# VARIABLES AUXILIAIRES ET PARADONNÉES POUR AMÉLIORER LE REDRESSEMENT D'UNE ENQUÊTE TELEPHONIQUE : LE CAS DE CAMME-TH

*Stéphane Legleye (*), Bénédicte Mordier, Amandine Nougaret, Marie Clerc (**), François Beck (***), Maxime Levesque (****)*

*(*) Insee, Division Recueil et traitement de l'information*
*(**) Insee, Département des ressources et conditions de vie des ménages*
*(***) Insee, UPCEM*
*(****) Insee, Département de la Conjoncture*

stephane.legleye@insee.fr

**Mots-clés** : Paradonnées, variables auxiliaires, redressement, téléphone, Camme

## Résumé

Le redressement des enquêtes peut se faire en une ou deux étapes, le dernier étant généralement plus efficace [1]. Le redressement en une étape est généralement un calage, mobilisant des variables socio-démographiques renseignées par les répondants et dont les totaux sont connus dans la population cible. Lorsqu'il est fait en deux étapes, la première consiste en une correction de la non-réponse totale (CNRT) reposant sur une modélisation de la réponse à l'enquête. Les variables auxiliaires de la base de sondage mais aussi les paradonnées (variables décrivant le processus de collecte et notamment les efforts entrepris pour contacter les personnes sélectionnées) y sont potentiellement intéressantes. Dans cette étude, nous évaluerons l'effet d'un redressement en deux étapes plutôt qu'une sur les estimations de plusieurs variables cibles d'une enquête téléphonique. Nous envisagerons trois CNRT mobilisant exclusivement ou en combinaison, variables auxiliaires et paradonnées. Les données utilisées sont celles de l'expérimentation menée en 2017 sur l'enquête mensuelle de conjoncture auprès des ménages (Camme). La sélection des variables pour les modèles de CNRT est faite en retenant celles qui apparaissent les plus liées aux dix variables cibles (parmi les répondants) et à la participation à l'enquête. Les résultats montrent que ni l'adjonction de variables auxiliaires (dont le revenu fiscal du foyer) ni celle de paradonnées ne modifie les estimations de soldes : aucun écart avec les estimations classiques obtenues par calage direct ne dépasse 1 point de pourcentage. Une analyse de robustesse est faite. Nous discutons la portée de ces résultats pour Camme mais aussi pour les enquêtes téléphoniques à génération aléatoire de numéros.

## Abstract

Survey weighting can be achieved in one or two stages. The two-stage method allows taking into account variables that are recorded in the sampling frame (usually sociodemographic and auxiliary variables) during the first stage, when these variables relate both to the target variables of the survey and the response probability. Paradata (i.e. the data generated by the fieldwork itself, such as the history of contact attempts) are also good candidates 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 study the case of the recent Camme telephone survey. The sample was drawn in the tax registry that gave access to some auxiliary variables as well as to telephone numbers. We compare four weighting processes: the classical one-stage calibration based on sociodemographics of the respondents only, opposed to three two stages methods with various variables in the first stage: 1/ auxiliary variables only; 2/ paradata only; 3/ paradata and auxiliary variables. Correlations with the target variables, influence on the estimates and weight statistics are compared. Conclusions concerning Camme and other random telephone survey are discussed.

## Article long

### 1. Introduction

Le redressement des enquêtes peut se faire en une ou deux étapes. Le principe est toujours de sélectionner des variables liées à la fois à la réponse à l'enquête ainsi qu'aux principales variables cibles de l'enquête, dans la mesure du possible. La technique en une étape repose généralement sur un calage, qui consiste à rendre les marges de l'échantillon identiques à celles de la population cible sur quelques variables, généralement des variables socio-démographiques lorsqu'il s'agit d'enquête ménage.
Le gros avantage de cette technique est qu'elle permet de n'utiliser que des totaux ou des pourcentages des variables socio-démographiques pour la population cible : aucune base de données n'est requise. Comme les variables socio-démographiques sont généralement liées, même faiblement, à toutes les variables que l'on peut collecter dans une enquête, procéder ainsi permet d'estimer sans biais toutes les statistiques d'intérêt de l'enquête [2]. De plus, cela permet de présenter un échantillon qui apparaît bien comme une photographie miniature de la population cible, ce qui accroît la confiance dans la qualité des résultats.
Il est toutefois possible d'adjoindre une première étape consistant à mobiliser d'autres variables pour prédire la réponse à l'enquête avant un calage, cette étape étant qualifiée de correction de la non-réponse totale (CNRT). Les variables mobilisables doivent posséder quatre caractéristiques [3] :

- 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 de la base de sondage sont évidemment une ressource de choix dans ce but. Elles peuvent être de type sociodémographique et donc partiellement redondantes avec celles du calage ultérieur, mais elles sont naturellement utiles dans une CNRT dès lors que leurs codages ou leurs combinaisons ne sont pas totalement redondantes avec celles du calage.
Mais un autre type de données peut aussi être mobilisé : les paradonnées. Ce sont les variables décrivant le processus de collecte : nombre de tentatives de contact, caractéristiques des enquêteurs, types de difficultés rencontrées sur le terrain, etc. Les efforts entrepris pour contacter les répondants sont parmi les plus connues [4, 5]. (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 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. 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.

Dans cette étude, nous nous concentrerons sur le cas de l'enquête Camme. Camme est une enquête téléphonique reposant sur une base de sondage tirée des fichiers fiscaux. Sa particularité est donc de disposer d'une base de sondage assez riche. Nous utiliserons plus particulièrement les données de l'expérimentation de l'utilisation des numéros de téléphone laissés par les contribuables à l'administration fiscale, disponibles pour la première fois en 2017. Camme est une enquête particulière en ce qu'elle n'utilise aucune variable de la base de sondage pour son redressement : ce dernier s'effectue par calage direct sur quelques variables socio-démographiques classiques.
Notre objectif est ultimement de voir si le recours aux variables auxiliaires de la base de sondage ainsi qu'éventuellement aux paradonnées permet d'améliorer le redressement de Camme. Nous allons donc étudier le lien entre ces variables et les variables cibles de l'enquête puis envisager trois CNRT : l'une

fondée uniquement sur les variables auxiliaires, une autre sur les paradonnées uniquement, et enfin une mixant les deux types de variables. Ces CNRT seront suivies du calage classique de Camme, et nous comparerons les estimations des variables cibles des redressements en deux étapes à celles obtenues via le redressement classique en une étape, soit le calage direct.
Cette étude sera aussi l'occasion de présenter brièvement le gain de couverture, de taux de réponse et de représentativité obtenus grâce au recours au numéro de téléphone contenus dans les fichiers fiscaux relativement à l'enquête Camme courante.

### 1.1. *Présentation de Camme*

L'enquête mensuelle de conjoncture auprès des ménages Camme répond à une demande de la Commission européenne. Elle fait partie du système européen harmonisé des enquêtes de conjoncture auquel contribuent maintenant les États membres de l'Union européenne, avec dix autres enquêtes de conjoncture que mène l'Insee auprès des entreprises. Créée en 1958, elle devient mensuelle à partir de janvier 1987. Au cœur de l'enquête, le module « conjoncture » s'articule autour d'une partie « opinion sur la conjoncture générale en France » et d'une partie « opinion sur la situation économique du ménage ». Depuis 2005, l'enquête est le support de plates-formes permettant de répondre rapidement à des besoins du système statistique national. Ces plates-formes abordent en une vingtaine de questions des thèmes aussi diversifiés que les pratiques environnementales, le bien-être subjectif ou les difficultés des ménages en matière de logement.

### 1.2. *Présentation de Camme TH*

De mai à septembre 2017, une expérimentation méthodologique (Camme TH) a été menée, qui visait à élargir le champ couvert par l'enquête, en utilisant les coordonnées téléphoniques présentes dans les fichiers de la taxe d'habitation, en complément des numéros de téléphone retrouvés dans l'annuaire. Ces deux sources d'information permettent actuellement de porter à plus de 70 % le champ couvert par l'enquête contre 35 % pour l'enquête Camme courante. L'expérimentation avait pour but d'étudier les difficultés liées à l'utilisation de ces numéros lors de la collecte et d'évaluer le potentiel impact de l'élargissement du champ de l'enquête sur la mesure actuelle du moral des ménages.

### 1.3. *Variables cibles*

Ces variables sont au nombre de 10, elles sont ternaires.
NVP = niveau de vie en France au cours des 12 derniers mois
NVF = niveau de vie en France au cours des 12 prochains mois
SFP = situation financière du foyer au cours des 12 derniers mois
SFF = situation financière du foyer au cours des 12 prochains mois
OE = opportunité d'épargner
PP = évolution des prix au cours des 12 derniers mois
PF = évolution des prix au cours des 12 prochains mois
CEA = situation financière du foyer
CEF = capacité à mettre de l'argent de côté au cours des 12 prochains mois
EF = évolution du nb de chômeurs au cours des 12 prochains mois

### 1.4. *Variables auxiliaires*

Les variables auxiliaires disponibles dans la base de sondage sont : le sexe et l'âge de la personne de référence (en tranches) ; le type de logement (maison/appartement ou autre), le statut d'occupation (propriétaire/locataire ou autre), la région de résidence (qui sera dichotomisée en Ile-de-France et reste de la France), la taille d'unité urbaine, le dépôt d'une adresse courriel (oui/non), le dépôt d'un numéro de téléphone fixe (oui/non), le dépôt d'un numéro de téléphone mobile.

### 1.5. *Paradonnées*

Il s'agit du nombre d'appels (1-20) et des nombres de refus, de non-contacts et de raccroches qui sont connus uniquement pour les 5 premiers appels, et de l'envoi d'une lettre de relance (oui/non) pour les impossibles à joindre et de l'envoi d'une lettre de relance pour les refusants (oui/non). D'autres variables sont également disponibles comme les dates et heures des appels. Mais comme les issues ne sont documentées que pour les 5 premiers, nous les avons délaissées.

## 2. Analyses statistiques

### 2.1. *Représentativité et distances standardisées*

Les taux de participation à l'enquête sont fournis dans le tableau 1. La représentativité des différents échantillons sélectionnés et des répondants est évaluée simplement par des distances standardisées à la base de sondage et des indicateurs R [6] (Tableau 2).

La différence standardisée *d* est définie suivant [7]. 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.

### 2.2. *Choix des variables pour la CNRT*

Les variables candidates à un modèle de CNRT doivent être associées à la fois à la réponse à l'enquête et aux variables cibles Y. Ces dernières étant au nombre de 10, nous avons retenu comme critères d'association qu'au moins 2 F devaient être supérieurs à 10 tandis que 2 F devaient être supérieurs à 3. Les variables candidates associées uniquement à la réponse ne sont pas retenues, alors que celles associées aux variables Y le sont systématiquement [3, 8]. La synthèse de ces associations est présentée dans le tableau 3.

### 2.3. *Les modèles de CNRT*

Le modèle de CNRT est estimé à l'aide d'une régression logistique modélisant la réponse à l'enquête à partir des variables sélectionnées plus haut. La CNRT proprement dite est effectuée par la méthode des groupes homogènes de réponse (GRH) définis par la méthode des quantiles à partir de la distribution de la probabilité prédite de réponse.
Le premier modèle de CNRT n'intègre que les variables auxiliaires. Il est effectué à partir de l'échantillon des répondants et non-répondants pondéré par le poids de sondage. Les variables candidates, et quelques interactions bivariées entre elles, sont testées suivant la méthode stepwise classique : le seuil d'entrée d'une variable candidate est fixé à 0.25, le seuil de sortie à 0.1 pour les tests de Wald. Le second modèle suit la même procédure de sélection automatique mais n'intègre en entrée que les paradonnées : toutes les interactions bivariées entre les candidates sont testées. Le troisième intègre les deux jeux de variables

auxiliaires et de paradonnées sélectionnés dans les deux premiers modèles sans procédure de sélection automatique, sans ajouter d'interaction.
Les estimations des pourcentages de modalités $Y_i$=1 pour chacune des 10 variables cibles $Y_i$ de Camme sont présentées tableau 5. Enfin, les statistiques de poids des différentes pondérations sont présentées tableau 6.

## 3. Résultats

### 3.1. *Taux de réponse*

Le taux de collecte varie selon l'échantillon (Tableau 1) : c'est dans les échantillons utilisant les numéros fournis par la TH qu'il est le plus élevé (échantillons 3 et 4), culminant auprès des ménages dont un numéro figurait dans la TH mais un autre (éventuellement le même) également dans l'annuaire (échantillon 3 : le taux de réponse de 77.6% est plus élevé que le 69.3% de l'échantillon 4 : p=0.011). L'apport des numéros TH s'avère donc important ; en particulier, appeler avec le numéro TH des ménages dont un numéro est aussi dans l'annuaire est la situation la plus favorable.

### 3.2. *Représentativité*

Le Tableau 2 présente la distribution de quelques variables présentes dans la base de sondage au sein des différents échantillons. Une approche de la représentativité (univariée) est fournie par les différences standardisées *d*, calculées pour toutes les modalités des variables (catégorielles) et synthétisées dans la moyenne des valeurs absolues calculées par variable et au global. La représentativité d'un échantillon sélectionné est toujours supérieure à celle de sa fraction de répondants et globalement, c'est bien l'échantillon complet de Camme TH qui apparaît offrir la meilleure couverture (*d* global moyen=3,6 vs 8,6 pour Camme courant ou Camme TH exclusivement). Tous les échantillons de répondants sous-représentent les jeunes, les faibles revenus etc. Ces biais sont classiques. Si on se limite aux répondants, c'est l'échantillon de répondants Camme TH exclusif qui apparaît le plus représentatif avec le *d* moyen global le plus faible (*d*=14,2 vs d=19,6 pour l'ensemble de l'expérimentation et 23,2 pour Camme courant). Ce résultat plutôt surprenant est confirmé par le calcul des indicateurs R : R=0,90 pour l'échantillon total, R=0,91 pour l'échantillon Camme courant et R=0,95 pour Camme TH exclusivement.

### 3.3. *Les modèles de CNRT*

Nous nous tournons maintenant vers la CNRT. La sélection des variables associées aux variables cibles montre que pour les variables auxiliaires : le sexe, l'âge de la personne de référence fiscale, son statut matrimonial, le type de logement et le statut d'occupation, le revenu fiscal (ou mieux : le décile), le dépôt d'un courriel, celui d'un numéro de téléphone fixe ou celui d'un mobile, l'échantillon d'appel et la résidence en Ile-de-France sont associées aux variables cibles (au moins 2 F>3). Pour les paradonnées, l'envoi d'une lettre de relance pour les injoignables ou d'une lettre à destination des refusants, ainsi que le nombre de raccroches sont associés aux variables Y suivant notre critère.
Pour l'association avec la réponse à l'enquête, les variables auxiliaires satisfont au critère de sélection, sauf : le dépôt d'un numéro de téléphone (fixe ou mobile) et le revenu fiscal du foyer (ou son décile). Pour ce qui est des paradonnées, elles satisfont toutes au critère du F>3. On note toutefois qu'aucune ne satisfait au critère d>10 (max=4,8), bien que tous les t-tests soient significatifs au seuil 0.001. Le tableau 3 synthétise ces résultats.

La liste détaillée des variables retenues au final à l'issue des modélisations est la suivante :
Modèle AUX : 9 effets
ageprcl ageprcl*ageprcl echant idf mcdvt1 occr r_foy_rev thtelfix thtelmob*echantth

Modèle PARA : 10 effets
echant nb_raccr nb_rates nb_rates*echant nb_rates*riajb nb_rates*rrefb riajb riajb*nb_raccr rrefb rrefb*nb_raccr

Modèle COMP : identique à Modèle AUX + Modèle PARA (18 effets)

Les pouvoirs prédictifs de ces modèles sont élevés : AUC=0.813 pour Modèle COMP, 0,785 pour Modèle PARA et 0.657 pour le Modèle AUX. Ces différences illustrent bien le fait que les variables auxiliaires sont moins reliées à la réponse que les paradonnées.
Les estimations calées des pourcentages de modalité 1 de chacune des variables Y de Camme sont présentées tableau 4. Les écarts sont minimes et tout à fait négligeables, pour toutes les variables et tous les redressements, lorsqu'on prend comme base le calage direct classique de Camme.

La structure des poids est correcte pour ces différents redressements, comme le montre le tableau 5.

### 3.4. *Analyse de robustesse*

Plutôt que de prendre comme critère de sélection des variables pour la CNRT les corrélations/associations bivariées entre les variables auxiliaires et paradonnées d'un côté et les variables cibles et la réponse de l'autre, nous pouvons nous éloigner de cette méthodologie classique afin d'orienter les choix vers les variables cibles. En effet, dès lors que les paradonnées et les variables auxiliaires s'avèrent fortement liées à la réponse (ce qui est le cas ici), toute combinaison le sera. Par conséquent, il est tentant de retenir la combinaison la plus liée aux variables cibles. Pour ce faire, nous avons synthétisé la structure des variables cibles au sein des répondants, à l'aide d'une ACP (pondérée par le poids de sondage). La structure est grossièrement bidimensionnelle en retenant le critère de Keyser (valeur propre supérieure à 1) : les trois premières valeurs propres valent respectivement 2,76, 1,32 et 0,97 (soit 27,6% d'inertie pour la première). Chaque CNRT utilisera la combinaison de variables qui explique au mieux la composante 1, considérée comme une variable continue : toutes les variables (auxiliaires ou paradonnées ou les deux) ainsi que leurs interactions bivariées sont soumises à sélection (sans hiérarchie). Les $R^2$ de ces modèles sont faibles : 0,07 pour les variables auxiliaires, 0,01 pour les paradonnées et 0,08 pour l'ensemble ; les corrélations des variables cibles avec les prédictions de la composante 1 sont comprises entre 0,00 et 0.28 : 2 corrélations seulement sont supérieures à 0.2 pour le modèle avec variables auxiliaires et celui avec auxiliaires et paradonnées (Tableau 6). En utilisant ces sélections de variables (ou les valeurs prédites de la première composante) pour opérer les CNRT via la méthode des GRH, les estimations obtenues post calage ne sont jamais différentes de plus d'un point de pourcentage de l'estimation initiale fournie par le calage direct (résultat disponible sur demande). Il semble donc difficile d'améliorer nos résultats initiaux.

### 3.5. *Une exploration rapide de Camme courant*

Si l'on réplique l'analyse présentée à l'enquête Camme courante, les résultats sont similaires : les estimations des variables de soldes sont inchangées ou presque par les différents redressements en deux étapes et l'usage et aucune pondération ne modifie très sensiblement les résultats (résultats disponibles sur demande).

## 4. Discussion

### 4.1. *Synthèse des résultats*

Le recours aux numéros de téléphone de la TH semble être très bénéfique pour le taux de participation et la représentativité de l'échantillon de répondants. Il semble utile de poursuivre les enquêtes Camme en suivant le protocole de l'expérimentation étudiée ici. Se restreindre aux seuls ménages ayant laissé un numéro TH pourrait même être envisagé.
Le redressement en deux étapes est généralement une bonne option pour les enquêtes. Dans le cas de Camme, toutefois, une telle méthode s'avère inutile. Quelles sont les raisons de ce résultat ?
D'abord on peut citer la très forte représentativité de l'échantillon de répondants du point de vue des variables de calage : toute tentative d'améliorer les choses sera donc limitée dans ses conséquences. Deuxièmement, on peut citer les corrélations faibles entre les variables auxiliaires et surtout les paradonnées et les variables cibles de Camme. C'est pour cette raison que nous avons choisi arbitrairement un seuil très bas pour retenir les variables auxiliaires et les paradonnées suivant leur lien

avec les variables cibles (F>3 et nb F>3=2). En termes de coefficient de corrélation, pour les paradonnées, le maximum est obtenu pour le nombre de refus/raccroche et la variable CEF (rho=0,05) ; pour les variables auxiliaires, les maxima sont rho=0,08 pour la taille d'unité urbaine et la variable OE, tandis que le décile de revenu du foyer fiscal apparaît plus important (rho=0,27 pour la variable CEA, 0,19 pour la variable CEF, 0,15 pour les variables PP et SFP). Autrement dit, seule l'adjonction du décile de revenu du foyer fiscal peut avoir un impact notable sur l'estimation des variables cibles si l'on retient de critère d'une corrélation supérieure à 0.5 énoncé dans [14]. Notre analyse initiale et notre analyse de robustesse montrent que le cumul de variables faiblement corrélées n'améliore en rien les choses : retenir des seuils plus bas encore pour sélectionner d'autres variables liées aux variables cibles ne sert à rien. Notre analyse confirme ainsi l'observation faite par Kreuter et alii : il est très difficile de trouver des variables auxiliaires ou des paradonnées satisfaisant aux 4 critères de Little et Vartivarian et notamment fortement corrélées aux variables cibles [9] : aucune ne présente un coefficient de corrélation supérieur à 0,5 tel que recommandé. Ces difficultés sont connues [10].
La troisième raison réside peut-être dans la nature des questions à l'origine des variables de Camme : il s'agit pour certaines de questions d'opinions que l'on pose ordinairement aux spécialistes de l'économie et autres experts médiatiques plutôt qu'aux citoyens ordinaires. Il est probable qu'une part importante des réponses soit fournie au hasard ou du moins sans trop de conviction. De fait, les taux de non-réponse sont relativement importants pour certaines questions (8% pour la prédiction du futur de l'économie française et de l'indice des prix). De plus, les échelles initiales sont de type Likert à 5 modalités, mais les variables cibles utilisées les regroupent en trois, avec les non réponses dans la modalité centrale. Au final, la nature et le traitement de ces questions sont très spécifiques, et pourraient expliquer la stabilité des estimations quels que soient les redressement utilisés (ou l'absence de redressement).
Enfin, il faut considérer l'étape de calage effectuée soit directement (redressement classique) soit à la suite de la CNRT. Bien que ce calage ne prenne pas en compte le revenu fiscal qui apparaît être la variable (parmi les auxiliaires et les paradonnées) la plus liée aux variables cibles, l'estimation du total des revenus fiscaux des ménages de la base de sondage à la suite du calage aboutit au score impressionnant de 99,2% de la somme totale déclarée par les contribuables de la base de sondage (alors que l'estimation en deux temps CNRT AUX –qui comprend pourtant le décile de revenu- et calage n'aboutit qu'à 95% du total). Autrement dit, ce redressement est très efficace malgré son apparence fruste. Le choix de variables fortement liées aux revenus (type de logement et statut d'occupation, groupe social et taille du ménage) explique probablement ce résultat.

### 4.2. *Enseignements pour les autres enquêtes*

Parce que l'expérimentation Camme TH est une enquête téléphonique tirée dans une base de sondage et disposant d'une bonne couverture du champ des ménages (70% des adresses), notre travail offre quelques pistes de réflexion pour les enquêtes téléphoniques aléatoires en général, y compris celles réalisées sans bases de sondage, comme le Baromètre santé de Santé publique France.
L'usage des paradonnées n'est d'aucun secours pour Camme dès lors que le calage est fait sur l'âge (en quatre modalités), le nombre de personnes dans le ménage (en quatre modalités), le groupe social (en six modalités), le type de logement (deux modalités) et le statut d'occupation (deux modalités), la taille d'Unité urbaine (en six modalités). Cela ne prouve toutefois pas que cela ne soit pas le cas pour toutes les enquêtes téléphoniques des variables cibles d'un autre type, comme des variables de comportement de santé (par exemple fumer du tabac, boire de l'alcool etc.) En revanche, l'estimation presque parfaite des revenus fiscaux des ménages obtenue avec ce calage suggère qu'en première approche, le calage direct utilisé est très fortement lié au revenu, lui-même lié à quantité d'autres variables cibles classiques dans les enquêtes (santé, comportements d'achats, d'épargne etc.) De fait, la corrélation linéaire entre le poids calé directement et le revenu fiscal du ménage vaut -0.18 et celle avec son décile vaut -0.34. Les variables du calage pourraient ainsi être utilisées pour tester le redressement classique des enquêtes type Baromètre santé : type de logement et statut d'occupation notamment, mais aussi groupe social.
En revanche, les paradonnées ne sont pas ou très peu liées au revenu fiscal : la corrélation maximale est obtenue pour le nombre de refus/raccroches et le décile de revenu (rho=0,08), tandis que celle du nombre d'appel est très faible et non significative (rho=-0,02, p=0,223). Ce qui souligne que les répondants difficiles à joindre ou à convaincre de participer ne sont pas très différents des autres répondants de ce

point de vue, bien que les répondants soient effectivement dotés d'un revenu plus élevé (38 300 € vs 32900 €, $p<0.001$). Cela est rassurant pour les enquêtes aléatoires téléphoniques : il ne semble pas utile de prendre en compte les paradonnées pour redresser l'échantillon de répondants si ce qui est visé est le niveau de vie approximé par le revenu fiscal.

**Tableau 1 : taux de collecte des différents échantillons Camme TH sur la période mai-août 2017**

| Echantillon | Non-répondants | Répondants | Total |
|---|---|---|---|
| 1 : Annuaire<br>Utilisé : Annuaire | 698 | 821 | 1519 |
| | 46,0% | 54,0% | 100,0% |
| 2 : Annuaire + TH<br>Utilisé : Annuaire | 1086 | 1740 | 2826 |
| | 38,4% | 61,6% | 100,0% |
| 3 : Annuaire + TH<br>Utilisé : TH | 49 | 170 | 219 |
| | 22,4% | 77,6% | 100,0% |
| 4: TH<br>Utilisé : TH | 479 | 1079 | 1558 |
| | 30,7% | 69,3% | 100,0% |
| Total | 2312 | 3810 | 6122 |
| | 38,8% | 62,3% | 100,0% |

**Lecture :** dans l'échantillon 1 des numéros retrouvés uniquement dans l'Annuaire, 54 % des ménages ont répondu aux premières interrogations des enquêtes Camme de mai à août 2017

Tableau 2 : représentativité des échantillons sélectionnés et des répondants par rapport à la base de sondage

| | Distribution (%) | | | | | | | Equilibrage (*d*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base de sondage | Ech 1 à 4 | | Ech 1 et 2 | | Ech 3 et 4 | | Ech 1 à 4 | | Ech 1 et 2 | | Ech 3 et 4 | |
| | | Ech | Rep | Ech | Rep | Ech | Rep | Ech | Rep | Ech | Rep | Ech | Rep |
| Sexe pers. Ref. | | | | | | | | 0.0 | **15.3** | 2.2 | **16.7** | 4.4 | **12.2** |
| Homme | 68.0 | 68.0 | 74.9 | 69.0 | 75.5 | 66.0 | 73.5 | 0.0 | **15.3** | 2.2 | **16.7** | -4.4 | **12.2** |
| Statut matrim. | | | | | | | | 6.4 | **13.8** | **10.2** | **20.2** | 7.1 | **10.3** |
| Célibataire | 35.9 | 29.4 | 20.8 | 25.3 | 17.0 | 37.8 | 29.2 | **-13.8** | **-33.9** | **-23.1** | **-43.7** | 3.9 | **-14.3** |
| Divorcé | 13.8 | 15.1 | 13.2 | 13.4 | 11.1 | 18.6 | 17.7 | 3.8 | -1.8 | -1.1 | -8.0 | **13.1** | **10.8** |
| Marié | 36.0 | 38.3 | 51.9 | 42.8 | 56.6 | 29.1 | 41.5 | 4.8 | **32.5** | **14.0** | **42.2** | **-14.7** | **11.3** |
| Autre | 4.0 | 3.9 | 3.9 | 4.1 | 3.6 | 3.7 | 4.7 | -0.6 | -0.7 | 0.1 | -2.5 | -2.0 | 3.0 |
| Veuf | 10.4 | 13.3 | 10.2 | 14.5 | 11.7 | 10.9 | 7.0 | 9.1 | -0.4 | **12.5** | 4.3 | 1.7 | **-12.0** |
| Type logement | | | | | | | | 2.1 | **29.9** | **10.2** | **37.9** | **14.1** | **13.1** |
| Maison | 52.0 | 53.1 | 66.5 | 57.1 | 70.2 | 45.0 | 58.5 | 2.1 | **29.9** | **10.2** | **37.9** | **-14.1** | **13.1** |
| Statut occupation | | | | | | | | 4.0 | **37.7** | **15.4** | **43.4** | **18.9** | **25.8** |
| Propriétaire | 54.5 | 56.5 | 72.4 | 62.1 | 74.8 | 45.2 | 67.0 | 4.0 | **37.7** | **15.4** | **43.4** | **-18.9** | **25.8** |
| Age pers. Ref. | | | | | | | | 8.7 | **14.9** | **10.7** | **17.8** | 7.2 | **12.9** |
| 18-24 | 3.7 | 1.7 | 0.4 | 1.3 | 0.3 | 2.4 | 0.7 | **-12.6** | **-23.3** | **-15.4** | **-24.6** | -7.6 | **-20.6** |
| 25-34 | 20.7 | 14.0 | 7.6 | 10.9 | 5.0 | 20.2 | 13.4 | **-17.8** | **-38.3** | **-27.1** | **-48.4** | -1.2 | **-19.5** |
| 35-44 | 14.3 | 17.8 | 14.5 | 16.4 | 12.5 | 20.6 | 18.9 | 9.5 | 0.5 | 5.8 | -5.3 | **16.6** | **12.4** |
| 45-54 | 16.4 | 17.3 | 19.7 | 17.3 | 19.0 | 17.3 | 21.3 | 2.3 | 8.6 | 2.3 | 6.8 | 2.3 | **12.6** |
| 55-64 | 15.4 | 15.8 | 19.5 | 16.8 | 20.2 | 13.8 | 18.1 | 1.1 | **10.8** | 3.8 | **12.5** | -4.7 | 7.1 |
| 65-74 | 14.0 | 12.1 | 20.5 | 13.2 | 22.1 | 9.9 | 17.0 | -5.5 | **17.3** | -2.2 | **21.2** | **-12.5** | 8.4 |
| 75-84 | 9.4 | 11.8 | 13.4 | 13.7 | 15.8 | 7.9 | 8.0 | 7.9 | **12.7** | **13.6** | **19.6** | -5.3 | -4.9 |
| 85+ | 6.2 | 9.6 | 4.4 | 10.4 | 5.2 | 8.0 | 2.6 | **12.8** | -8.0 | **15.4** | -4.2 | 7.1 | **-17.3** |
| Revenu fiscal | | | | | | | | 2.1 | 9.3 | 2.5 | 9.6 | 3.2 | 9.4 |
| d1 | 9.9 | 7.4 | 3.6 | 6.5 | 3.6 | 9.2 | 3.6 | -9.0 | **-25.4** | **-12.5** | **-25.4** | -2.3 | **-25.5** |
| d2 | 10.0 | 10.4 | 6.5 | 9.8 | 6.1 | 11.5 | 7.5 | 1.3 | **-12.5** | -0.6 | **-14.3** | 5.0 | -8.7 |
| d3 | 10.0 | 10.5 | 7.4 | 10.4 | 7.7 | 10.9 | 6.6 | 1.9 | -9.2 | 1.4 | -7.8 | 2.9 | **-12.3** |
| d4 | 10.0 | 10.7 | 9.0 | 10.6 | 9.2 | 10.8 | 8.6 | 2.2 | -3.4 | 2.0 | -2.6 | 2.5 | -5.0 |
| d5 | 10.0 | 10.1 | 9.8 | 10.4 | 9.2 | 9.7 | 11.0 | 0.6 | -0.6 | 1.4 | -2.5 | -1.1 | 3.5 |
| d6 | 10.0 | 10.2 | 10.8 | 10.2 | 10.8 | 10.2 | 10.7 | 0.8 | 2.6 | 0.7 | 2.7 | 0.8 | 2.4 |
| d7 | 10.0 | 9.7 | 12.1 | 11.0 | 12.4 | 7.1 | 11.4 | -1.0 | 6.6 | 3.1 | 7.5 | **-10.3** | 4.6 |
| d8 | 10.0 | 10.7 | 12.5 | 10.5 | 12.8 | 11.2 | 12.1 | 2.4 | 8.0 | 1.5 | 8.6 | 4.0 | 6.5 |
| d9 | 10.1 | 9.8 | 14.8 | 10.1 | 15.0 | 9.2 | 14.3 | -1.0 | **14.3** | 0.0 | **15.0** | -2.9 | **12.8** |
| d10 | 10.1 | 10.5 | 13.5 | 10.6 | 13.2 | 10.1 | 14.3 | 1.2 | **10.6** | 1.7 | 9.6 | 0.1 | **12.7** |
| Taille UU | | | | | | | | 2.0 | **15.9** | 9.3 | **16.5** | 5.3 | **15.9** |
| Rural | 20.7 | 20.1 | 25.0 | 21.1 | 27.4 | 18.1 | 19.9 | -1.5 | **10.3** | 0.9 | **15.7** | -6.6 | -2.1 |
| 2 000-4 999 h. | 6.4 | 6.0 | 7.2 | 6.4 | 7.7 | 5.0 | 6.2 | -2.0 | 3.2 | 0.0 | 5.0 | -6.2 | -0.9 |
| 5 000-9 999 h. | 5.6 | 5.8 | 6.5 | 5.4 | 6.1 | 6.7 | 7.4 | 1.2 | 4.1 | -0.8 | 2.5 | 4.9 | 7.6 |
| 10 000-19 999 h. | 4.9 | 5.5 | 5.3 | 5.6 | 5.8 | 5.3 | 4.4 | 2.8 | 2.0 | 3.3 | 3.9 | 1.9 | -2.4 |
| 20 000-49 999 h. | 6.4 | 6.0 | 6.5 | 6.3 | 7.1 | 5.4 | 5.3 | -1.7 | 0.6 | -0.5 | 2.8 | -4.4 | -4.7 |
| 50 000-99 999 h. | 7.7 | 7.5 | 7.4 | 7.6 | 7.5 | 7.2 | 7.1 | -0.8 | -1.2 | -0.3 | -0.7 | -1.8 | -2.4 |
| ≥ 100 000 h. | 5.7 | 5.2 | 5.2 | 4.9 | 5.1 | 5.7 | 5.2 | -2.2 | -2.1 | -3.3 | -2.3 | 0.1 | -1.8 |
| ≥200 000 h. | 26.7 | 26.2 | 23.1 | 25.7 | 21.7 | 27.3 | 26.1 | -1.2 | -8.5 | -2.3 | **-11.8** | 1.2 | -1.4 |
| UU de Paris | 16.0 | 17.8 | 13.7 | 17.0 | 11.6 | 19.3 | 18.5 | 4.8 | -6.4 | 2.7 | **-12.8** | 8.8 | 6.5 |
| n= | 50454 | 7083 | 3810 | 4694 | 2561 | 2389 | 1249 | | | | | | |
| Moyenne | | | | | | | | 3.6 | **19.6** | 8.6 | **23.2** | 8.6 | **14.2** |

**Légende :** Ech 1 et 2 : Camme courant ; Ech 3 et -4 : Expérimentation TH ; Ech 1 à 4 : échantillon complet
Ech : échantillon ; Rep=répondants de l'échantillon
*d*=distance standardisée (voir section Analyses statistiques)

Tableau 3 : sélection des variables auxiliaires et paradonnées associées aux 10 variables cibles et à la réponse à l'enquête

| | | | Lien avec les variables cibles $Y_i$ | | Lien avec la réponse |
|---|---|---|---|---|---|
| | signification | Df | Nb F>3 | Nb p<0.05 | *d moyen* |
| *Auxiliaires* | | | | | |
| titre | Sexe de la pers. de ref. | 2 | **7** | 7 | **14.6** |
| ageprcl | Age de la pers. de ref | 8 | **4** | 7 | **10.7** |
| mcdvt1 | Statut matrimonial pers. ref. | 5 | **6** | 7 | **11.3** |
| natlocr | Maison/appartement | 2 | **3** | 3 | **19.6** |
| occr | Locataire/propriétaire | 2 | **6** | 6 | **23.8** |
| courriel | Courriel dans la TH | 2 | **8** | 7 | **17.2** |
| thtelmob | Tel mobile dans la TH | 2 | **5** | 5 | 0.5 |
| thtelfix | Tel fixe dans la TH | 2 | **5** | 3 | **22.8** |
| echant | Echantillon | 4 | **5** | 5 | **12.4** |
| reg | Région | 21 | 0 | 3 | 3.0 |
| idf | Ile-de-France | 2 | **5** | 5 | **11.5** |
| tu_n | | 9 | 1 | | |
| *Paradonnées* | | | | | |
| echant | Echantillon | 4 | 5 | 7 | **12.4** |
| Nb_appels | | 1 | 1* | 0 | -4,3 |
| rrefb | Lettre refus | 2 | **3** | 5 | **38.6** |
| riajb | Lettre imp. à joindre | 2 | **4** | 1 | **74.7** |
| nb_refus | Nb_refus exprimés | 3 | **2** | 2 | **40.7** |
| nb_nonco | Nb non contacts | 6 | **1** | 0 | **21.3** |
| nb_raccr | Nb raccroches | 3 | **5** | 0 | **19.4** |
| nb_rates | Nb appels ratés** | 6 | **5** | 2 | **29.6** |
| nb_refrac | Somme refus/raccroches | 3 | **4** | 1 | **42.4** |

En gras : variables retenues suivant les critères définies dans la section Analyses statistiques
*d* de niveau variable (variables catégorielles) ou d (variables continues) : voir section Analyses statistiques
* : Pour le lien avec les Y, pour Nb_appels, Nb F>3 est en fait une valeur *d* car la variable est continue.
** : répondeurs, problèmes ou occupés

Tableau 4 : proportions de modalités 1 des variables cibles suivant les redressements et écart au calage direct

| | Calage direct | | CNRT_AUX + calage | | CNRT_PARA + calage | | CNRT_COMP + calage | |
|---|---|---|---|---|---|---|---|---|
| Y* | % | StdErr | % | Deff | % | Deff | % | Deff |
| NVP | 8.5 | 0.49 | 8.4 | 1.01 | 8.5 | 1.12 | 8.4 | 1.11 |
| NVF | 19.6 | 0.68 | 19.5 | 1.01 | 20.2 | 1.13 | 20.1 | 1.13 |
| SFP | 12.8 | 0.59 | 12.7 | 1.02 | 13.0 | 1.12 | 12.7 | 1.10 |
| SFF | 17.4 | 0.66 | 17.5 | 1.02 | 17.6 | 1.13 | 17.7 | 1.15 |
| OE | 51.3 | 0.86 | 51.4 | 1.02 | 51.3 | 1.10 | 51.1 | 1.12 |
| PP | 60.6 | 0.84 | 59.9 | 1.03 | 61.6 | 1.08 | 61.1 | 1.10 |
| PF | 43.2 | 0.85 | 43.1 | 1.02 | 43.1 | 1.10 | 43.3 | 1.12 |
| CEA | 37.1 | 0.83 | 36.3 | 1.01 | 36.7 | 1.10 | 36.0 | 1.10 |
| CEF | 47.9 | 0.84 | 47.4 | 1.02 | 47.6 | 1.10 | 47.1 | 1.12 |
| EF | 27.1 | 0.76 | 27.1 | 1.02 | 27.3 | 1.11 | 27.3 | 1.13 |
| Ecart absolu maximum | | | 0.8 | | 1.1 | | 1.1 | |

* : voir la section Présentation de Camme TH pour la signification des variables

Tableau 5 : Statistiques de poids des différents redressements

| Variable | Min | P1 | P5 | P95 | P99 | Max | CV | Somme | Max / min |
|---|---|---|---|---|---|---|---|---|---|
| Uniforme | 7500.6 | 7500.6 | 7500.6 | 7500.6 | 7500.6 | 7500.6 | 0.0 | 28577312 | 1.0 |
| Poids de sondage | 314.6 | 314.6 | 314.6 | 614.6 | 614.6 | 614.6 | 18.5 | 2175312 | 2.0 |
| Calage direct | 2396.7 | 4286.3 | 4752.3 | 12747.9 | 17140.1 | 29022.4 | 36.6 | 28577312 | 12.1 |
| CNRT_AUX_cal | 2131.2 | 3847.6 | 4291.8 | 13735.4 | 18165.5 | 34728.0 | 42.4 | 28577312 | 14.9 |
| CNRT_PARA_cal | 1600.1 | 3101.9 | 3576.4 | 16908.3 | 24157.2 | 47799.7 | 61.4 | 28577312 | 31.1 |
| CNRT_COMP_cal | 1699.9 | 3406.7 | 3701.1 | 17265.1 | 25512.1 | 50348.6 | 64.3 | 28577312 | 16.6 |

Tableau 6 : corrélations entre les variables cibles et les prédictions de la première composante principale des variables cibles par les variables auxiliaires et de paradonnées et leur combinaison (analyse de robustesse)

| Variable Y | **AUX** | **PARA** | **COMP** |
|---|---|---|---|
| **nvp** | 0.06 | 0.00 | 0.06 |
| **nvf** | 0.09 | 0.02 | 0.09 |
| **sfp** | 0.18 | 0.05 | 0.18 |
| **sff** | 0.15 | 0.08 | 0.15 |
| **oe** | 0.08 | 0.03 | 0.08 |
| **pp** | 0.17 | 0.08 | 0.18 |
| **pf** | 0.02 | 0.01 | 0.03 |
| **cea** | **0.24** | 0.05 | **0.24** |
| **cef** | **0.28** | 0.11 | **0.28** |
| **ef** | 0.08 | 0.05 | 0.09 |

En grisé : corrélation supérieure à 0,10 ; en gras : corrélation supérieure à 0,2.

## **Bibliographie** *Calibri 13*


1. Haziza, D. and E. Lesage, *A discussion of weighting procedures for unit nonresponse.* Journal of Official Statistics, 2016. **32**(1): p. 129-145.
2. Deville, J. and C.-E. Sarndal, *Calibration estimators in survey sampling.* Journal of the American Statistical Association, 1992. **87**(418): p. 376-382.
3. Little, R.J.A. and S. Vartivarian, *Does weighting for nonresponse increase the variance of survey means?* Survey methodology, 2005. **31**(2): p. 161ï 68.
4. Couper, M., *Measuring survey quality in a CASIC environment*, in *Survey Research Methods Section of the American Statistical Association*, A.s. association, Editor. 1998, American statistical association. p. 41-49.
5. Olson, K., *Paradata for Nonresponse Adjustment.* The Annals of the American Academy of Political and Social Science, 2013. **645**(1): p. 142-170.
6. Bethlehem, J., F. Cobben, and B. schouten, *Des indicateurs de la représentativité aux enquêtes.* Techniques d'Enquêtes, 2009(Recueil du symposium 2008 de Statistique Canada): p. 1-10.
7. Austin, P.C. and E.A. Stuart, *Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies.* Stat Med, 2015. **34**(28): p. 3661-79.
8. Myers, J.A., et al., *Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates.* American Journal of Epidemiology, 2011. **174**(11): p. 1213-1222.
9. Kreuter, F., et al., *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**(2): p. 389-407.
10. Biemer, P.P., P. Chen, and K. Wang, *Using level-of-effort paradata in non-response adjustments with application to field surveys.* Journal of the Royal Statistical Society Series A, 2013. **176**(1): p. 147-168.