DÉCRIRE LES ABORDS ET L'ÉTAT DES LOGEMENTS POUR REDRESSER UNE ENQUÊTE EN FACE-À-FACE. L'EXEMPLE DE L'ENQUÊTE SANTÉ ESPS 2014

Stéphane Legleye (*), Stéphanie Guillaume (**) et Paul Dourgnon (**)

(*) Insee, Division Recueil et traitement de l'information
(**) Institut de recherche et de documentation en économie de la santé (Irdes)

stephane.legleye@insee.fr

Mots-clés : Paradonnées, redressement, face-à-face, SNIIRAM, ESPS

Résumé

Le redressement d'une enquête peut bénéficier d'une première étape de correction de la non-réponse en amont d'un calage. Dans cette première étape, les variables auxiliaires de la base de sondage, mais aussi les paradonnées (c.-à-d. les données générées par la collecte, comme l'historique des tentatives de contact) sont des variables attrayantes parce qu'elles sont enregistrées pour les répondants et les non-répondants et qu'elles peuvent se rapporter à la probabilité de réponse et aux variables cibles. Nous testons la pertinence de quatre variables de paradonnées dans le sous-échantillon de l'enquête de 2014 interrogé en face-à-face Santé, les soins de santé et l'assurance (ESPS) : le nombre de tentatives de visite, le type de logement, son apparence et la présence de personnes qui découragent son accès. La base de sondage est par ailleurs riche de variables auxiliaires, dont les indicateurs de consommation de santé (dépenses en ambulatoire et chez des spécialistes, nombre de consultations etc.). Nous comparons le calage direct classique d'ESPS à trois procédures en deux étapes, mobilisant chacune uniquement les variables auxiliaires, les paradonnées ou une combinaison des deux.

Les résultats montrent que le calage direct assure une très bonne estimation des totaux de dépenses de santé, et ne modifie que très marginalement les estimations de l'état de santé déclarée au sein de l'enquête. Les limites et les généralisations possibles sont discutées.

Abstract

Auxiliary variables and survey paradata (i.e. the data generated by the fieldwork itself, such as the history of contact attempts) are attractive variables for non-response survey weighting, when they relate to the target variables because they are recorded for the respondents and the non-respondents and relate to the response probability as well. We test the relevance of four paradata variables in the face-to-face subsample of the 2014 health, health care and insurance survey (ESPS): number of visit attempts, the type of housing, its appearance, and the presence of individuals discouraging its access. True health consumption indicators are available in the sampling frame such as the global health expenditure, the number of visits to a general (resp. specialist) practitioner. We compare the classic direct calibration of ESPS to three two-step procedures, each using only auxiliary variables, paradata or a combination of both However, preliminary results show that the estimates of these health indicators do not vary much when paradata are used in the weighting process. Other indicators (e.g. distributions, extreme values) will be analysed. Conclusion for ESPS and other face-to-face health survey are proposed.

1. Introduction

Le redressement des enquêtes ménages peut se faire en une ou deux étapes, le redressement en deux étapes étant généralement plus efficace [1]. Le calage, réalisé lors de la seconde étape, mobilise des variables socio-démographiques renseignées par les répondants, dont les totaux sont connus dans la population cible. L'éventuelle première étape consiste en une correction de la non-réponse totale (CNRT) reposant sur une modélisation de la réponse à l'enquête. Les variables mobilisables doivent posséder quatre caractéristiques [2] :

- Etre renseignées pour les répondants et les non-répondants, sans valeurs manquantes
- Etre liées à la réponse à l'enquête
- Etre liées avec les variables Y de l'enquête
- Etre renseignées sans erreur de mesure

Les variables auxiliaires de la base de sondage sont évidemment des variables de choix, mais les paradonnées le sont aussi. Ce sont les variables décrivant le processus de collecte lui-même. Les efforts entrepris pour contacter les répondants sont parmi les plus connues [3,4]. (Il existe un autre type de paradonnées, disponibles uniquement sur les répondants (temps de passation, débit de la voix, hésitations, contexte de passation, présence d'un tiers etc.). Ces dernières ne sont utiles que pour l'étude fine de la qualité des données et l'apurement de l'échantillon de répondants et nous ne les considérerons pas ici.) Les paradonnées sont renseignées pour les répondants comme les non-répondants et souvent liées aux variables cibles, faisant craindre l'existence de biais importants liés à leur ignorance dans le redressement en une étape [5-11].

Elles partagent avec les données de la base de sondage éventuelle le fait d'être disponibles pour les répondants et les non-répondants ; elles sont également très liées à la réponse à l'enquête, par définition et elles sont généralement collectées sans erreur. Lorsqu'elles sont liées aux variables cibles, alors elles remplissent les quatre conditions requises pour être utilisables dans un redressement, donc ici une CNRT. Précisons que les paradonnées existent même dans les enquêtes sans base de sondage, comme par exemple les enquêtes téléphoniques à génération aléatoire de numéros, pour lesquelles seul un calage direct est classiquement opéré pour le redressement. Ceci en fait un sujet d'étude important dans ce cadre.

Notre objectif est de voir si le recours à une CNRT sur paradonnées et sur variables auxiliaires peut se révéler utile dans le cadre d'une enquête aléatoire en face-à-face fondée sur une base de sondage.

Pour ce faire, nous utiliserons le volet face-à-face de l'exercice 2014 de l'enquête Santé et protection sociale, de l'Irdes (Institut de recherche et de documentation en économie de la santé). Cette enquête a la particularité d'utiliser la base de l'assurance maladie comme base de sondage, qui fournit de nombreux indicateurs médicoadministratifs décrivant les consommations de soin des personnes, par définition très liés aux comportements de santé étudiés, ce qui en fait une situation assez unique.

Nous comparerons, pour le volet face-à-face de l'enquête, le redressement en une étape par calage direct à des redressements en deux étapes : CNRT sur les variables auxiliaires puis calage et CNRT sur paradonnées puis calage.

2. Méthode

ESPS est un panel quadriennal tiré dans la base du régime général des bénéficiaires du régime général de l'assurance maladie, augmenté de quelques régimes spéciaux ; l'enquête existe depuis 1988 et nous utiliserons son dernier exercice de 2014. Il s'agit d'une enquête ménage : les ménages sont tirés au sort (sondage aléatoire simple) au du SNII-RAM. Par conséquent, les ménages nombreux ont plus de chances d'être tirés. Tous les éligibles du ménage y participent. La collecte est multimode (face-à-face, papier et téléphone suivant les parties du questionnaire) et occasionne deux visites en face-à-face (ou deux questionnaires téléphoniques) : nous nous concentrerons ici sur la première visite en face-à-face. Ce mode de collecte est attribué aux ménages dont le bénéficiaire tiré est âgé ou allocataire de la CMU (couverture maladie universelle) ou les ménages dont on n'a pas pu trouver un numéro de téléphone dans l'annuaire.

2.1. *Paradonnées*

En 2014, les enquêteurs devaient renseigner trois variables durant leur repérage du logement :

Logement : « S'agit-il d'un logement… 1/ Individuel (maison, pavillon) ; 2/ Collectif mais de petite taille (maison avec plusieurs sonnettes) ; 3/ Collectif de grande taille (immeuble, HLM, cité) ».

Allure : « Quelle est l'allure générale du bâtiment et de ses accès ? 1/ Bien tenu en apparence ; 2/ Quelque peu dégradé mais pas inaccueillant ; 3/ Très dégradé et inhospitalier (très sale, abimé, taggé). »

Hostile : « Indiquez s'il y a présence d'individus hostiles, menaçant, gênants ou surveillant l'entrée, décourageant l'accès (jeunes, bandes, chiens...). 1/ Oui ; 2/ Non. »
Le nombre de visites en face-à-face décrit les efforts entrepris lors de la collecte : *Nb_Visite_Faf*.

2.2. *Variables cibles*

Nous retiendrons les variables cibles suivantes :

- Recours aux soins (ces données sont renseignées dans la base SNII-RAM et non déclarées)
    - Nombre de consultations chez un spécialiste
    - Nombre de consultations chez un généraliste
    - Total des dépenses chez les généralistes
    - Total des dépenses chez les spécialistes

- Etat de de santé (déclaré par les répondants)
    - Se déclarer limité dans les gestes quotidiens
    - Se déclarer en mauvaise santé
    - Se déclarer en affection de longue durée
    - Se déclarer affecté d'une maladie chronique

2.3. *Variables de calage*

Le calage ordinaire de SPS est un calage simple sur 6 variables : Taille du ménage (1, 2, 3, 4+) ; sexe ; classe d'âge (14 modalités) ; régime d'assurance maladie (4 modalités) ; CMU (2 modalités) ; IdF (2 modalités).

2.4. *Variables auxiliaires*

Les adresses de la base de sondage sont géolocalisées à la commune (environ 50%) ou à l'IRIS, ce qui permet d'enrichir la base avec de très nombreux indicateurs contextuels tirés des enquêtes de l'Insee (plusieurs dizaines). Nous avons testé quelques-uns de ces indicateurs qui seront présentés plus loin. L'âge du bénéficiaire, son régime d'assurance maladie, son inscription à la CMU seront considérés comme des variables auxiliaires.

2.5. *Procédure*

Pour juger de la représentativité de l'échantillon de répondants, nous calculons un indicateurs R [12] à partir de la modélisation de l'appartenance à l'échantillon de répondants au sein de la base de sondage via une régression logistique relativement aux 5 variables de calage socio-démographiques. Le premier indicateur (R^1) prend en compte les effets propres des 5 variables SD, tandis que le second (R^2) prend en compte en plus toutes les interactions bivariées entre ces dernières, soit 15 effets.

Les variables utiles à un modèle de CNRT sont sélectionnées en fonction de leur lien avec la première composante principale des 4 variables médico-administratives et avec la réponse à l'enquête et avec la première composante des quatre variables d'état de santé déclaré.
Chaque modèle de CNRT utilisera la sélection de variables la plus prédictive de la composante principale médico-administrative, car ces variables sont renseignées pour les non-répondants et les répondants. La CNRT se fera par la méthode des quantiles (ici des déciles).
Les poids CNRT seront ensuite calés. L'appréciation de l'efficacité d'un redressement en deux temps se fera à partir de l'estimation des totaux des variables médico-administratives d'une part ; des proportions de variables d'état de santé déclaré d'autre part, en comparaison du redressement en une étape (calage direct).
L'association entre une variable binaire (réponse à l'enquête par exemple) et une autre est donnée par la distance standardisée *d* [13]. Ainsi, pour une variable continue,

Pour la modalité i d'une variable catégorielle :

La distance d permet ainsi d'évaluer simplement la similitude de distribution d'une variable (binaire ou continue) donnée dans deux échantillons (ou suivant les deux modalités d'une variable binaire), ce qu'on appelle souvent l'équilibrage. Inversement, un d>10 signera une association entre deux variables (une binaire et

une autre continue ou binaire). Pour fournir une mesure synthétique de l'équilibrage d'une variable catégorielle ayant plus de 2 modalités nous proposons également de calculer la moyenne des valeurs absolues des *d* de toutes les modalités.
Pour évaluer l'association d'une variable cible Y (ternaires) avec une autre, on a recours à une analyse de variance : on retiendra alors qu'une valeur F supérieure à 3 signe une association à prendre en compte. Cette valeur est arbitraire et son choix sera discuté plus loin. Le recours aux distances standardisées ou aux valeurs F permet de s'abstraire en partie des problèmes de puissance statistique liés à la taille des échantillons : lorsque celle-ci augmente, tous les écarts de distribution sont significatifs aux seuils habituels (0.05 ou 0.01) alors même que les différences peuvent être minimes.
Les modèles de CNRT utiliseront des combinaisons de variables maximisant le lien avec les données médico-administratives, à partir de la liste des variables sélectionnées sur ces critères. En effet, ces variables auxiliaires étant d'une part connues pour les répondants et les non-répondants, et d'autre part des reflets objectifs de l'état de santé, nous les privilégierons dans un premier temps relativement aux variables cibles d'état de santé déclaré. Pour produire une seule CNRT valable pour les 4 variables cibles retenues, nous proposons de retenir la première composante de l'analyse en composante principales (ACP) de ces quatre variables médico-administratives. C'est cette composante que l'on utilisera comme synthèse et que l'on cherchera à prédire à l'aide de combinaisons de variables auxiliaires et de paradonnées retenues suivant les critères exposés ci-dessus. Pour cela, les effets propres plus quelques interactions bivariées seront testées. La combinaison retenue sera ensuite utilisée pour prédire la réponse à l'enquête. L'idée est que la sélection initiale de variables, liées à la réponse à l'enquête, devrait assurer une bonne prédiction de cette réponse quelle qu'en soit la combinaison : autant alors tenter de maximiser le lien avec les variables cibles. A titre de test de robustesse, nous procèderons aux mêmes analyses mais en utilisant les variables d'état de santé déclaré en lieu et place des données médico-administratives dans l'ACP. Dans un souci de description de l'échantillon, nous présenterons également les secondes composantes issues des ACP des variables déclarées et médico-administratives.

3. Résultats

3.1. *Description de l'échantillon*

La distribution des variables de santé déclarée et des paradonnées est présentée Tableau 1. Sur la base initiale sélectionnée pour le face-à-face, 705 observations sont éliminées en raison du non renseignement des variables médico-administratives. Ces valeurs manquantes sont plus fréquentes parmi les non-répondants à l'enquête que parmi les répondants (7,9% 5,1%, $p<0.001$) et aussi parmi les ménages en CMU, ou dont on n'a pas retrouvé le téléphone ou qui appartiennent à un ménage de plus de 6 bénéficiaires (8% vs 2% pour les ménages dont la personne sélectionnée a plus de 65 ans) ; ces valeurs manquantes sont également plus fréquentes parmi les logements aux abords à l'allure un peu dégradée (8% vs 6% pour ceux jugés accueillants et bien entretenus), en Île-de-France (8% vs 6%) ou en immeubles collectifs qu'en pavillon (7% vs 6%). En revanche, parmi les répondants, la proportion de valeurs manquantes est plus faible parmi les gens se déclarant en bonne santé, sans limitation ni ALD etc. Ce biais sera discuté plus tard.
On notera que plus de la moitié des adresses sont enquêtées en face-à-face parce qu'aucun numéro de téléphone n'avait été retrouvé : les autres motifs d'orientation vers le face-à-face sont l'âge, la taille de la grappe de bénéficiaires de l'assurance maladie au sein du ménage et la perception de la CMU. Cette variable est donc très importante pour notre propos.
On peut noter que l'indicateur R de représentativité est très élevé : R^1=0,98 si l'on retient les 5 variables de calage sans interaction et R^2=0,84 si l'on considère toutes les interactions bivariées entre elles. L'enquête bénéficie donc d'un échantillon de répondants très représentatif du point de vue des variables de calage.

3.2. *Etat de santé déclaré et consommations de soins*

Les ACP des deux types de variables cibles sont représentées Tableau 2. La composante 1 des données médico-administratives (PrinY_1) est le cumul de visites et de dépenses médicales : elle cumule 55% de la variance. La seconde oppose les consultations/dépenses auprès de spécialistes à celles auprès de généralistes (omnipraticiens). La composante 1 des indicateurs d'état de santé déclaré (Prinsante_1) est déterminée par le cumul des variables d'états de santé (la variable Limite est en effet codé dans le sens contraire des autres) : elle cumule 62% de la variance. La composante 2 opposant les personnes se percevant en bon état de santé et sans limitation fonctionnelle à celles déclarant une ALD. La corrélation linéaire entre les composantes PrinY_1&2 et Prinsante_1&2 est modeste : seule la première composante décrivant l'état de santé déclaré est corrélée avec

les composantes médico-administratives 1 et 2. Retenir la composante médico-administrative 1 pour sélectionner les variables pour les CNRT semble donc pertinent : on assure une bonne liaison avec l'état de santé déclaré (du moins la première composante).
Les associations entre les variables auxiliaires et les paradonnées et les composantes principales des variables de dépenses de santé et d'état de santé d'un côté et la participation à l'enquête de l'autre, sont représentées Tableau 3. Pour les variables continues, les associations avec les composantes principales PrinY_1 et Prinsante_1 sont très faibles et les coefficients de corrélation linéaire sont très rarement supérieurs à 0.03 en valeur absolue : sauf les relations avec l'âge (ou son carré) et le nombre de visites atteignent des coefficients supérieurs (0.27 pour l'âge et 0.07 pour les nombres de visites). Le nombre de visites apparaît en revanche assez fortement lié à l'état de santé déclaré des répondants (rho=0,17). Pour les paradonnées catégorielles, les liens sont très modestes avec les composantes principales mais plus avec la participation. On peut noter que la raison de l'interrogation en face-à-face n'est pas liée à la réponse mais est fortement liée aux composantes principales PrinY_1 et Prinsante_1.
Les modèles de sélection des variables retiennent, pour la composante principale de dépenses de santé sont détaillés Tableau 4. Ils expliquent très peu de la variance de la composante PrinY_1 : les R^2 ajustés ne dépassent pas 0,1.

3.3. *Modèles de CNRT*

Les modèles de CNRT utilisant ces combinaisons de variables pour prédire la participation sont également peu explicatifs de la réponse : l'aire sous la courbe ROC [14] vaut 0.56 (pseudo R^2=0,01) pour le modèle sur variables auxiliaires, 0,59 pour le modèle sur paradonnées (pseudo R^2=0,03) et 0,58 pour le modèle complet (pseudo R^2=0,03).

3.4. *Distributions des poids et estimations*

La distribution des poids obtenus à la suite des GRH sur déciles puis calage est présentée Tableau 5 : les dispersions sont très semblables à celle des poids simples calés. Cela s'explique par les faibles pouvoirs prédictifs entre les variables utilisées pour les CNRT et la réponse. Les résultats des estimations des totaux des 4 variables médico-administratives de consommation de soins et de dépenses médicales sont représentés Tableau 6. Les redressements en deux temps ne modifient guère l'estimation obtenue par calage direct sur le poids uniforme, même si l'usage du modèle de CNRT avec paradonnées et variables auxiliaires se révèle meilleur que tous les autres redressements dans trois cas sur 4.
Enfin, les estimations des variables d'état de santé déclaré sont présentées Tableau 7. Là encore, il n'y a aucun changement notable dans les estimations.

3.5. *Analyse de robustesse*

Plutôt que de tenter de maximiser le lien entre les données médico-administratives et les paradonnées et variables auxiliaires dans la CNRT, nous avons tenté de maximiser le lien avec les données de santé déclarées : nous reprenons la procédure précédente mais substituons la première composante principale Prinsante_1 à PrinY_1. Les R^2 expliquant Prinsante_1 sont plus élevés : R^2=0,22 pour le modèle avec les variables auxiliaires uniquement, R^2=0,03 pour le modèle avec uniquement les paradonnées et R^2=0,23 pour le modèle complet. En revanche, les pseudo R^2 des modèles de CNRT restent faibles (0,02, 0,01 et 0,04 respectivement). Les estimations des variables médico-administratives tout comme les proportions des variables d'état de santé déclaré sont pratiquement inchangées, ce qui conforte nos conclusions précédentes.

4. Discussion

4.1. *Synthèse*

Les tentatives de redressement en deux étapes dans l'échantillon interrogé en face-à-face de l'enquête Santé protection sociale de 2014 ne montrent aucune amélioration par rapport au calage direct classique : l'estimation des totaux des indicateurs médico-administratifs du nombre de consultations médicales, ainsi que de dépenses de santé (en ambulatoire de spécialistes) est quasiment inchangée, tout comme les estimations des indicateurs de santé déclarée étudiés.
L'échantillon interrogé en face-à-face est donc bien redressé par le calage direct. Celui-ci prend en effet en compte l'âge et le fait d'être bénéficiaire de la CMU et le régime d'assurance maladie, qui apparaissent liés aux indicateurs médico-administratifs de consommations de soins. Néanmoins, on notera que croiser la CMU, le régime assurantiel ou l'âge ne modifient guère les résultats. A l'inverse, aucune des paradonnées ni autres variables auxiliaires n'apparaît fortement lié à ces indicateurs ni aux variables d'état de santé déclaré. On peut donc raisonnablement considérer que les biais liés à la non-réponse dans l'enquête sont minimes au regard des variables cibles étudiées et qu'ils n'en affectent que faiblement la mesure.
Une des raisons de l'efficacité du calage est la représentativité de l'échantillon de répondants analysé relativement aux 5 variables de calage utilisées (les indicateurs R sont à peine modifiés si l'on prend en compte en plus les 4 variables médico-administratives décrivant la consommation et les dépenses de soins).

Parmi les limites, on pourra noter la présence de valeurs manquantes sur les variables médico-administratives mais aussi sur les variables auxiliaires : 871 observations sans géocalisation à la commune ont été éliminées en amont de notre analyse (sur les 14385 observations initiales), tandis que la géolocalisation des adresses n'a pas permis d'obtenir plus de 57% de données à l'IRIS, et nous avons dû imputer de nombreuses valeurs communales. Il est probable que des données auxiliaires plus fines, notamment à l'IRIS, auraient été plus performantes.
Par la suite, 2537 observations ont dû être éliminées sur les 13514 restantes car elles n'ont pas donné lieu à un enregistrement des paradonnées. Nous avons donc restreint l'analyse à un sous-échantillon sans doute non représentatif, mais sur lequel il était aisé de travailler. Toutefois, il est peu probable, bien que cela reste possible, que les associations observées et utilisées dans notre étude entre les variables auxiliaires et les paradonnées d'un côté et les variables cibles retenues et la réponse à l'enquête d'autre part soient très différentes sur les observations éliminées, mais nous ne pouvons garantir la généralisabilité de nos conclusions à l'ensemble de l'échantillon. Nos résultats confirment donc en partie (nonobstant les remarques limitatives sur la précision et les non-réponses affectant les variables auxiliaires et les paradonnées) la conclusion de Kreuter et alii, selon laquelle il est effectivement difficile de trouver des paradonnées ou variables auxiliaires satisfaisant aux 4 critères énoncés en introduction [15].

4.2. *Enseignements pour les autres enquêtes*

Il est impératif d'améliorer la qualité du recueil des paradonnées pour les étudier. Toutefois, les nombres de visites, qui sont les paradonnées les mieux renseignées, sont certes liés à la participation à l'enquête, mais faiblement aux variables de santé (médico-administratives ou déclarées) : leur prise en compte est donc pratiquement inutile. Leur corrélation négative avec les données de santé déclarées ou médico-administratives souligne qu'on interroge facilement les personnes un peu malades à leur domicile en raison de leur état de santé un peu dégradé : au contraire, les bien portants seront plus souvent absents ou injoignables. Toutefois, ce biais est très modeste au vu de nos analyses ; et il est peu probable qu'il soit modifié en intensité ou du moins en sens en prenant en compte toutes les observations éliminées en amont dans notre étude. A l'opposé du spectre, les personnes hospitalisées manquent nécessairement dans l'enquête mais elles ne figurent pratiquement jamais dans les populations cibles des enquêtes. Les paradonnées semblent donc davantage pouvoir servir à la documentation de la collecte qu'au redressement.
Le calage de SPS ne fait pas appel à certaines variables classiques comme la PCS ou le diplôme ; en revanche, il fait appel à la CMU et au régime d'assurance maladie, ce qui pourrait largement expliquer son efficacité pour reconstituer les totaux des consultations ou des dépenses de soins renseignés dans le Sniiram. Pour tester la robustesse de nos conclusions, on pourrait envisager un autre calage plus classique, ainsi que des analyses complémentaires portant sur d'autres indicateurs de santé déclarée, comme la consommation d'alcool et de tabac.

Tableau 1
Distribution des variables de santé déclarées

	Frequency	Percent
Limité dans les activités quotidiennes		
Oui, fortement	642	11.6
Oui, limité mais pas fortement	1264	22.8
Non, pas limité du tout	3629	65.6
Etat de santé général		
Très bon	1025	18.5
Bon	2273	41.1
Assez bon	1644	29.7
Mauvais	505	9.1
Très mauvais	88	1.6
Maladie chronique		
Oui	2510	45.4
Affection de longue durée (ALD)		
Oui	1594	28.8
Logement		
Individuel	6751	65.7
Collectif petite taille	1057	10.3
Collectif de grande taille	2464	24.0
Allure		
Bien tenu en apparence	8924	86.9
Quelque peu dégradé mais pas inaccueillant	1173	11.4
Très dégradé et inhospitalier (très sale, abîmé, taggé)	175	1.7
Hostile (présence d'individus décourageant l'accès)		
Oui	280	2.7
Non	9992	97.3
Raison du face-à-face		
65+ ans	1087	10.6
70+ ans	2472	24.1
CMU	1108	10.8
6 bénéficiaires et +	259	2.5
Pas de téléphone	5346	52.0
Réponse à l'enquête		
Oui	5535	53.9

Tableau 2 : ACP des variables cibles médico-administratives et des variables cibles déclarées et corrélations entre les premières composantes

N=5535	prinsante_1	prinsante_2
Etasante	0.52	-0.40
Limite	-0.51	0.47
Chronique	0.51	0.21
ald	0.45	0.76
Inertie	0,62	0,17
N=10272	PrinY_1	PrinY_2
Dépense en spécialistes	0.54	-0.49
Dépense en ambulatoire	0.48	0.36
Nombre de séances (Omnipraticien)	0.41	0.70
Nombre de séances (Spécialiste)	0.56	-0.36
Inertie	0,55	0,22
N=5535	PrinY_1	PrinY_2
Corrélations linéaires (Pearson)		
prinsante_1	0.38	0.22
prinsante_2	0.07	0.00

Tableau 3 : association entre variables auxiliaires et paradonnées et composantes principales des variables cibles

		PrinY_1 n=10272		Prinsante_1 n=5535		Réponse n=10272
		Rho	F	Rho	F	*d*
Numériques						
QCHOM1524	Décile de chômage 15-24 ans communal	0.02	2.7	0.03	**6.3**	0.4
QCHOM1564	Décile de chômage 15-64 ans communal	0.01	0.3	0.02	2.0	6.6
QICAPBEP	Décile à l'Iris CAP-BEP	0.03	**6.9**	-0.01	1.2	6.7
QCCAPBEP	Décile communal CAPBEP	-0.03	**7.0**	0.02	2.5	**24.3**
QCSUP	Décile communal dipl. supérieur	0.02	**5.4**	-0.06	**23.3**	**17.6**
QCS1	Décile agriculteurs exploitants	-0.03	**10.5**	-0.01	0.6	**11.0**
QCS2	Décile artisans	0.00	0.0	-0.03	**4.5**	4.3
QCS3	Décile cadres	0.02	**4.7**	**-0.06**	**21.3**	**15.4**
QCS4	Décile prof. Intermédiaires	0.00	0.2	-0.05	**13.0**	1.9
QCS5	Décile employés	0.02	**3.0**	0.02	2.7	**14.0**
QCS6	Décile ouvriers	-0.03	**9.8**	0.01	0.8	**13.7**
QCS7	Déciles retraités	0.03	**6.5**	**0.06**	**19.8**	**12.7**
QCS8	Déciles autres inactifs	0.00	0.1	0.02	2.3	**13.6**
QFMONO	Décile familles monparentales	0.00	0.0	0.02	2.2	**15.4**
QCDD	Décile emplois en CDD	-0.01	0.3	**0.04**	**7.7**	4.9
QCDI	Décile emplois en CDI	0.02	**3.8**	-0.03	**4.9**	5.6
age	Age quinquennal	**0.27**	**816.6**	**0.45**	**1366.7**	1.8
Age²	Age quinquennal²	**0.27**	**807.2**	**0.44**	**1363.4**	0.9
Nb_visites*		**-0.07**	**56.8**	**-0.17**	**169.0**	**11.0**
Nb_visites²*		**-0.07**	**51.7**	**0.15**	**130.0**	**11.4**
Catégorielles						
sexe	Sexe de la personne sélectionnée		**113.2**		**4.2**	0.7
Logement*	Type de logement		2.4		0.8	**16.7**
Allure*	Allure des abords		**8.3**		0.9	9.4
Hostile*	Présence de personnes hostiles		1.1		1.2	3.9
d75	Résidence à Paris		0.6		0.1	2.7
d93	Résidence dans le 93		0.2		1.8	**12.3**
d94	Résidence dans le 94		0.0		0.0	**10.9**
idf	Résidence en Île-de-France		0.9		0.5	**20.3**
Raison du face-à-face			**140.8**		**205.8**	3.7
Cmu	Personne sélectionnée à la CMU		**14.9**		3.5	0.4
Regime	Régime d'assurance maladie		0,8		**12,4**	3,4

En gras : rho>0.03, F>3 ou d>10, critères de sélection pour l'entrée dans les modèles de CNRT
* : paradonnées

Tableau 4 : variables retenues dans les modèles de CNRT

CNRT	nb_effects	Effects	R^2 ajusté
CNRT	nb_effects	Effects	0,10
		LOGEMENT sexe age age2*sexe	
		PQ_FAF_r*sexe QCCAPBEP*QCS1	0,01
Modèle COMP	7	QICAPBEP*QCS1	
Modèle PARA	3	LOGEMENT nbvf nbvf*ALLURE	0,10

Tableau 5 : distribution des poids calés

Variable	N	Min	P1	P5	P95	P99	Max	CV	Sum
p_un_cal	5535	1222.0	2758.4	3512.7	19883.5	26881.9	68038.5	64.0	48759018.0
pcnrt_aux_cal	5535	1202.2	2652.2	3423.6	19948.7	28369.8	71516.1	64.9	48759018.0
pcnrt_para_cal	5535	891.1	2285.4	2922.7	20690.4	32045.5	74058.5	72.6	48759018.0
pcnrt_comp_cal	5535	1152.4	2626.6	3409.7	19986.0	28609.7	60724.9	65.6	48759018.0

Tableau 6 : Totaux des variables de dépenses de santé après calage

	FREQ	Ratio dép. ambulatoire	Ratio dép. spécialistes	Ratio consult. omnipraticiens	Ratio consult. spécialistes
psniiram	10272	1.000	1.000	1.000	1.000
p_uniforme	5535	0.547	0.546	0.555	0.543
p_uniforme_cal	5535	0.894	0.981	0.954	0.992
P_cnrt_aux_cal	5535	0.895	0.979	0.955	0.994
pcnrt_para_cal	5535	0.894	0.980	0.967	0.996
pcnrt_comp_cal	5535	0.896	0.981	0.957	0.997

Tableau 7 : Estimations des variables d'état de santé déclaré

	Calage direct		CNRT_AUX_CAL			CNRT_PARA_CAL			CNRT_COMP_CAL		
	%	StdErr	%	StdErr	Deff	%	StdErr	Deff	%	StdErr	Deff
etasante_nrb	8.5	0.4	8.5	0.4	1.01	8.7	0.4	1.06	8.6	0.4	1.02
limite_nrb	28.1	0.7	28.1	0.7	1.00	28.1	0.7	1.04	28.2	0.7	1.01
ald_nr	22.1	0.6	22.0	0.6	1.00	22.1	0.6	1.04	22.1	0.6	1.01
chronique_nr	39.2	0.8	39.3	0.8	1.01	39.2	0.8	1.04	39.4	0.8	1.01

Deff=ratio des erreurs standards relativement au calage simple

Bibliographie *Calibri 13*

1 Haziza D, Lesage E: A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics 2016;32:129-145.

2 Little RJA, Vartivarian S: Does weighting for nonresponse increase the variance of survey means? Survey methodology 2005;31:161ï 168.

3 Couper M: Measuring survey quality in a CASIC environment; in association As (ed): Survey Research Methods Section of the American Statistical Association, American statistical association, 1998, pp 41-49.

4 Olson K: Paradata for Nonresponse Adjustment. The Annals of the American Academy of Political and Social Science 2013;645:142-170.

5 Beck F, Legleye S, Peretti-Watel P: Le recours au téléphone dans les enquêtes en population générale sur les drogues: Journées de Méthodologie Statistique,. Paris, Insee, 2002,

6 Legleye S, Charrance G, Razafindratsima N, Bohet A, Bajos N, Moreau C: Improving survey participation: cost effectiveness of call-backs to refusals and increased call attempts in a national telephone survey in France. Public Opinion Quarterly 2013;77:666-695.

7 Legleye S, Razakamanana N, Charrance G, Juillard H: L'utilisation des historiques d'appels pour redresser une enquête téléphonique : une étude par simulation à partir de l'enquête Fecond XIIème Journées de méthodologie statistique de l'INSEE. Paris, France, 2015,

8 Kreuter F (ed): Improving surveys with paradata, New York, John Wiley & Sons, 2013.

9 Maitland A, Cordero CC, Kreuter F: An exploration into the use of paradata for nonresponse adjustment in a health survey; JSM proceedings. Alexandria, VA, American Statistical Association, 2009, pp 370-378.

10 Legleye S, Razakamanana N, Charrance G, Juillard H: Is it worth using paradata to correct for total non-response in telephone survey? A simulation study based on real data: ESRA. Reykjavik, 2015,

11 Blom AG: Nonresponse Bias Adjustments: What Can Process Data Contribute? Institute for Social and Economic Research.

12 Bethlehem J, Cobben F, schouten B: Des indicateurs de la représentativité aux enquêtes. Techniques d'Enquêtes 2009:1-10.

13 Austin PC, Stuart EA: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine 2015;34:3661-3679.

14 Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

15 Kreuter F, Olson K, Wagner J, Yan T, Ezzati-Rice T, Casas-Cordero C, Lemay M: Using proxy measures and other correlates of survey outcomes to adjust for non-response: exemple from multiple surveys. Journal of the Royal Statistical Society Series A 2010;173:389-407.