Variables auxiliaires et paradonnées pour améliorer le redressement d'une enquête téléphonique : 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 des conditions de vie des ménages

(***) Insee, Unité des prix à la consommation et des enquêtes ménages

(****) *Insee, Département de la Conjoncture*

Plan de la présentation

Objectif : établir l'utilité des variables auxiliaires et des paradonnées pour la correction de la non-réponse de Camme

- Présentation de Camme et de l'expérimentation TH
- La correction de la non-réponse totale (CNRT)
- Résultats
- Conclusions

Présentation de Camme

- Base de sondage : taxe d'habitation
- Sondage aléatoire stratifié suivant l'âge de la personne de référence fiscale (< 30 ans surreprésentés)
- A partir du nom/prénom et de l'adresse de la personne de référence du ménage, recherche d'un numéro de téléphone
- Couverture : 30 % des adresses TH
- Panel : trois interrogations successives

  Enquête ménage et non individu (Une seule personne par ménage)
- Questionnement : opinions sur la conjoncture économique passée et sur la situation future
  - Tirées d'échelles ordonnées à 5 modalités, regroupées en -1, 0, 1.
- Redressement actuel : calage direct sur
  - sexe, âge, taille du ménage, groupe social, taille d'UU, type de logement

L'expérimentation Camme-TH

- Base de sondage identique
- Plan de sondage identique
- Couverture : utilisation en plus de l'annuaire, des numéros de téléphone donnés dans le fichier TH :

  70% des adresses TH couvertes
- Expérimentation de mai à septembre 2017.

  - Échantillons analysés ici:
    - Première interrogation uniquement
    - tous les échantillons sur les 4 premiers mois
    - n=6122

Les échantillons

Echant	Non-répondants	Répondants	Total	Taux de réponse	
1 : Num. Annuaire	698	821	1 519	54,0%	Camme Courant
2 : Num. Annuaire + TH Tel Annuaire	1 086	1 740	2 826	61,6%	
3 : Num. Annuaire + TH Tel TH	49	170	219	77,6%	Expérimentation TH
4 : Num. TH	479	1 079	1 558	69,3%	
Total	2 312	3 810	6 122	62,3%	

Taux de réponse supérieur dans les numéros TH: apport très positif

Redressement : en une ou deux étapes ?

- Une étape : calage direct
  - Nécessite uniquement des variables renseignées par les répondants et des totaux/pourcentages de la population cible pour ces mêmes variables → CAMME actuel
- Deux étapes:
  - correction de la non-réponse totale (CNRT): utiliser des variables recueillies pour les répondants et non-répondants pour modéliser la réponse; repondérer pour que les répondants représentent les non-répondants
  - Calage : à partir du poids CNRT, caler comme dans le calage direct
  - Souvent de meilleure qualité (Haziza & Lesage, 2016) → À tester

La correction de la non-réponse totale :
Critères des variables de redressement

4 critères (Little & Vartivarian, 2005)

- Pas de valeurs manquantes (renseignées pour les répondants et les non-répondants)
- Lien avec la non-réponse
- Lien avec les variables Y de l'enquête (rho≥0,5)
- Pas d'erreur de mesure

La correction de la non-réponse totale
1) Les variables auxiliaires

12 variables auxiliaires disponibles dans la base de sondage de CAMME

- Sexe, âge et statut matrimonial de la personne de référence fiscale
- nature du logement (maison/autre); statut d'occupation (propriétaire/autre)
- Région; taille d'unité urbaine
- Numéro de téléphone fixe, Numéro de téléphone mobile, Courriel présents dans les fichiers fiscaux
- Échantillon (provenance du numéro appelé)
- Revenu fiscal du ménage

La correction de la non-réponse totale :
2) Les paradonnées

- Efforts pour joindre le répondant (Couper, 1998; Olson, 2013)
    - Séquences des issues des contacts (enquête transversale) ou séquences des participations antérieures (enquête longitudinale)
- Descriptions du bâti, du quartier etc. faites par l'enquêteur

➔ Connues pour les répondants et non-répondants
➔ Utiles en CNRT

- Temps de passation (global, par module, par question), pauses, hésitations, débit de voix de l'enquêteur, etc.
- Eye-tracking, rythme des frappes sur le clavier etc.

➔ Connues pour les répondants uniquement
➔ Utiles en monitorage qualité

La correction de la non-réponse totale :
2) Les paradonnées disponibles dans Camme

6 variables

- Nombre d'appels (1-20)
- Nombre de refus
  (5 premiers appels uniquement)
- Nombre de raccroches
  (5 premiers appels uniquement)
- Nombre de non-contacts
  (5 premiers appels uniquement)
- Lettre de relance pour les impossibles à joindre
- Lettre de relance pour les refus

La correction de la non-réponse totale :
2) Des exemples de redressement avec paradonnées

- Beaucoup d'exemples de monitoring, responsive design
- Quelques redressements : (Wisner, Phillips, Baribeau, & Lévesque, 2009)
- La conclusion est toutefois qu'il est difficile de trouver des variables qui remplissent les quatre conditions (Kreuter et al., 2010); en particulier le lien avec les variables d'intérêt

La correction de la non-réponse totale

3) Sélection des variables liées aux variables d'intérêt Y

Variables continues :

Corrélations des Y avec les variables auxiliaires et paradonnées
(Rho de Pearson)

	CEA	CEF	EF	NVF	NVP	OE	PF	PP	SFF	SFP	Moy(abs)
Revenu fiscal	**0.17**	**0.13**	0.06	**0.10**	0.03	0.04	0.02	0.09	0.03	0.09	0.08
Décile revenu fiscal	**0.26**	**0.19**	**0.10**	0.09	0.03	0.02	0.03	**0.15**	0.01	**0.10**	**0.10**
Taille UU	-0.01	0.01	0.05	0.03	-0.01	0.08	-0.02	0.04	0.05	0.02	0.03
NbAppels	0.00	0.01	0.00	0.00	-0.02	0.01	0.00	0.03	0.03	0.02	0.01
nb_refus	-0.01	-0.03	-0.02	-0.01	-0.02	-0.03	0.00	0.00	-0.02	-0.01	0.02
nb_noncontacts	-0.01	-0.01	0.00	0.02	0.01	0.00	-0.03	0.00	0.02	0.02	0.01
nb_raccroches	-0,02	-0.04	-0.03	-0.01	0.02	-0.01	-0.03	-0.03	-0.02	-0.02	0.02

Lecture : En grisé : |rho|≥ 0,05; en gras : |rho|≥ 0,1

Les plus corrélées : Revenu du foyer fiscal (surtout en déciles), taille d'unité urbaine

La correction de la non-réponse totale

3) Sélection des variables liées aux variables d'intérêt Y

Variables catégorielles :

Lien des Y avec les variables auxiliaires et paradonnées
(ANOVA : valeurs F - statistique de Fisher)

	CEA	CEF	EF	NVF	NVP	OE	PF	PP	SFF	SFP	Nb F≥5
Titre (sexe)	**33.2**	**18.8**	1.6	4.5	5.8	1.2	6.3	**12.6**	2.4	6.2	5
Age_Pref	1.3	**25.6**	2.6	1.2	2.4	5.8	0.4	2.9	**32.8**	**12.1**	4
Statut matri.	**17.1**	**22.5**	0.7	0.4	1.0	5.0	1.5	7.1	**18.0**	**12.0**	**6**
Maison/appart	**10.5**	0.1	0.1	0.0	0.1	**16.1**	0.2	0.2	**29.5**	1.4	3
Propriét./locat.	**86.4**	**19.6**	5.9	4.3	0.7	2.9	0.7	**26.9**	**28.0**	0.4	**5**
Courriel	**36.7**	**93.6**	7.4	3.1	1.5	8.5	0.4	**49.1**	**18.9**	**22.6**	**7**
TH_tel_mob	0.1	**39.1**	0.0	0.4	0.4	**10.4**	0.0	4.7	**48.4**	**14.2**	**4**
TH_tel_fixe	**15.2**	0.0	3.5	2.5	0.8	6.3	0.6	3.1	**12.1**	1.0	3
Echant_TH	3.6	**13.0**	3.3	1.3	0.7	2.2	0.0	5.0	**10.1**	1.6	3
Région	2.4	2.3	1.2	1.5	1.4	2.0	1.1	1.6	1.5	1.2	0
IdF	0.0	0.2	8.7	6.1	1.5	**23.8**	2.0	**12.7**	5.5	0.0	5
Lettre refus	2.2	2.9	5.0	0.1	0.0	0.5	0.4	3.1	3.5	1.4	1
Lettre injoignable	0.3	0.2	1.1	3.7	0.1	0.9	0.3	3.5	5.2	4.0	1

Lecture : en grisé : F≥5, en gras: F≥10. Les variables retenues sont en grisé.

La correction de la non-réponse totale

4) Sélection des variables liées à la réponse

Variables catégorielles

Différences standardisées *d*

	modalités	*d*	P-value
Titre (sexe)	2	14.6	***
Age_Pref	8	10.7	***
Maison/appart	2	19.6	***
Statut matrimonial	5	11.3	***
Propriétaire/locataire	2	23.8	***
Courriel	2	17.2	***
TH_tel_mob	2	0.5	***
TH_tel_fixe	2	22.8	***
Echant_TH	4	12.4	***
Région	21	3.0	***
IdF	2	11.5	***
Lettre refus	2	38.6	***
Lettre injoignable	2	74.7	***
nb_non contacts	6	21.3	***
nb_refus_raccroches	3	42.4	***

$$d_{bin} = 100 \times \frac{p_1 - p_0}{\sqrt{0.5 \times (p_1 q_1 + p_0 q_0)}}$$

$$d_{multi} = \frac{1}{N} \sum |d_{bin}|$$

Lecture : en grisé : d ≥ 10

Les plus corrélées : type de logement et statut d'occupation; téléphone fixe dans la TH, lettres aux refus et aux injoignables, nombre de non contacts et de raccroches

La correction de la non-réponse totale

4) Sélection des variables liées à la réponse

Variables continues :

Différences standardisées *d* (critère: d≥ 10)

	d	P-value
Revenu fiscal	0.6	***
Décile revenu fiscal	1.1	***
Taille d'UU	-0.5	***
NbAppel	-4.3	***
nb_refus	-2.2	***
nb_non contacts	-2.1	***
nb_raccroches	-1.1	***

Aucune liaison importante

CNRT puis calage

3 modèles de CNRT

- auxiliaires, paradonnées, auxiliaires + paradonnées
- GRH avec méthode des quantiles
  - Définition de déciles (noniles pour les paradonnées)
  - Repondération par l'inverse du % de répondants dans chaque quantile

Calage : calage classique de Camme

- Nombre de personnes dans le ménage (1, 2, 3, 4+)
- Âge de la personne de référence du ménage (4 tranches)
- CS de la personne de référence du ménage (6 catégories)
- Type de logement (2 catégories)
- Taille d'UU (6 catégories)

Modèles de CNRT

1. Variables auxiliaires

- On introduit toutes les variables auxiliaires retenues précédemment dans les analyses de liaison
- On ajoute les carrés de : décile de revenu, âge de la personne de référence
- On croise les variables de types de téléphone et celle de l'échantillon TH
- Modélisation logistique stepwise avec le poids de sondage normalisé (slstay=0,1)

<u>Résultat : 9 variables</u>

- Age, Age², échantillon, Idf, statut matrimonial, statut d'occupation, telfix, telmob*échantillon, revenu du foyer

Modèles de CNRT

2. Paradonnées uniquement

La stratégie est identique à celle de la sélection des auxiliaires

- On inclut également la variable échantillon
- On ajoute les interactions

<u>Résultat : 10 variables</u>

- Echant, nb raccroches, nb non-contacts, Lettre IAJ, Lettre refus, Echant*nb non-contacts, Lettre IAJ*nb non-contacts, Lettre refus*nb non-contacts, Lettre IAJ*nb raccroches, Lettre refus*nb raccroches

Modèles de CNRT

3. Variables auxiliaires + paradonnées

- On garde les variables auxiliaires retenues; on teste l'inclusion des paradonnées identifiées dans l'analyse des liens
- Modèle logistique stepwise avec le poids de sondage normalisé (slstay=0,1)
- Résultat : les 18 variables

Résultats

Comparaison des pourcentages de modalité « 1 » des variables Y

Variables d'intérêt	Calage direct		CNRT auxiliaires		CNRT paradonnées		CNRT complet	
	%	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	1.08	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

Deff=ratio des erreurs standards CNRT/calage direct

Bilan : AUCUN EFFET du type de CNRT sur les résultats
(sauf variable CEA de situation financière actuelle du ménage)

Résultats

Statistiques de poids

	Min	Max	Max/min	Coeff de Variation	Somme
Poids uniforme	7501	7501	1.0	0	28 577 312
Poids de sondage	315	615	2.0	18.5	2 175 312
Calage direct	2397	29022	12.1	36.6	28 577 312
CNRT aux. + calage	2174	34567	15.9	42.5	28 577 312
CNRT para. + calage	1860	50479	27.1	57.6	28 577 312
CNRT aux/para + calage	1772	50166	28.3	58.8	28 577 312

Calage direct : dispersion des poids la plus basse

Bilan : apport de la CNRT limité pour CAMME, y compris lorsqu'on inclue des paradonnées

Limites : particularités de Camme
La couverture et le taux de réponse

- 70 % de couverture
- 62 % de taux de réponse
- Pas de sélection aléatoire du répondant

  → La population répondante est sans doute particulière

- Sur Camme courant seul ? Idem
  - Paradonnées et variables auxiliaires sans effet notable

Limites : particularités de Camme
La définition des soldes

Questions très particulières

Regroupement des modalités de 1 à 5 en -1, 0, 1 qui écrase les différences et la non-réponse partielle

Deux stratégies d'agrégation :

- Pour toutes les variables sauf pp et pf:
  1, 2 ➔ 1
  3, manquant ➔ 0
  4, 5 ➔ -1
- Pour pp (évolution des prix passés) et pf (perception de l'évolution future des prix) :
  1 ➔ 1
  Manquant, 2 ➔ 0
  3, 4, 5 ➔ -1

Taux de NR compris entre 0,4% (situation financière passée et situation financière future) et 8% (niveau de vie futur en France et évolution future des prix).

Limites : particularités de Camme

Un redressement à l'effet réduit…

Variables d'intérêt	Calage direct		Poids de sondage		Poids uniforme	
	%	StdErr	%	Ecart	%	Ecart
nvp	8.5	0.49	8.0	-0.4	8.2	-0.2
nvf	19.6	0.68	20.1	0.5	19.9	0.3
sfp	12.8	0.59	11.3	-1.5	12.1	-0.7
sff	17.4	0.66	14.6	-2.8	16.0	-1.4
oe	51.3	0.86	50.0	-1.3	50.7	-0.7
pp	60.6	0.84	60.6	0.0	60.7	0.1
pf	43.2	0.85	43.8	0.6	43.8	0.6
cea	37.1	0.83	37.3	0.2	37.9	0.8
cef	47.9	0.84	45.7	-2.3	47.3	-0.6
ef	27.1	0.76	27.6	0.6	27.5	0.4

Quelles conclusions pour d'autres enquêtes ?

- Utilisation des paradonnées pour le redressement d'enquêtes sans base de sondage ?
- Les paradonnées peuvent elles remplacer le revenu (souvent lié aux variables d'intérêt lors d'un calage ou celui-ci est manquant ?

Modélisation ANOVA des déciles de revenus par les variables auxiliaires et les paradonnées

Variable	F	P
TH courriel	861	0,000
TH tel fixe	76	0,000
Echantillon	32	0,000
Lettre de refus	22	0,000
TH tel mob	15	0,000

Bilan :
→ très peu de paradonnées explicatives pour des enquêtes non tirées dans la TH
→ envoi d'une lettre de refus

Côté variables quantitatives :
→ agrégation refus et raccroches légèrement corrélée au revenu (rho=-0,08).
→ Le nombre de tentatives d'appel n'est pas corrélé au revenu

Conclusion

- Utiliser les numéros TH améliore beaucoup la couverture et le taux de réponse
- Camme est robuste : variables auxiliaires et paradonnées modifient peu les estimations des variables de solde (en fait, rien ne les affecte sensiblement)
- Peu de liens entre les paradonnées et les variables d'intérêt ou auxiliaires → leur utilité semble limitée pour la CNR

Références

- Couper, M.P. 1998. Measuring survey quality in a CASIC environment. Proceedings of the Survey Research Methods Section of the American Statistical Association, 41–49.
- Haziza, D., & Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics, 32*(1), 129-145. doi:10.1515/JOS-2016-0006
- Kreuter, F. (Ed.) (2013). *Improving surveys with paradata*. New York: John Wiley & Sons.
- Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T., Casas-Cordero, C., & Lemay, M. (2010). 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, 173*(2), 389-407.
- Little, R. J. A., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey methodology, 31*(2), 161–168.
- Olson, K. (2013). Paradata for Nonresponse Adjustment. *The Annals of the American Academy of Political and Social Science, 645*(1), 142-170. doi:10.1177/0002716212459475
- Wisner, J., Phillips, O., Baribeau, B., & Lévesque, A. (2009). *Using paradata to manage nonresponse in the Survey of Labour and Income Dynamics*. Paper presented at the Symposium 2008.

Variables auxiliaires et paradonnées pour améliorer le redressement d'une enquête téléphonique : 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 des conditions de vie des ménages

(***) Insee, Unité des prix à la consommation et des enquêtes ménages

(****) *Insee, Département de la Conjoncture*

Plan de la présentation

Objectif : établir l'utilité des variables auxiliaires et des paradonnées pour la correction de la non-réponse de Camme

- Présentation de Camme et de l'expérimentation TH
- La correction de la non-réponse totale (CNRT)
- Résultats
- Conclusions

Présentation de Camme

- Base de sondage : taxe d'habitation
- Sondage aléatoire stratifié suivant l'âge de la personne de référence fiscale (< 30 ans surreprésentés)
- A partir du nom/prénom et de l'adresse de la personne de référence du ménage, recherche d'un numéro de téléphone
- Couverture : 30 % des adresses TH
- Panel : trois interrogations successives

  Enquête ménage et non individu (Une seule personne par ménage)
- Questionnement : opinions sur la conjoncture économique passée et sur la situation future
  - Tirées d'échelles ordonnées à 5 modalités, regroupées en -1, 0, 1.
- Redressement actuel : calage direct sur
  - sexe, âge, taille du ménage, groupe social, taille d'UU, type de logement

Exemples de questions dans Camme :

A votre avis, au cours des 12 derniers moins, le niveau de vie en France dans l'ensemble
1 s'est nettement améliorée
2 s'est un peu améliorée
3. est restée stationnaire
4. s'est un peu dégradée
5. s'est nettement dégradée

Est-ce le bon moment pour épargner ?
Oui certainement
Oui peut-être ….

L'expérimentation Camme-TH

- Base de sondage identique
- Plan de sondage identique
- Couverture : utilisation en plus de l'annuaire, des numéros de téléphone donnés dans le fichier TH :
  70% des adresses TH couvertes
- Expérimentation de mai à septembre 2017.

  - Échantillons analysés ici:
    - Première interrogation uniquement
    - tous les échantillons sur les 4 premiers mois
    - n=6122

Les échantillons

Echant	Non-répondants	Répondants	Total	Taux de réponse	
1 : Num. Annuaire	698	821	1 519	54,0%	Camme Courant
2 : Num. Annuaire + TH Tel Annuaire	1 086	1 740	2 826	61,6%	
3 : Num. Annuaire + TH Tel TH	49	170	219	77,6%	Expérimentation TH
4 : Num. TH	479	1 079	1 558	69,3%	
Total	2 312	3 810	6 122	62,3%	

Taux de réponse supérieur dans les numéros TH: apport très positif

Redressement : en une ou deux étapes ?

- Une étape : calage direct
  - Nécessite uniquement des variables renseignées par les répondants et des totaux/pourcentages de la population cible pour ces mêmes variables → CAMME actuel
- Deux étapes:
  - correction de la non-réponse totale (CNRT): utiliser des variables recueillies pour les répondants et non-répondants pour modéliser la réponse; repondérer pour que les répondants représentent les non-répondants
  - Calage : à partir du poids CNRT, caler comme dans le calage direct
  - Souvent de meilleure qualité (Haziza & Lesage, 2016) → À tester

La correction de la non-réponse totale :
Critères des variables de redressement

4 critères (Little & Vartivarian, 2005)

- Pas de valeurs manquantes (renseignées pour les répondants et les non-répondants)
- Lien avec la non-réponse
- Lien avec les variables Y de l'enquête (rho≥0,5)
- Pas d'erreur de mesure

La correction de la non-réponse totale
1) Les variables auxiliaires

12 variables auxiliaires disponibles dans la base de sondage de CAMME

- Sexe, âge et statut matrimonial de la personne de référence fiscale
- nature du logement (maison/autre); statut d'occupation (propriétaire/autre)
- Région; taille d'unité urbaine
- Numéro de téléphone fixe, Numéro de téléphone mobile, Courriel présents dans les fichiers fiscaux
- Échantillon (provenance du numéro appelé)
- Revenu fiscal du ménage

La correction de la non-réponse totale :
2) Les paradonnées

- Efforts pour joindre le répondant (Couper, 1998; Olson, 2013)
    - Séquences des issues des contacts (enquête transversale) ou séquences des participations antérieures (enquête longitudinale)
- Descriptions du bâti, du quartier etc. faites par l'enquêteur

    ➔ Connues pour les répondants et non-répondants

    ➔ Utiles en CNRT
- Temps de passation (global, par module, par question), pauses, hésitations, débit de voix de l'enquêteur, etc.
- Eye-tracking, rythme des frappes sur le clavier etc.

    ➔ Connues pour les répondants uniquement

    ➔ Utiles en monitorage qualité

La correction de la non-réponse totale :
2) Les paradonnées disponibles dans Camme

6 variables

- Nombre d'appels (1-20)
- Nombre de refus
  (5 premiers appels uniquement)
- Nombre de raccroches
  (5 premiers appels uniquement)
- Nombre de non-contacts
  (5 premiers appels uniquement)
- Lettre de relance pour les impossibles à joindre
- Lettre de relance pour les refus

La correction de la non-réponse totale :
2) Des exemples de redressement avec paradonnées

- Beaucoup d'exemples de monitoring, responsive design
- Quelques redressements : (Wisner, Phillips, Baribeau, & Lévesque, 2009)
- La conclusion est toutefois qu'il est difficile de trouver des variables qui remplissent les quatre conditions (Kreuter et al., 2010); en particulier le lien avec les variables d'intérêt

La correction de la non-réponse totale

3) Sélection des variables liées aux variables d'intérêt Y

Variables continues :

Corrélations des Y avec les variables auxiliaires et paradonnées
(Rho de Pearson)

	CEA	CEF	EF	NVF	NVP	OE	PF	PP	SFF	SFP	Moy(abs)
Revenu fiscal	**0.17**	**0.13**	0.06	**0.10**	0.03	0.04	0.02	0.09	0.03	0.09	0.08
Décile revenu fiscal	**0.26**	**0.19**	**0.10**	0.09	0.03	0.02	0.03	**0.15**	0.01	**0.10**	**0.10**
Taille UU	-0.01	0.01	0.05	0.03	-0.01	0.08	-0.02	0.04	0.05	0.02	0.03
NbAppels	0.00	0.01	0.00	0.00	-0.02	0.01	0.00	0.03	0.03	0.02	0.01
nb_refus	-0.01	-0.03	-0.02	-0.01	-0.02	-0.03	0.00	0.00	-0.02	-0.01	0.02
nb_noncontacts	-0.01	-0.01	0.00	0.02	0.01	0.00	-0.03	0.00	0.02	0.02	0.01
nb_raccroches	-0,02	-0.04	-0.03	-0.01	0.02	-0.01	-0.03	-0.03	-0.02	-0.02	0.02

Lecture : En grisé : |rho|≥ 0,05; en gras : |rho|≥ 0,1

Les plus corrélées : Revenu du foyer fiscal (surtout en déciles), taille d'unité urbaine

La correction de la non-réponse totale

3) Sélection des variables liées aux variables d'intérêt Y

Variables catégorielles :

Lien des Y avec les variables auxiliaires et paradonnées
(ANOVA : valeurs F - statistique de Fisher)

	CEA	CEF	EF	NVF	NVP	OE	PF	PP	SFF	SFP	Nb F≥5
Titre (sexe)	**33.2**	**18.8**	1.6	4.5	5.8	1.2	6.3	**12.6**	2.4	6.2	5
Age_Pref	1.3	**25.6**	2.6	1.2	2.4	5.8	0.4	2.9	**32.8**	**12.1**	4
Statut matri.	**17.1**	**22.5**	0.7	0.4	1.0	5.0	1.5	7.1	**18.0**	**12.0**	**6**
Maison/appart	**10.5**	0.1	0.1	0.0	0.1	**16.1**	0.2	0.2	**29.5**	1.4	3
Propriét./locat.	**86.4**	**19.6**	5.9	4.3	0.7	2.9	0.7	**26.9**	**28.0**	0.4	**5**
Courriel	**36.7**	**93.6**	7.4	3.1	1.5	8.5	0.4	**49.1**	**18.9**	**22.6**	**7**
TH_tel_mob	0.1	**39.1**	0.0	0.4	0.4	**10.4**	0.0	4.7	**48.4**	**14.2**	**4**
TH_tel_fixe	**15.2**	0.0	3.5	2.5	0.8	6.3	0.6	3.1	**12.1**	1.0	3
Echant_TH	3.6	**13.0**	3.3	1.3	0.7	2.2	0.0	5.0	**10.1**	1.6	3
Région	2.4	2.3	1.2	1.5	1.4	2.0	1.1	1.6	1.5	1.2	0
IdF	0.0	0.2	8.7	6.1	1.5	**23.8**	2.0	**12.7**	5.5	0.0	5
Lettre refus	2.2	2.9	5.0	0.1	0.0	0.5	0.4	3.1	3.5	1.4	1
Lettre injoignable	0.3	0.2	1.1	3.7	0.1	0.9	0.3	3.5	5.2	4.0	1

Lecture : en grisé : F≥5, en gras: F≥10. Les variables retenues sont en grisé.

Variables les plus corrélées :

Courriel,
Statut matrimonial de la personne de référence,
Statut d'occupation du logement,
Sexe,
Île-de-France/Autre
Laisser un numéro de mobile à l'administration fiscale
Age de la personne de référence fiscale,
Type de logement,
Echantillon

La correction de la non-réponse totale

4) Sélection des variables liées à la réponse

Variables catégorielles

Différences standardisées *d*

	modalités	*d*	P-value
Titre (sexe)	2	14.6	***
Age_Pref	8	10.7	***
Maison/appart	2	19.6	***
Statut matrimonial	5	11.3	***
Propriétaire/locataire	2	23.8	***
Courriel	2	17.2	***
TH_tel_mob	2	0.5	***
TH_tel_fixe	2	22.8	***
Echant_TH	4	12.4	***
Région	21	3.0	***
IdF	2	11.5	***
Lettre refus	2	38.6	***
Lettre injoignable	2	74.7	***
nb_non contacts	6	21.3	***
nb_refus_raccroches	3	42.4	***

$$d_{bin} = 100 \times \frac{p_1 - p_0}{\sqrt{0.5 \times (p_1 q_1 + p_0 q_0)}}$$

$$d_{multi} = \frac{1}{N} \sum |d_{bin}|$$

Lecture : en grisé : d ≥ 10

Les plus corrélées : type de logement et statut d'occupation; téléphone fixe dans la TH, lettres aux refus et aux injoignables, nombre de non contacts et de raccroches

La correction de la non-réponse totale

4) Sélection des variables liées à la réponse

Variables continues :

Différences standardisées *d* (critère: d≥ 10)

	d	P-value
Revenu fiscal	0.6	***
Décile revenu fiscal	1.1	***
Taille d'UU	-0.5	***
NbAppel	-4.3	***
nb_refus	-2.2	***
nb_non contacts	-2.1	***
nb_raccroches	-1.1	***

Aucune liaison importante

CNRT puis calage

3 modèles de CNRT
- auxiliaires, paradonnées, auxiliaires + paradonnées
- GRH avec méthode des quantiles
  - Définition de déciles (noniles pour les paradonnées)
  - Repondération par l'inverse du % de répondants dans chaque quantile

Calage : calage classique de Camme
- Nombre de personnes dans le ménage (1, 2, 3, 4+)
- Âge de la personne de référence du ménage (4 tranches)
- CS de la personne de référence du ménage (6 catégories)
- Type de logement (2 catégories)
- Taille d'UU (6 catégories)

Modèles de CNRT
1. Variables auxiliaires

- On introduit toutes les variables auxiliaires retenues précédemment dans les analyses de liaison
- On ajoute les carrés de : décile de revenu, âge de la personne de référence
- On croise les variables de types de téléphone et celle de l'échantillon TH
- Modélisation logistique stepwise avec le poids de sondage normalisé (slstay=0,1)

<u>Résultat : 9 variables</u>

- Age, Age², échantillon, Idf, statut matrimonial, statut d'occupation, telfix, telmob*échantillon, revenu du foyer

Modèles de CNRT

2. Paradonnées uniquement

La stratégie est identique à celle de la sélection des auxiliaires

- On inclut également la variable échantillon
- On ajoute les interactions

Résultat : 10 variables

- Echant, nb raccroches, nb non-contacts, Lettre IAJ, Lettre refus, Echant*nb non-contacts, Lettre IAJ*nb non-contacts, Lettre refus*nb non-contacts, Lettre IAJ*nb raccroches, Lettre refus*nb raccroches

Modèles de CNRT

3. Variables auxiliaires + paradonnées

- On garde les variables auxiliaires retenues; on teste l'inclusion des paradonnées identifiées dans l'analyse des liens
- Modèle logistique stepwise avec le poids de sondage normalisé (slstay=0,1)
- Résultat : les 18 variables

Résultats

Comparaison des pourcentages de modalité « 1 » des variables Y

Variables d'intérêt	Calage direct		CNRT auxiliaires		CNRT paradonnées		CNRT complet	
	%	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	1.08	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

Deff=ratio des erreurs standards CNRT/calage direct

Bilan : AUCUN EFFET du type de CNRT sur les résultats
(sauf variable CEA de situation financière actuelle du ménage)

Résultats

Statistiques de poids

	Min	Max	Max/min	Coeff de Variation	Somme
Poids uniforme	7501	7501	1.0	0	28 577 312
Poids de sondage	315	615	2.0	18.5	2 175 312
Calage direct	2397	29022	12.1	36.6	28 577 312
CNRT aux. + calage	2174	34567	15.9	42.5	28 577 312
CNRT para. + calage	1860	50479	27.1	57.6	28 577 312
CNRT aux/para + calage	1772	50166	28.3	58.8	28 577 312

Calage direct : dispersion des poids la plus basse

Bilan : apport de la CNRT limité pour CAMME, y compris lorsqu'on inclue des paradonnées

Limites : particularités de Camme

La couverture et le taux de réponse

- 70 % de couverture
- 62 % de taux de réponse
- Pas de sélection aléatoire du répondant

  → La population répondante est sans doute particulière

- Sur Camme courant seul ? Idem
  - Paradonnées et variables auxiliaires sans effet notable

Limites : particularités de Camme
La définition des soldes

Questions très particulières

Regroupement des modalités de 1 à 5 en -1, 0, 1 qui écrase les différences et la non-réponse partielle

Deux stratégies d'agrégation :

- Pour toutes les variables sauf pp et pf:
  1, 2 ➔ 1
  3, manquant ➔ 0
  4, 5 ➔ -1
- Pour pp (évolution des prix passés) et pf (perception de l'évolution future des prix) :
  1 ➔ 1
  Manquant, 2 ➔ 0
  3, 4, 5 ➔ -1

Taux de NR compris entre 0,4% (situation financière passée et situation financière future) et 8% (niveau de vie futur en France et évolution future des prix).

Limites : particularités de Camme

Un redressement à l'effet réduit…

Variables d'intérêt	Calage direct		Poids de sondage		Poids uniforme	
	%	StdErr	%	Ecart	%	Ecart
nvp	8.5	0.49	8.0	-0.4	8.2	-0.2
nvf	19.6	0.68	20.1	0.5	19.9	0.3
sfp	12.8	0.59	11.3	-1.5	12.1	-0.7
sff	17.4	0.66	14.6	-2.8	16.0	-1.4
oe	51.3	0.86	50.0	-1.3	50.7	-0.7
pp	60.6	0.84	60.6	0.0	60.7	0.1
pf	43.2	0.85	43.8	0.6	43.8	0.6
cea	37.1	0.83	37.3	0.2	37.9	0.8
cef	47.9	0.84	45.7	-2.3	47.3	-0.6
ef	27.1	0.76	27.6	0.6	27.5	0.4

Quelles conclusions pour d'autres enquêtes ?

- Utilisation des paradonnées pour le redressement d'enquêtes sans base de sondage ?
- Les paradonnées peuvent elles remplacer le revenu (souvent lié aux variables d'intérêt lors d'un calage ou celui-ci est manquant ?

Modélisation ANOVA des déciles de revenus par les variables auxiliaires et les paradonnées

Variable	F	P
TH courriel	861	0,000
TH tel fixe	76	0,000
Echantillon	32	0,000
Lettre de refus	22	0,000
TH tel mob	15	0,000

Bilan :
→ très peu de paradonnées explicatives pour des enquêtes non tirées dans la TH
→ envoi d'une lettre de refus

Côté variables quantitatives :
→ agrégation refus et raccroches légèrement corrélée au revenu (rho=-0,08).
→ Le nombre de tentatives d'appel n'est pas corrélé au revenu

Conclusion

- Utiliser les numéros TH améliore beaucoup la couverture et le taux de réponse
- Camme est robuste : variables auxiliaires et paradonnées modifient peu les estimations des variables de solde (en fait, rien ne les affecte sensiblement)
- Peu de liens entre les paradonnées et les variables d'intérêt ou auxiliaires → leur utilité semble limitée pour la CNR

Références

- Couper, M.P. 1998. Measuring survey quality in a CASIC environment. Proceedings of the Survey Research Methods Section of the American Statistical Association, 41–49.
- Haziza, D., & Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics, 32*(1), 129-145. doi:10.1515/JOS-2016-0006
- Kreuter, F. (Ed.) (2013). *Improving surveys with paradata*. New York: John Wiley & Sons.
- Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T., Casas-Cordero, C., & Lemay, M. (2010). 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, 173*(2), 389-407.
- Little, R. J. A., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey methodology, 31*(2), 161–168.
- Olson, K. (2013). Paradata for Nonresponse Adjustment. *The Annals of the American Academy of Political and Social Science, 645*(1), 142-170. doi:10.1177/0002716212459475
- Wisner, J., Phillips, O., Baribeau, B., & Lévesque, A. (2009). *Using paradata to manage nonresponse in the Survey of Labour and Income Dynamics*. Paper presented at the Symposium 2008.