, la corrélation entre non-réponse et paradonnées P1 est également d'environ 0.10 ; dans NMAR, la corrélation entre non-réponse et variables d'intérêt est également de 0.10. En revanche, nous n'avons pas calculé les corrélations entre les autres dimensions prises deux à deux dans ces mécanismes, pas plus que dans MAR3, qui est un

mixte de MAR1 et MAR2. Il est évident toutefois que les ordres de grandeur des corrélations entre non-réponse, variables sociodémographiques, paradonnées et variables d'intérêt sont plutôt inférieurs à 0.1 dans tous les mécanismes.

Performance des corrections de la non-réponse totale
Pour synthétiser les résultats, nous avons calculé les BR et les MSE pour toutes les variables d'intérêt pour : chaque mécanisme, chacune des 5 méthodes de correction et les calages qui s'ensuivent, plus deux méthodes témoins : le calage direct (méthode 0), et l'absence de traitement (pas de calage ni de correction de la non-réponse, méthode 00), soit 12 méthodes.

Nous avons alors calculé la moyenne des BR et des MSE sur l'ensemble des variables d'intérêt par mécanisme de non-réponse et chacune des 12 méthodes. Nous avons repéré la meilleure méthode sur cette somme, par mécanisme de non-réponse et calculé le gain relatif par rapport à l'absence de traitement (pas de poids) et par rapport au calage direct (méthode témoin). Les résultats sont présentés dans le tableau 3.
Ils montrent que du point de vue du MSE, le calage direct (méthode témoin) n'est jamais la meilleure option ; ils montrent aussi que l'usage de méthodes alternatives à la régression logistique pour la constitution de GRH n'est pas non plus optimale. En revanche, le choix du jeu de paradonnées a une certaine importance : le jeu P1 est optimal pour les mécanismes générés à partir de lui (MAR2 et MAR3) sauf le MAR1 et NMAR, où c'est le jeu P2 qui se montre légèrement supérieur par rapport au calage direct.
Enfin, le calage est toujours supérieur à l'absence de calage sauf pour le mécanisme MAR2, pour lequel une correction par régression logistique sur le jeu P1 sans calage ultérieur (Log_P1) est préférable. Ce résultat n'est pas très surprenant puisque MAR2 repose sur le jeu P1. Pour le mécanisme MAR3, qui est une sorte de moyenne entre MAR1 et MAR2, c'est encore sans surprise le jeu de paradonnéesP1 qui est optimal. Pour MCAR, c'est une utilisation d'une ACP sur le jeu P2 qui est optimale, mais l'écart avec le calage direct est quasi-nul et l'écart avec l'absence de traitement est également très faible : ces résultats sont sans doute dus au hasard.
Pour MAR1, le gain par rapport au calage direct obtenu par Calage_log_P2 est de 3,8%, tandis qu'il s'élève à 1.1% pour NMAR.
Si l'on compare ces résultats à ceux obtenus pour les biais (résultats non présentés), on observe que les classements et les gains sont identiques, ce qui souligne la faiblesse de la variance dans des échantillons de 4300 individus. Les gainsde MSEobtenus par toutes les méthodes sont donc des gains sur la dimension biais.

Tableau 3 : Somme des MSE pour toutes les variables d'intérêt et par mécanisme.

	calage_ acp P1	calage_ acP_P2	calage_ _ahq_P3	calage_iog _P1	calage_log _P2	acP P1	ACP_P2	ahq_P3	Log_P1	log_P2	Sans poids	calage_sec	**Meilleur** (2ème)	Gain relatif/ méthode 00	Gain relatif/ méthode 0
Ensemble des mécanismes	3.83	3.88	3.75	3.63	3.64	4.32	4.56	4.42	4.26	4.26	4.54	3.83	Calage_log_P1 (Calage_log_P2)	20.0% (19.9%)	5.3% (5.1%)
MAR1	0.41	0.40	0.40	0.40	0.39	0.55	0.56	0.55	0.55	0.54	0.56	0.41	Calage_log_P2 (Calage_log_P1)	30.1% (29.0%)	3.8% (2.3%)
MAR2	0.18	0.22	0.14	0.10	0.10	0.10	0.24	0.15	0.10	0.10	0.21	0.18	Log_P1 (Calage_log_P1)	52.7% (52.4%)	45.9% (45.6%)
MAR3	0.53	0.54	0.51	0.44	0.46	0.70	0.78	0.74	0.64	0.66	0.78	0.53	Calage_log_P1 (Calage_log_P2)	43.2% (41.1%)	16.1% (13.0%)
MCAR	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	Calage_acp_P2	5.1%	0.0%
NMAR	2.69	2.68	2.67	2.66	2.66	2.95	2.95	2.95	2.94	2.93	2.97	2.69	Calage_log_P2	10.3%	1.1%

Légende : calage_acp_Pi : calage à partir de GRH par ACP sur le jeu de paradonnées i ; calage_log_Pi : calage sur logistique à partir du jeu Pi ; calage_ahq_P3 : calage à partir de GRH par AHQ sur le jeu de paradonnées 3 ; etc. Méthode 00 : pas de redressement ; méthode 0 : calage direct sans correction de la non-réponse totale préalable.

6. Conclusion / discussion

Notre analyse montre que, dans le cadre de notre simulation, le recours aux paradonnées pour une étape de correction de la non-réponse préalable à un calage est toujours bénéfique en termes d'erreur quadratique moyenne (alors que la correction de la non-réponse totale seule est moins efficace que lorsqu'elle est couplée à un calage). Ce gain est maximal lorsque la non-réponse est liée aux paradonnées, et faible sinon. Notre analyse montre aussi que, contrairement à ce que nous pensions, les corrections fondées sur des classifications sur la base d'ACP ou d'AHQ sont moins efficaces que celles opérées par modélisation logistique classique. En revanche, le choix des paradonnées a une certaine importance : le jeu P2 qui brosse un bilan anhistorique des appels par plage horaire se montre parfois supérieur au jeu P1 qui en est une simplification sans plage horaire, bien que l'écart soit faible.

Dans leur analyse de simulation, (Little and Vartivarian 2005) montrent que l'usage de données auxiliaires pour redresser de la non-réponse totale réduit le biais si leur corrélation avec la propension à participer ainsi que les variables d'intérêt est supérieur à 0.5, mesure effectuée par le coefficient de corrélation de Pearson. Dans une enquête réelle, le lien entre les paradonnéeset la participation est fort, mais celui avec les variables d'intérêt est largement plus faible. Dans le cas de nos données, le lien entre nos paradonnées et les variables d'intérêt est faible, de même qu'avec la participation, sauf dans le cas des mécanismes de non-réponse mobilisant les paradonnées (MAR2 et MAR3). C'est aussi très souvent le cas dans les enquêtes multithématiques. Cette situation de faible lien entre les paradonnées et les variables d'intérêt est courante(Maitland, Cordero et al. 2009), mais les paradonnées s'avèrent néanmoins utiles au redressement dans d'autres contextes, comme l'enquête sociale européenne (Blom 2009).Pour autant, dans des cas où la participation est très élevée, leur utilisation peut être contre-productive car conduisant à des redressements instables et augmentant la variance (Wagner, Valliant et al. 2013).
Les mécanismes MAR1 et NMAR lient la non-réponse aux variables sociodémograhiques utilisées pour le calage (MAR1) et aux variables d'intérêt (NMAR) : le lien entre paradonnées et réponse y est par construction très faible comme le montre la matrice de corrélation calculée sur la population totale. Ce point est une faiblesse : dans la réalité, les paradonnées sont très nettement liées à la non-réponse, ce qui augmente leur intérêt pour le redressement. Toutefois, notre analyse montre que même dans ces cas défavorables, il y a intérêt à utiliser les paradonnées pour la correction de la non-réponse totale préalablement au calage. L'intérêt des paradonnées dans les cas plus favorables s'en trouve donc renforcé.
Dans le mécanisme MAR2, en revanche, le lien entre paradonnées et réponse est très fort, par construction, mais le lien entre les paradonnées et les variables d'intérêt est toujours très faible. Dans ce cas, la correction de la non-réponse est évidemment très profitable, mais peu instructive, car le cas est irréaliste. Le mécanisme MAR3 qui lie la non-réponse aux variables sociodémographiques utilisées pour le calage et aux paradonnées est plus réaliste. Les études montrent en effet à la fois que les non-répondants réels ont des paradonnées un peu particulières mais aussi que les répondants qui leurs sont proches du point de vue des paradonnées (grand nombre d'appels, refus préalables etc.) ont également des caractéristiques sociodémographiques particulières. Or les effets sur les variables d'intérêt sont irréductibles aux caractéristiques sociodémographiques (Legleye, Charrance et al. 2013; Legleye, Charrance et al. 2014). Nous confirmons ce point : alors que le mécanisme MAR3 est fondé pour partie sur les variables sociodémographiques utilisées dans le calage, utiliser les paradonnées dans la correction de la non-réponse avant calage est très bénéfique en termes d'erreur quadratique moyenne.

Nos résultats présentent aussi différentes limites.
D'abord, ils sont obtenus par simulation à partir de l'échantillon des répondants : or le taux de participation dans Fecond est proche de 45%. Par conséquent, nous ne pouvons ignorer que les non-répondants sont nombreux et que nos données sont déjà filtrées : la simulation de la non-réponse y est relativement artificielle et pourrait ne pas reproduire la réalité. Néanmoins, procéder ainsi, sur un échantillon de taille importante, permet d'avoir une large variabilité des réponses, variables d'intérêt et paradonnées dont le lien est réel, complexe et non pas artificiel.

Ensuite, nous n'avons tenu compte que d'une fraction de l'information contenue dans les paradonnées. Seule une nomenclature agrégée des issues a été utilisée : 4 issues ont été retenues alors que les données brutes en livrent près de 50. Il reste possible, bien qu'improbable, qu'une autre combinaison de paradonnées se montre plus efficace que celles que nous avons choisies ici.
Troisièmement, nous n'avons pas calculé les coefficients de corrélation entre la participation, les variables d'intérêt et les paradonnées dans nos échantillons. De même, nous n'avons pas pu spécifier, dans nos mécanismes de génération de non-réponse, la corrélation entre ces variables.

Développements futurs
Une prochaine version de la macro est en cours, qui intègrera ces fonctionnalités. Il sera alors possible de proposer des mécanismes à façon, contrôlant le taux de non-réponse, ainsi que, partiellement (dès lors que nos variables ne sont pas simulées mais que seule la non-réponse l'est), les corrélations entre paradonnées, variables sociodémographiques, variables d'intérêt et participation. Elle fournira un outil pratique pour estimer l'intérêt d'utiliser les paradonnées dans une enquête téléphonique.

Bibliographie

Bajos, N., A. Bohet, et al. (2012). "La contraception en France : nouveau contexte, nouvelles pratiques ? [Contraception in France: new context, new practices?]." Population et sociétés(492): 1-4.

Bajos, N. and M. Bozon, Eds. (2009). La sexualité en France, pratiques, genre et santé [Sexuality in France, practices, gender and health]. Paris, La découverte.

Bajos, N., A. Spira, et al. (1992). "Analysis of sexual behavior in France (ACSF). A comparison between two modes of investigation: telephone survey and face to face survey." **6**(315-323).

Beck, F., S. Legleye, et al. (2005). "Aux abonnés absents : liste rouge et téléphone portable dans les enquêtes en population générale sur les drogues." Bulletin de méthodologie sociologique**n°86**(Avril 2005): 5-29.

Bigot, R. and P. Croutte (2011). La diffusion des technologies de l'information et de la communication dans la société française. Paris, CREDOC.

Blom, A. G. (2009) "Nonresponse Bias Adjustments: What Can Process Data Contribute?" Institute for Social and Economic Research.

Brick, J. M., I. F. Cervantes, et al. (2011). "Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double." Techniques d'enquête**vol. 37**(n°1): 1-16.

Deville, J. and G. Saporta (1979). Analyse harmonique qualitative. Data analysis and informatics. E. Didays. Amsterdam, North-Holland**:** 375-389.

Durrant, G. B. and F. Steele (2009). "Multilevel Modelling Of Refusal And Noncontact In Household Surveys: Evidence From Six Uk Government Surveys." Journal Of The Royal Statistical Society, Series A**172**(2): 361-381.

Gautier, A., F. Beck, et al. (2006). Téléphones portables exclusifs : résultats d'une méthode de génération partielle de numeros. Methodes d'enquêtes et sondages - Pratiques européenne et Nord-Américaine, Québec. P. Lavallée and L. Rivest. Québec, Dunod**:** 60-64.

Groves, R. M. (2006). "Nonresponse Rates And Nonresponse Bias In Household Surveys." Public Opinion Quarterly**70**(5): 646-675.

Haziza, D. and J.-F. Beaumont (2007). "On the Construction of Imputation Classes in Surveys." International statistical review**75**(1): 25-43.

Lan, R. L. (2009). "Enquêtes ménages : vers la fin de la baisse des taux de réponse ?" Courrier des statistiques**n°128**(septembre-décembre 2009).

Legleye, S., G. Charrance, et al. (2014). How can we improve participation in telephone surveys? The FECOND survey experience. Proceedings of Statistics Canada Symposium 2013.

Legleye, S., G. Charrance, et al. (2013). "Improving survey participation: cost effectiveness of call-backs to refusals and increased call attempts in a national telephone survey in France." Public Opinion Quarterly**77**(3): 666-695.

Little, R. J. A. and S. Vartivarian (2005). "Does weighting for nonresponse increase the variance of survey means?" Survey methodology**31**(2): 161–168.

Maitland, A., C. C. Cordero, et al. (2009). An exploration into the use of paradata for nonresponse adjustment in a health survey. JSM proceedings. Alexandria, VA, American Statistical Association**:** 370-378.
Sarndal, C.-E. and S. Lundstrom (2005). Estimation in surveys with nonresponse. Chichester, UK, John Wiley.
Schouten, B., F. Cobben, et al. (2009). "Indicators for the representativeness of survey response." Survey methodology**n°35**(Juin 2009): pp. 101-113.
Wagner, J., R. Valliant, et al. (2013). Journal of survey statistics and methodology**2**(4): 410-430.