Redresser l'échantillon d'une enquête en ligne. Un exemple à partir de l'enquête Vico

Léo Joubert, Olivier Lê van Truoc, Pierre Mercklé, Benoît Tudoux

▶ **To cite this version:**

Léo Joubert, Olivier Lê van Truoc, Pierre Mercklé, Benoît Tudoux. Redresser l'échantillon d'une enquête en ligne. Un exemple à partir de l'enquête Vico. Bulletin de Méthodologie Sociologique / Bulletin of Sociological Methodology, 2023, 158 (1), pp.116-142. ⟨10.1177/07591063231160287⟩. ⟨halshs-04117897⟩

HAL Id: halshs-04117897

https://shs.hal.science/halshs-04117897v1

Submitted on 6 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-ND 4.0 - Attribution - No Derivative Works - International License

Redresser l'échantillon d'une enquête en ligne
Un exemple à partir de l'enquête Vico
Version Auteur

Léo Joubert
DYSOLAB, Université de Rouen, France

Olivier Lê Van Truoc
Institut d'études politiques de Grenoble, France

Pierre Mercklé
Université Grenoble Alpes, PACTE, Grenoble, France

Benoît Tudoux
CNRS, ISP, Nanterre, France

Abstract
Correcting the sample of an online survey. An example from the Vico survey. The Vico survey, as is often the case in online surveys, is characterized by a number of biases: respondents tend to belong to higher income groups and have received higher education than the general population, and women represent a much higher proportion of respondents than men. A technique for adjusting for these biases is proposed in this article, and its effectiveness is discussed.

Résumé
Comme c'est souvent le cas dans les enquêtes en ligne, et *a fortiori* dans celles reposant sur des échantillons « spontanés », celui de l'enquête Vico est caractérisé par un certain nombre de biais : les répondant·es sont beaucoup plus souvent issu·es des catégories favorisées et diplômées que l'ensemble de la population, et les femmes y sont également bien plus nombreuses que les hommes. Cet article propose une technique de redressement de ces biais, et en discute l'efficacité.

Keywords
Online survey, sample, bias, weighting, survey methodology

Mots-clés
Enquête en ligne, sondage, biais, redressement, méthodologie d'enquête

Introduction

Pour des raisons à la fois de coût et de difficultés de plus en plus grandes pour réaliser des enquêtes en face à face ou même joindre des répondant·es par téléphone, les enquêtes par questionnaire diffusé en ligne se sont multipliées au cours des deux dernières décennies. Les outils pour le faire se sont également multipliés, et ils sont désormais extrêmement faciles d'accès : il est facile d'élaborer puis de diffuser un questionnaire à l'aide de plateformes grand public comme Framaforms, ou d'outils professionnels comme le logiciel Lime Survey.

Ce mode de diffusion des enquêtes et de collecte des données présente de nombreux avantages, mais aussi un certain nombre d'inconvénients, au premier rang desquels des risques accrus de « biais d'échantillonnage » (Frippiat et Marquis, 2010). Dans cet article, nous présentons les biais rencontrés lors de la première vague de l'enquête par questionnaire sur la « vie en confinement » (Vico, avril-mai 2020)[1], et les procédures utilisées pour les « redresser ». Nous y défendons, ce faisant, une approche des questions de redressement qui est centrée sur les problèmes pratiques que pose cette technique du redressement. Nous serons parfois amenés à nous appuyer sur des résultats statistiques bien établis en matière de méthodes de redressement, et parfois nous élaborerons à partir d'une expérience acquise par la pratique du « métier de sociologue », plus à distance donc des seules préoccupations de la statistique théorique. Notre but est de donner des clés aux praticien.nes des enquêtes sociologiques pour « hésiter dans la bonne direction », c'est-à-dire pour articuler au cas par cas l'exigence de « représentativité » avec la production de raisonnements sociologiquement pertinents.

Avant d'entrer dans le détail, il convient de rappeler qu'un redressement peut être envisagé à deux conditions liminaires : d'une part, l'échantillon est volontairement[2] ou involontairement déformé par rapport à la population étudiée ; et d'autre part il existe des statistiques de cadrage fiables sur ces variables (Dussaix et Grosbras, 1993)[3]. Si ces deux conditions sont remplies, un redressement devient envisageable – sans pour autant être nécessairement pertinent.

Dans le but de présenter un retour d'expérience réflexif sur le redressement d'une enquête en ligne à partir de l'exemple de l'enquête Vico, nous suivrons un raisonnement en quatre temps. Nous présenterons d'abord le contexte et le mode de passation très spécifiques de cette enquête, menée sur le confinement pendant le confinement et par un collectif de chercheur·es eux-mêmes confiné·es. Puis, nous détaillerons la technique de redressement choisie, celle du « calage sur marges », ainsi que les enjeux de mesure de la qualité de la pondération produite. Ensuite, nous détaillerons les aspects les plus pratiques du redressement en présentant les choix finalement effectués. Enfin, nous proposerons une manière sociologique d'appréhender un problème statistique particulièrement complexe : l'utilisation ou non des pondérations[4] calculées par le redressement.

---

[1] La première vague de l'enquête Vico a été réalisée par questionnaire entre le 15 avril et le 10 mai 2020. Elle visait à recueillir des informations sur la situation des Français vis-à-vis du logement et du travail avant et pendant le confinement, ainsi que sur leurs activités et les évolutions de leurs relations personnelles pendant le confinement. Le questionnaire de l'enquête a été administré en ligne à un échantillon final de 16 224 personnes âgées de 18 ans et plus résidant en France pendant le confinement. Enquête « La vie en confinement – Vague 1 », 2020, Équipe Vico. Pour obtenir plus de détails sur le projet : https://vico.hypotheses.org.

[2] A titre d'exemple, nous pouvons citer les enquêtes Génération du Céreq, où des sur-représentations régionales ou de secteurs d'activités économiques sont volontairement mises en place afin d'avoir suffisamment d'observations statistiques sur ces catégories.

[3] Que ces statistiques de calage soient issues d'enquêtes externes, comme l'enquête Emploi 2018 dans le cas de VICO, ou d'une base de sondage construite pour l'enquête.

[4] La pondération d'un jeu de données renvoie en fait à deux principes qu'il faut bien distinguer : d'une part, un principe de correction de la structure de l'échantillon afin que celui-ci soit au final un modèle réduit de la population étudiée tout en gardant la taille de l'échantillon observé. On peut alors parler de redressement et de poids de coefficient de pondération normalisé. Et d'autre part, un principe d'extrapolation des résultats au niveau

Enquêter sur la « vie en confinement » quand on est soi-même confiné·e : d'accord, mais comment faire ?

La diffusion d'un questionnaire

Le mode d'administration de l'enquête Vico est en grande partie le résultat de la situation très particulière dans laquelle l'enquête a été réalisée. Nous avons imaginé et mis en œuvre cette enquête par questionnaire pendant le premier confinement du printemps 2020, afin de saisir « à chaud » les effets de la crise sanitaire et des restrictions de circulation sur les conditions de vie, de travail et les relations personnelles. Mais étant données les circonstances et l'urgence, nous n'avions ni les moyens financiers ni la possibilité matérielle de réaliser une enquête par questionnaire traditionnelle, issue d'un sondage aléatoire (échantillon probabiliste) ou d'un sondage empirique (méthodes des quotas où l'on cherche à reproduire les profils sociodémographiques observés dans la population de référence en bonne proportion). Il n'était en particulier pas possible, dans des délais aussi courts et avec des moyens relativement limités, de disposer d'un échantillon tiré au sort dans la totalité de la population, comme pour les enquêtes de la statistique publique.

C'est la raison pour laquelle nous avons choisi de diffuser le questionnaire en ligne sur un site internet créé pour cette occasion[5]. Quatre modes de diffusion de l'invitation à remplir le questionnaire ont ensuite été mis en œuvre simultanément ou successivement. Le premier a simplement consisté à prendre appui sur les différents réseaux personnels des chercheur·es membres de l'équipe pour obtenir, par effet de « boule de neige », la constitution d'un premier sous-groupe. La règle consistait alors à relayer le questionnaire à des connaissances ayant des caractéristiques (de milieu social, d'âge, de conditions de confinement) différentes des nôtres qui sont celles de personnes appartenant toutes à la fonction publique, plutôt diplômées, et en général économiquement et socialement privilégiées. Les réseaux familiaux et amicaux des étudiant·es des facultés françaises de sociologie ont ensuite été mobilisés, avec l'aide active de plus d'une soixantaine de collègues à travers toute la France. En l'occurrence, nous avions fait l'hypothèse que l'origine sociale plutôt populaire des étudiant·es de sociologie nous permettrait de toucher des milieux sociaux modestes.

On le voit, il ne s'agit pas à ce stade de s'appuyer sur une stratégie de passation extrêmement précise, mais plutôt de « faire feu de tout bois ». Au départ, l'objectif était d'obtenir un maximum de réponses au questionnaire. Il aurait été fort possible que parmi les étudiant·es « globalement modestes » que nous visions, seul un sous-échantillon issu de milieux aisés fasse effectivement passer les questionnaires. Nous aurions alors dû adapter notre démarche.

Dans un troisième registre, le lien vers le questionnaire a été diffusé et relayé en ligne (Bès et al., 2020), sur des pages personnelles ou des pages de groupes Facebook choisis de façon que puissent être ainsi atteintes des personnes dont il est habituellement plus difficile d'obtenir les réponses à de telles enquêtes (comme par exemple des pages Facebook de professions ou de branches rassemblant des ouvriers et des employés de sexe masculin, dans les secteurs de l'industrie, du bâtiment et des travaux publics…). Ce mode de diffusion a été quantitativement très efficace : un quart (25,6 %) des enquêté·es déclarent avoir eu connaissance du questionnaire par une information trouvée sur Internet (Tableau 1). Enfin, nous avons également obtenu de nombreuses réponses (plus de 30 % du total) en sollicitant les rédactions des quotidiens régionaux. L'idée, là encore, était que leur lectorat nous permettrait

de la population étudiée (estimation de la taille d'une population ou d'un stock). On parlera alors de poids ou de coefficient d'extrapolation. Dans la littérature ou la documentation des données d'enquête, cette distinction n'est pas toujours évidente.

[5] https://enqueteconfinement.wixsite.com/site.

de toucher un grand nombre de gens et d'atteindre les populations les plus diversifiées possibles, tant géographiquement que socialement. Au total, plus d'une vingtaine de titres ont répondu positivement à notre requête (voir annexe), ainsi que le site Actu.fr qui regroupe, pour la France entière, un ensemble de 93 journaux locaux (dont 15 gratuits et 77 hebdomadaires).

Tableau 1. Comment avez-vous eu connaissance de ce questionnaire ?

	Effectif	
Par un article de la presse régionale ou locale (imprimé ou numérique)	**5 175**	**1,9**
Par une information trouvée sur internet (site internet, Facebook…)	**4 157**	**5,6**
Par un article de la presse nationale (imprimé ou numérique)	**1 512**	**,3**
Par un(e) ami(e) proche	**1 357**	**,4**
Par un(e) collègue de travail, un(e) membre d'une association ou d'une organisation que je fréquente	**989**	**,1**
Par un(e) enseignant(e) ou un(e) chercheur(e) qui participe à l'enquête	**865**	**,3**
Par un message électronique (SMS, email, Messenger, WhatsApp…)	**852**	**,3**
Par un autre membre de ma famille	**718**	**,4**
Par une connaissance plus éloignée	**587**	**,6**
Par mon fils ou ma fille	**338**	**,1**
Par mon conjoint ou ma conjointe	**212**	**,3**
Par mon père ou ma mère	**117**	**,7**
Par une personne que je ne connaissais pas auparavant	**83**	**,5**
Par un(e) voisin(e)	**77**	**,5**

Source : Enquête Vico, avril-mai 2020.
Champ : Personnes résidant habituellement en France et âgées de 18 ans ou plus (N = 16 224).
Lecture : 31,9 % des répondant·es ont eu connaissance du questionnaire par un article de la presse régionale.
Note : la somme des pourcentages est supérieure à 100%, les répondant·es ayant la possibilité de cocher plusieurs cases.

Qui a répondu à l'enquête ?

En termes strictement quantitatifs, la combinaison de ces modes de diffusion a été une vraie réussite puisque plus de 16 000 personnes ont rempli l'intégralité du questionnaire. La taille de l'échantillon ainsi obtenu est donc particulièrement élevée. C'est beaucoup plus en effet que pour les sondages d'opinion dont les médias sont habituellement friands, mais qui ne portent le plus souvent que sur environ un millier de personnes. Et c'est même significativement plus que dans les enquêtes habituelles de la recherche publique en sciences sociales, dont les échantillons comportent généralement moins de 10 000 personnes. La grande taille de l'échantillon de l'enquête Vico présente des avantages certains : outre qu'elle permet de documenter un plus grand nombre de « vies en confinement » individuelles, elle offre également la possibilité d'analyser beaucoup plus finement et précisément les conditions de confinement et les évolutions des relations de groupes sociaux ou de profils minoritaires, dont les effectifs sont trop faibles pour être étudiés statistiquement. Les données récoltées ont ainsi permis de produire des analyses spécifiques sur les étudiant.es et les soignant.es (Mariot, Mercklé, Perdocin (dir), 2021).

La médaille a bien sûr son revers. Les modalités de diffusion en ligne de l'enquête ont permis de recueillir un très grand nombre de réponses, mais elles ont aussi produit, sans surprise, ce qu'on appelle des « biais d'échantillonnage ». Si ceux-ci sont classiques dans ce genre d'enquêtes, ils apparaissent tout de même relativement forts dans Vico. Cela veut dire que sur un certain nombre de critères, les caractéristiques des personnes qui ont répondu à l'enquête Vico diffèrent significativement de celles de l'ensemble de la population à laquelle elles appartiennent, et dont l'enquête ambitionnait de rendre compte. Ces écarts et décalages par rapport à la population nationale concernent principalement quatre caractéristiques : le sexe, l'âge, le niveau de diplôme et le milieu social (Tableau 2). Ainsi dans notre enquête, il aurait dû y avoir, à l'image de la population de référence (celle dont l'enquête ambitionne de parler), 51 % de femmes et 49 % d'hommes ; mais dans notre échantillon, 73 % des répondants sont… des répondantes. De même, alors qu'environ un tiers des gens sont titulaires d'un diplôme universitaire en France, ils sont plus de deux fois plus nombreux (69 %) à être dans ce cas parmi nos enquêté·es, ce qui contribue à déformer l'échantillon en faveur des plus diplômé·es et des catégories professionnelles les plus qualifiées. À l'inverse, les populations ouvrières sont sous-représentées dans notre étude. Moins marquées, on note également des différences concernant la répartition géographique de l'échantillon, avec des distorsions sur certaines régions et une surreprésentation des zones urbaines de province au détriment des habitant·es des communes rurales et de l'agglomération parisienne.

Tableau 2. Distribution de l'échantillon de l'enquête Vico et de la population de la France

	Echantillon brut		**Echantillon apuré**		**Echantillon redressé**		**France**
	n	**%**	**n**	**%**	**n**	**%**	**%**
Total	**16 224**	**100,0**	**15 083**	**100**	**15 083**	**100,0**	**100,0**
Sexe							
Un homme	4 385	27,0	4 078	27	7 322	48,5	51,3
Une femme	11 780	72,6	11 005	73	7 761	51,5	48,7
Autre définition de genre	59	0,4	0	0	0	0,0	0,0
Âge							
De 18 à 24 ans	1 796	11,1	1 302	8,6	1 724,9	11,4	11,5
De 25 à 29 ans	1 576	9,7	1 496	9,9	1 285,5	8,5	8,1
De 30 à 34 ans	1 598	9,8	1 561	10,3	1 262,0	8,4	8,8
De 35 à 39 ans	1 768	10,9	1 746	11,6	1 425,0	9,4	9,1
De 40 à 44 ans	1 663	10,3	1 640	10,9	1 345,4	8,9	8,9
De 45 à 49 ans	1 741	10,7	1 717	11,4	1 432,3	9,5	9,8
De 50 à 54 ans	1 490	9,2	1 473	9,8	1 517,5	10,1	9,7
De 55 à 59 ans	1 280	7,9	1 267	8,4	1 433,0	9,5	9,4
De 60 à 64 ans	1 205	7,4	1 196	7,9	1 277,7	8,5	8,8
De 65 à 69 ans	1 003	6,2	999	6,6	1 402,1	9,3	8,6
De 70 à 74 ans	690	4,3	681	4,5	969,9	6,4	7,2
75 ans et plus	409	2,5			hors-champ		
NA	5	0,0	5	0,0	7,6	0,1	0,0

Tableau 2. *Suite*

	Echantillon brut		Echantillon apuré		Echantillon redressé		France
	n	%	n	%	n	%	%
Total	**16 224**	**100,0**	**15 083**	**100**	**15 083**	**100,0**	**100,0**
Taille de l'unité urbaine							
Commune rurale	2 515	15,5	2 431	16,1	3 328,2	22,1	23,4
De 2000 à 4999 hab.	688	4,2	659	4,4	551,1	3,7	7,4
De 5 000 à 9 999 hab.	879	5,4	837	5,5	647,4	4,3	4,8
De 10 000 à 19 999 hab.	786	4,8	743	4,9	673,6	4,5	4,9
De 20 000 à 49 999 hab.	995	6,1	943	6,3	769,4	5,1	5,2
De 50 000 à 99 999 hab.	1 092	6,7	1 044	6,9	954,9	6,3	6,4
De 100 000 à 199 999 hab.	847	5,2	761	5,0	578,3	3,8	5,0
De 200 000 à 1 999 999 hab.	5 962	36,7	5 490	36,4	4 647,8	30,8	26,1
Unité urbaine de Paris	1 815	11,2	1 580	10,5	2 277,6	15,1	16,9
PCS							
Agriculteurs, agricultrices	53	0,3	50	0,3	60,3	0,4	0,9
Artisan(e)s, commerçant(e)s, chef(fe)s d'entreprise	515	3,2	506	3,4	636,7	4,2	4,3
Cadres et professions intellectuelles supérieures	4 146	25,6	4 049	26,8	1 719,6	11,4	11,4
Professions intermédiaires	3 075	19,0	3 043	20,2	2 496,9	16,6	16,1
Employé(e)s	2 946	18,2	2 915	19,3	3 585,9	23,8	17,7
Ouvrier(e)s	472	2,9	467	3,1	981,2	6,5	13,7
Retraité(e)s	2 641	16,3	2 233	14,8	2 927,6	19,4	19,1
Élèves, étudiant(e)s	1 493	9,2	957	6,3	1 489,3	9,9	5,3
Autres inactif(ve)s	706	4,4	690	4,6	1 016,7	6,7	11,5
NA	177	1,1	173	1,1	168,9	1,1	0,0
Situation conjugale							
En couple	11 793	72,7	11 077	73,4	10 722,4	71,1	62,6
Célibataire	4 362	26,9	3 945	26,2	4 285,6	28,4	37,4
NA	69	0,4	61	0,4	75,0	0,5	0,0
Niveau de diplôme							
Supérieur au baccalauréat	11 158	68,8	10 520	69,7	9 052,7	60,0	33,0
Baccalauréat	2 999	18,5	2 629	17,4	3 292,9	21,8	20,6
Inférieur au baccalauréat	1 994	12,3	1 872	12,4	2 657,4	17,6	46,5
NA	73	0,4	62	0,4	80,0	0,5	0,0

Source : Enquête Vico, avril-mai 2020 ; Enquête Emploi en continu (version FPR) - 2018, INSEE (producteur), ADISP (diffuseur).
Champ : Population résidant habituellement en France métropolitaine et âgée de 18 à 74 ans.
Lecture : 27,0 % des répondant·es de l'enquête Vico sont des hommes, contre 51,3 % des personnes résidant habituellement en France. La colonne "échantillon apuré" correspond aux réponses retenues pour effectuer le redressement. Les critères de sélection sont précisés dans la section le redressement en pratique.

Comment expliquer cette surreprésentation des diplômé·es et des catégories sociales supérieures, mais aussi des jeunes et des femmes parmi celles et ceux qui ont bien voulu répondre ? Au premier abord, l'ampleur de la surreprésentation des femmes peut sembler la plus surprenante. Mais elle ne l'est pas tant que ça, car on sait que de façon générale les femmes sont plus facilement enclines à répondre aux enquêtes que les hommes. Le biais scolaire évoqué précédemment a pu y contribuer également, dans la mesure où les plus diplômés sont aussi plus souvent des femmes. Enfin, et peut-être surtout, il faut faire l'hypothèse que c'est le sujet même de l'enquête qui est à l'origine de cet écart. L'enquête porte en effet en bonne part sur la sociabilité et les relations personnelles, dont la gestion domestique incombe souvent aux femmes, ce qui fait qu'à l'intérieur de chaque ménage ce sont elles qui, deux fois plus souvent que leur conjoint, ont répondu à l'enquête. On notera du reste que les auteurs d'une enquête en ligne assez similaire, puisque portant sur les réseaux personnels à San Francisco entre 2016 et 2018, ont observé un *sex ratio* pratiquement identique (Fischer et Bayham, 2019). Cela suggère que ce que nous observons dans notre étude constitue peut-être le résultat d'un effet général de ce dispositif d'enquête et du type de questions posées, portant sur la sociabilité et les relations sociales.

Plus généralement, les biais d'échantillonnage les plus importants observés dans l'enquête Vico tiennent sans doute largement aux particularités du dispositif de recueil des données. D'abord, l'échantillon de l'enquête est « spontané » : sa construction ne repose pas sur une base de sondage dans laquelle des individus sont tirés aléatoirement, et peuvent être relancés s'ils ne répondent pas à l'enquête. La réponse à l'enquête suppose au contraire d'une part d'avoir été informé·e de son existence, et d'avoir décidé volontairement de le faire, deux sources potentielles de biais. Ensuite, il s'agit d'un questionnaire sociologique dont le format, d'allure administrative ou scolaire, peut apparaître trop aride voire rébarbatif aux moins diplômé·es. Sa taille a sans doute accentué encore cette tendance à l'abandon en cours de route parmi les moins disposé·es à l'exercice : le remplir entièrement a pris 27 minutes en moyenne aux enquêté·es, et presque trois quarts d'heure à plus de 10 % d'entre eux. Le questionnaire est par ailleurs auto-administré, ce qui laisse toute latitude aux gens qui le commencent de s'arrêter dès qu'ils s'ennuient, alors que dans les enquêtes en face-à-face ou téléphoniques ils ou elles se sentent lié·es par une sorte de contrat moral avec l'enquêteur·trice qui les encourage à aller au terme de l'exercice.

Également, ce questionnaire est rempli en ligne, ce qui suppose de disposer d'un accès à internet et d'une forme de familiarité avec le numérique, plus rares chez les plus de 70 ans. Il n'est donc pas étonnant de constater que ceux et celles-ci ont moins souvent répondu à l'enquête que les autres : on en trouve seulement 7 % dans notre échantillon, alors qu'ils et elles représentent 19 % de la population totale de la France âgée de 18 ans ou plus. La sous-représentation des personnes âgées se concentre par ailleurs principalement sur les tranches d'âges les plus élevées, pour lesquelles nous comptons très peu de répondant·es. Nous avons donc choisi de restreindre la population de référence (ou univers) de l'enquête aux personnes âgées de 18 à 74 ans. La moindre inclusion des plus âgé·es dans l'échantillon participe aussi à la sous-représentation des personnes peu diplômées.

Enfin, les écarts constatés sur la répartition géographique sont certainement liés au mode de diffusion de l'enquête (méthode boule de neige initiée depuis des centres universitaires, par définition urbains, et relais dans la presse quotidienne régionale).

Au-delà de l'enquête VICO, les remarques que nous venons de faire posent la question plus générale de la représentativité en sciences sociales. La notion de représentativité peut être définie comme une exigence de couverture des principales caractéristiques sociodémographiques de la population étudiée. Cette notion polysémique n'est pas toujours acceptée chez les statisticien·nes d'enquête (Selz, 2013). En effet, certain·es statisticien·nes considèrent qu'un échantillon représentatif peut être défini comme un « échantillon non

biaisé », issu d'un plan de sondage aléatoire (Lejeune, 2021), tandis que d'autres, comme Olivier Sautory, estiment qu'un échantillon ne peut jamais être représentatif en soi, mais qu'il peut seulement l'être des variables à estimer (Gerville-Réache, Couallier et Paris, 2011). Une position intermédiaire pourrait être de considérer qu'un échantillon est représentatif s'il permet d'inférer avec fiabilité les variables d'intérêt de l'enquête.

Visant à être « représentative de la population », selon la formule (parfois trop vite) consacrée, les responsables de l'enquête VICO pouvaient s'appuyer sur les chiffres édités par l'INSEE dans de nombreuses enquêtes pour disposer des « marges » de cette population. C'est d'ailleurs grâce à ces marges que nous avons pu mesurer les biais d'échantillonnage inventoriés ci-dessus et conclure à la nécessité d'un redressement.

En cela, VICO montre combien la disponibilité des marges – donc possibilité théorique de définir un plan de sondage – n'avance en rien quand il s'agit de surmonter la difficulté fondamentale de n'importe quelle enquête par questionnaire : comment interroger le plus d'individus pertinents possible ? C'est bien pour combler cet écart entre plan de sondage théorique et contraintes pratiques de passation que le redressement a vocation à faire partie du métier de sociologue. Il faut cependant être au clair sur ce que cela implique, à savoir que la « population de référence » à laquelle il est souvent fait allusion dans le cas d'une enquête par questionnaire *redressée* est toujours, par définition, une population *ex post* plutôt qu'une population *ex ante*. Cela vaut même si, bien sûr, la question de la population de référence est souvent tranchée en amont, ne serait-ce que pour des raisons tenant à l'organisation de projets de recherche d'une ampleur organisationnelle comme celle de VICO.

Si, dans le cas d'un sondage proportionnel, c'est-à-dire sans sur- ou sous-représentation volontaire, les coefficients de pondération font correspondre la fréquence d'un profil avec sa probabilité d'inclusion avant passation du questionnaire, alors la population *ex post* est strictement identique à son équivalent *ex ante*. Lorsque cette correspondance est imparfaite, on pourra plus prudemment parler de correction du biais de non-réponse pour signifier que seules quelques-unes des contraintes pratiques de passation ont été surmontées. En pratique, le calage sur marge aboutit fréquemment à une correspondance parfaite, *mais uniquement par rapport aux variables qui ont été sélectionnées pour calculer les poids*.

Sur d'autres variables qui n'ont pas pu être incluses, le redressement n'améliore pas significativement la correspondance entre les marges de l'échantillon et les marges de la population de référence. Dans de rares cas, notamment lorsque l'échantillon présente de graves biais de construction, le redressement peut même dégrader la représentativité sur ces variables connexes. Par exemple, si aucun·e des retraité·es interrogé·es n'est un·e ancien.ne cadre, on peut raisonnablement penser que les retraité·es de l'échantillon auront un revenu et un niveau de diplôme inférieur à ceux de la population de référence. Donc, un redressement admettant comme variable de calage le revenu et le niveau de diplôme dégraderait la représentativité au regard des PCS.

Voilà pourquoi la frontière est ténue entre représentativité et correction du biais de non-réponse : même une représentativité parfaite en fonction de certains critères peut n'être qu'une correction modeste ou, plus rarement, une aggravation du biais de non-réponse sous d'autres rapports. Ces distinctions statistiques sont cruciales pour le/la sociologue, car elles incitent à déconstruire la notion de représentativité. Cette dernière n'est pas une caractéristique que possède, ou non, un échantillon, mais plutôt un horizon théorique à questionner avant, pendant et après l'analyse. L'analyse de données d'échantillon ne peut se faire qu'en tenant compte des éventuels biais de couverture et de non réponse, deux biais qui éloignent l'échantillon de sa population de référence Nous verrons plus bas que le calage sur marges auquel nous avons procédé redresse de façon excellente la plupart des variables sélectionnées pour le calage (en particulier l'âge, le sexe et la catégorie socio-professionnelle), qu'il a également un impact sur

le diplôme, mais incomplet, et qu'il se révèle insuffisamment opérant pour d'autres variables importantes.

Techniques, enjeux et qualité du redressement

Comment redresser une enquête en ligne ?

Dans l'ensemble donc, l'échantillon de l'enquête Vico est marqué par un certain nombre de biais d'échantillonnage. Dans cet article, on traitera uniquement des biais visibles sur des critères connus au niveau de la population et mesurés dans l'échantillon. D'autres biais de sélection peuvent exister sans que l'on puisse les redresser. Lorsqu'ils sont « redressables », les biais d'échantillonnage peuvent l'être en utilisant les techniques habituelles de « pondération ». Deux étapes doivent être distinguées ici. La première, techniquement assez simple mais sociologiquement déjà engageante, consiste à traiter les valeurs manquantes des variables de calage (Neiter et Buisson, 2010), dans la mesure où une variable ne peut être utilisée pour caler un échantillon que si elle est renseignée pour tous les individus. Deux solutions sont possibles ici : le retrait des non-répondant·es, ou l'imputation des valeurs manquantes. Dans le cas de Vico, nous verrons plus loin que nous avons opté pour la seconde méthode.

La seconde étape est le calcul des « coefficients de pondération » mesurant le « poids » de chaque individu compte tenu de son profil. Cette étape utilise des informations sur l'univers de l'enquête (post-stratification) afin de réduire les distorsions dues à des erreurs de non-réponse (biais de sélection, biais de couverture). Les « coefficients de pondération » sont affectés à chaque répondant·e, en s'appuyant sur une comparaison précise des caractéristiques de l'échantillon avec celles de la population de la France.

Pour calculer ces coefficients, nous avons utilisé la technique classique du « calage sur marges » : la distribution de la population de référence (celle de la totalité des personnes résidant en France âgées de 18 ans à 74 ans) selon un certain nombre de caractéristiques étant connue, on calcule des coefficients de pondération pour chaque individu de l'échantillon, de telle façon que leur application permet de minimiser l'écart entre la distribution de l'échantillon et celle de la population de référence. L'algorithme de redressement va chercher à atteindre les objectifs fixés en faisant en sorte que le poids de redressement pour chaque individu reste le plus proche possible de son poids initial (donc le plus proche possible de 1, du moins dans le cadre d'un sondage aléatoire simple où les observations ont la même probabilité d'inclusion).

Formellement, le calage sur marges s'écrit de la manière suivante (Deville, Särndal et Sautory, 1993) : on cherche W_i minimisant

$$\left\{\sum_{i\in S} d_i G\left(\frac{W_i}{d_i}\right) s.c. \sum_{i\in S} W_i X_i = T_X\right\}$$

Où S représente l'échantillon étudié, W_i le poids de redressement de l'individu i, d_i son poids de sondage, X_i une variable de référence et T_X le vecteur de ses marges. La fonction G est une fonction de distance entre les poids de redressement et les poids de sondage. À travers sa minimisation, la résolution de cette équation vise à limiter la déformation de l'échantillon. Une fonction G adéquate admet deux propriétés :

1. $G(1) = 0$
2. G est positive et convexe, ce qui entraîne que $G\left(\frac{W_i}{d_i}\right)$ est d'autant plus élevée que le rapport $\frac{W_i}{d_i}$ est éloigné de 1. Autrement dit, G augmente avec la déformation de l'échantillon.

Plusieurs fonctions G sont bien sûr possibles et ont été introduites par les praticien·nes du calage sur marge au fil du temps. Chaque fonction va se distinguer en fonction de son mérite

respectif : certaines permettent de fixer des poids minimaux et maximaux à l'intérieur desquels la solution optimale devra être trouvée, tandis que d'autres optimisent la distribution des poids au détriment d'autres facteurs… On se tournera vers des manuels de référence à jour pour choisir la bonne fonction.

Comment évaluer la qualité d'un redressement ?

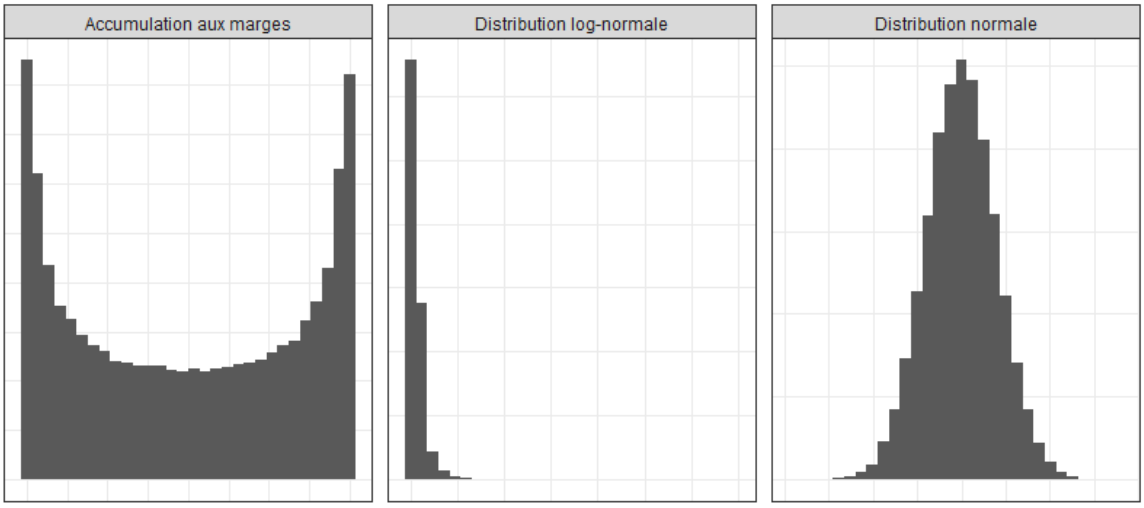
Avant d'entrer dans l'exemple pratique offert par le redressement de l'échantillon de l'enquête Vico, nous proposons trois critères opérationnels permettant de réfléchir, dans le cas de n'importe quelles données d'enquête, sur cette tension entre représentativité et non-réponse.

Premier critère : la « convergence » de l'algorithme

Le premier critère, technique, semble trivial : il faut s'assurer que la procédure de calage sur marges a « convergé », autrement dit qu'elle est bien parvenue à produire un jeu de coefficients de pondération respectant les contraintes qui auront été indiquées. On trouve deux types de contraintes, en plus de la contrainte strictement mathématique de minimisation de l'écart entre les marges de l'échantillon et les marges de la population. D'une part, le choix des variables peut exercer une contrainte, dans la mesure où il peut se révéler trop ambitieux compte tenu des biais d'échantillonnage de l'enquête. Le praticien se lance alors dans une série de transactions, retirant les variables une par une, en ajoutant d'autres, créant des variables uniques pour combiner deux variables existantes (par exemple, une variable « âge et sexe » avec des modalités comme « homme de 18-25 ans »). Ces multiples essais, qui ont pour but final la convergence de l'algorithme, débutent un dialogue entre « raison statistique » et « raison sociologique » (Martin, 1999). Nous renvoyons ici à ce qui a été dit plus haut sur la question de la population de référence par définition *ex post* dans les enquêtes redressées : la non-convergence de l'algorithme peut être interprétée comme un signal indiquant que la raison sociologique demande des opérations mathématiquement impossibles. Cela peut être le signe d'un problème survenu au moment de l'enquête de terrain, qui aurait occasionné des biais de construction trop importants. C'est ici un impératif de faisabilité de l'enquête que l'on doit à nouveau affronter, alors qu'il est habituellement présent bien en amont, au moment de la détermination de la question de recherche (Quivy et Van Campenhoudt, 2018).

D'autre part, il y a la contrainte des bornes de calage, qui peuvent être exigeantes au vu de la qualité de l'échantillon avant redressement. La spécification des bornes de calage permet de contrôler la dispersion des poids en fixant un poids minimal et un poids maximal. Si les bornes peuvent être une solution à un redressement dont la dispersion « dérape », celle-ci n'est pas parfaite. En effet, des bornes trop exigeantes empêcheront l'algorithme de converger ou, dans le cas de très fortes sous ou sur-représentations, produiront un phénomène d'accumulation aux marges (cf. Figure 1 ci-après).

Figure 1. Distributions possibles des coefficients de redressement



Deuxième critère : l'amélioration de la représentativité de variables non incluses dans la liste des variables de calage

Le deuxième critère, essentiel, est celui d'une amélioration incidente de la représentativité sur d'autres variables : on mesure la qualité d'un redressement à sa capacité à rapprocher les caractéristiques de l'échantillon de celles de la population de référence, non seulement bien sûr pour celles de ces caractéristiques utilisées pour le calage, mais également pour d'autres caractéristiques. On sera donc attentif aux effets du redressement sur les principales caractéristiques sociodémographiques qui n'auront pas été introduites dans le calage. Par exemple, si on ne redresse que sur le sexe et sur l'âge, alors on examinera les effets du redressement sur les distributions des régions de résidence, des catégories socioprofessionnelles, des niveaux de diplômes…

On sera donc attentif aux effets du redressement sur d'autres variables d'intérêt dont on connaît la distribution dans la population de référence. Par exemple, le premier tour des élections municipales de 2020, organisé le dimanche 15 mars 2020, soit la veille du début du confinement, a suscité un taux record d'abstention de 55,3 % à l'échelle nationale. Dans l'enquête Vico, le taux d'abstention déclaré par les répondants était de seulement 43,5 %. On regardera donc dans quelle mesure le redressement permet « d'augmenter » le taux d'abstention « prédit » par l'enquête.

Ce deuxième critère illustre la complexité des liens entre représentativité et correction du biais de non-réponse. Le redressement peut « marcher », au sens où les marges de l'échantillon et les marges de la population de référence peuvent coïncider, tout en ayant un « effet pervers » dégradant – ou n'améliorant pas – d'autres marges qui n'ont pas pu être incluses. Il faut cependant préciser que ce second critère n'est généralement pertinent que dans des situations où la passation du questionnaire a donné lieu à des distorsions très fortes. Dans les cas où le plan de sondage – quand il en existe un ! – est adéquatement suivi, il est peu fréquent que le redressement de l'échantillon augmente l'écart entre les marges de l'échantillon et de la population, même pour des variables non incluses dans le calage.

Cependant, nous ne saurions trop conseiller, dans l'évaluation de la qualité d'un redressement, procéder à cette vérification, au titre d'une précaution. Ces redressements « incidents » au calage peuvent sans doute être considérés comme un critère de validation externe. Par exemple, si un calage réalisé à l'aide de marges portant sur l'âge dégrade l'échantillon au regard du vote à un scrutin, alors les individus auquel le redressement a accordé un fort poids ont peut-être voté de façon atypique par rapport à leur classe d'âge. On voit ainsi

comme la réalisation du redressement est une étape offrant de multiples occasions de faire une myriade de vérifications permettant de mieux connaître son échantillon.

Troisième critère : la dispersion des coefficients de pondération

Dans l'échantillon redressé, chaque répondant·e va être affecté·e d'un poids différent, en fonction de la fréquence relative de ses caractéristiques dans l'échantillon. Les répondant·es dont les caractéristiques sont surreprésentées vont avoir un coefficient plus faible, ce qui contribue en quelque sorte à diminuer la taille du sous-échantillon qu'ils constituent. Réciproquement, certain·es répondant·es vont avoir un coefficient plus élevé parce qu'ils et elles sont sous-représenté·es.

Il faut être particulièrement vigilant sur ce point. En effet, il n'y a pas de garantie que les répondant·es disposant d'un coefficient fort « représentent » correctement les individus de mêmes caractéristiques appartenant à la population de référence : on considère au contraire généralement que les biais d'échantillonnage qui ont produit leur sous-représentation font que ceux ayant répondu malgré tout ont des caractéristiques particulières.

Dans l'enquête Vico par exemple, la diffusion du questionnaire par internet, qui a empêché des individus âgés de répondre, fait que celles et ceux qui ont tout de même répondu ont probablement une maîtrise d'internet telle qu'ils sont différents de la population de leur âge, par exemple en termes de niveau de diplôme, de qualification ou de catégorie socioprofessionnelle.

Comment se prémunir de ce biais ? S'il est difficile d'élaborer une règle générale ici, on pourra considérer que ce biais est d'autant plus fort lorsque les sous-populations les moins bien représentées dans l'échantillon ont un poids fort. Cela revient à dire que le biais est d'autant plus fort que les poids sont dispersés. Les redressements aboutissant à une dispersion forte, dans le cas d'échantillons proportionnels, doivent donc être évalués avec une grande prudence.

Il y a plusieurs façons d'apprécier la dispersion des coefficients de pondération. On peut mobiliser des mesures habituelles de dispersion : le rapport entre extrêmes (autrement dit entre le coefficient le plus élevé et le coefficient le plus faible) donne une première indication. On va être attentif à maintenir les poids dans un intervalle de plus ou moins 4 fois le poids moyen, donc dans le cas d'un sondage proportionnel, à maintenir l'amplitude entre un coefficient minimum autour de 0,25[6] et un coefficient maximum autour de 4, le point essentiel est de veiller à ce qu'une petite proportion d'individus ne représente pas une part trop importante du poids total, pesant trop lourdement dans les estimations et conduisant à augmenter les marges d'erreur (Lejeune, 2021).

On sera également attentif à l'allure générale de la distribution : on veillera en particulier à ce que son logarithme approche une loi normale ou au moins à ce qu'il n'y ait pas de concentration trop forte des coefficients aux extrêmes de la distribution (voir à nouveau Figure 1).

Enfin, on pourra se reporter à un indicateur synthétique qui mesure « l'efficacité » du redressement en normalisant la variance des poids. Exprimé en pourcentage, la formule en est la suivante, avec n la taille de l'échantillon et x_i les coefficients de pondération :

$$efficacité = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2} \text{ x } 100$$

L'efficacité, qui est donc le rapport entre le carré de la somme des coefficients et la somme des carrés des coefficients, divisé par le nombre de coefficients et exprimé en

[6] Dans toute la suite, on utilise des coefficients de pondération calculés de telle sorte que leur moyenne est de 1, donc obtenus en divisant les coefficients bruts par leur moyenne. De cette façon, la taille totale de l'échantillon reste inchangée, que l'on utilise les effectifs bruts ou les effectifs redressés, ce qui évite de fausser les tests de significativité.

pourcentage, mesure l'écart global entre les poids de redressement et 1. Plus l'efficacité est proche de 100 %, moindre est la déformation de l'échantillon. Il n'y a pas de valeur de référence pour l'efficacité, même si généralement on conseille justement de rester au-dessus de ce seuil de 70 % (Guilbert, 2001).

Le redressement en pratique

« Apurer » les données

Avant de procéder au calcul des coefficients de pondération, il est possible de réaliser un certain nombre d'opérations d'apurement des données qui peuvent améliorer significativement la qualité du redressement. L'apurement peut comporter deux opérations distinctes : la correction d'incohérences de réponses manifestes ; et le retrait d'un certain nombre d'individus de l'échantillon.

Dans le cas de l'enquête Vico, les premières analyses des données ont permis d'identifier quelques dizaines de réponses à certaines questions qui ne sont pas réalistes et qui ont été assimilées à des erreurs de saisie (comme par exemple des logements de 3 500 pièces, ou des répondants ayant 54 enfants). La suppression de ces réponses irréalistes (effectuée en les remplaçant par des valeurs manquantes), en particulier quand elle concerne des variables utilisées pour le calage, va avoir un effet direct sur la qualité du redressement, en diminuant la dispersion des coefficients. L'apurement des données va également nettement favoriser les chances que l'algorithme converge, or nous avons vu plus haut qu'il s'agissait d'une première étape parfois critique.

La deuxième opération d'apurement consiste à retirer du redressement un certain nombre d'individus ou de groupes d'individus, soit parce que dans un très petit nombre de cas individuels leurs réponses sont aberrantes ou fantaisistes, soit parce que collectivement ils biaisent fortement l'échantillon ou ils appartiennent à des catégories trop peu nombreuses pour être correctement représentées par l'enquête. Dans ce dernier cas, on n'efface pas ces individus de l'échantillon, mais on se contente de les « retirer » du redressement en leur attribuant d'office un coefficient de pondération nul. La variable de pondération devient ainsi une variable de filtre qui permet de travailler sur un échantillon redressable ou non. Dans le cas de l'enquête Vico, on a ainsi retiré du redressement : les 59 individus ayant déclaré un genre « autre » ; les 27 enseignant·es-chercheur·es en sociologie ayant répondu à l'enquête pour tester le questionnaire ; les 517 étudiant·es ayant rempli le questionnaire à la demande de ces enseignant·es ; les 90 femmes qui ont répondu à l'enquête à la demande de leur mari, pour corriger le biais de genre et réduire le nombre de répondants par ménage ; les 53 ultra-marin·es ; et enfin, les 414 personnes âgées de 75 ans et plus.

Trop peu nombreuses par rapport à la proportion dans la population française, les personnes âgées de 75 ans et plus présentes dans l'échantillon possèdent en outre des caractéristiques qui les différencient nettement des individus de leur classe d'âge (principalement en raison de l'administration du questionnaire par internet), ce qui fait qu'elles ne peuvent pas les représenter de façon acceptable, même au prix d'un redressement. Il en résulte au total le retrait de 1 146 individus, soit 7,1 % de l'échantillon. L'échantillon final après les retraits et suppressions de répondants est donc au total constitué de 15 078 individus.

Choisir et préparer les variables de calage

Il n'existe pas de règle absolue pour sélectionner les variables de calage, et c'est aux enquêteurs et enquêtrices de les choisir judicieusement, en fonction du sujet de l'étude[7] et des biais d'échantillonnage qu'ils peuvent constater ou auxquels ils peuvent s'attendre. En théorie toutefois, il faut choisir les variables les plus discriminantes pour le sujet de l'enquête. Plus les variables de calage seront liées aux variables d'intérêt, et plus le redressement en améliorera la précision (Lejeune, 2021).

Le plus souvent, on utilise une combinaison de plusieurs critères, pouvant parfois être croisés entre eux (par exemple, le sexe et l'âge). A cela s'ajoute une précaution sur le nombre des modalités utilisées au titre des marges. Un trop grand nombre de modalités aura pour effet d'augmenter exagérément la contrainte sur le redressement et donc la dispersion des poids. Une règle empirique (Lejeune, 2021) est que la taille de l'échantillon divisée par le nombre de modalités reste supérieure à 30[8].

En plus du contrôle de l'inflation du nombre de modalités, deux critères statistiques peuvent être utilisés pour décider, au cas par cas, d'ajouter une variable ou de ne pas le faire :

- Le lien entre la variable et les variables d'intérêts. Dans le cas où ce lien ne peut pas être déterminé à l'avance, il doit cependant être plausible pour garantir que le redressement améliore véritablement les résultats de recherche.
- Le lien entre la variable et le biais d'échantillonnage. Plus les variables retenues pour le calage sont liées aux biais d'échantillonnage, plus le redressement sera efficace. Notons tout de même qu'il est parfois impossible de retenir l'ensemble des variables pertinentes, car les biais sont trop forts pour que le redressement soit efficace. Appliquer ce critère demande donc un équilibre entre faisabilité et couverture des dimensions sociales, spatiales, démographiques…

Dans l'enquête Vico, nous avons calé l'échantillon sur les cinq variables suivantes, choisies en raison de leur rôle important dans la détermination des comportements et des attitudes analysées par l'enquête (Lavallée et Beaumont, 2016) : le sexe (2 modalités), l'âge (5 modalités), la ZEAT de la commune de résidence (8 modalités), la taille de l'unité urbaine d'appartenance de ce cette commune de résidence (2 modalités), et la catégorie socioprofessionnelle (6 modalités). Les distorsions observées dans l'échantillon brut, nous ont contraints à limiter le nombre de critères considérés. Par exemple, nous pensions inclure le niveau d'étude, fortement biaisé, en plus de ces cinq variables, mais avons dû y renoncer : son ajout conduisait à augmenter fortement la dispersion des coefficients de pondération, donc à réduire considérablement l'efficacité du redressement.

Réalisation du redressement

Nous avons effectué le redressement en utilisant le package R « icarus »[9] développé dans la continuité de la « macro » -CALMAR pour SAS (Sautory, 1991), une des premières solutions informatiques portant le calage sur marges tel qu'imaginé par Yannick Lemel et développé ensuite par Jean-Claude Deville (Lemel, 1976 ; Deville et Särndal, 1992).

---

[7] On utilise fréquemment des variables socio-démographiques ou géographiques, mais ce n'est pas une règle absolue. En fonction de la thématique de l'enquête et des données de cadrage disponibles, on pourra mobiliser d'autres critères de redressements (taux d'équipements, types de logements, comportements électoraux…).

[8] Iyadh Gacem, dans sa thèse de doctorat « Sondages : la post-stratification et ses limites » (2006), a montré sur des simulations sur l'enquête Emploi, que la marge d'erreur finit par augmenter de manière forte lorsqu'on augmente le nombre de critères et de catégories, et validait la règle de ne pas dépasser (N/30) critères.

[9] Rebecq Antoine, 2015, « icarus : un package R pour le calage sur marges et ses variantes », http://paperssondages16.sfds.asso.fr/submission_54.pdf.

Afin de pouvoir procéder au calage sur les marges de nos cinq variables, il faut connaître leur distribution dans la population de référence : nous avons utilisé pour cela l'Enquête Emploi 2018 de l'INSEE[10], qui était la source la plus fiable et la plus récente dont il était possible de disposer pour décrire ces caractéristiques pour la population de la France âgée de 18 ans et plus. En pratique, le choix des variables de calage est bien souvent contraint par leur disponibilité dans des enquêtes de référence comme celles-ci.

Les 5 variables de calage retenues comportent des proportions variables de non-réponses : le sexe est entièrement renseigné, mais les autres variables présentent des taux de non-réponse pouvant aller jusqu'à 3,9 %. Avant de pouvoir les utiliser pour le calage, il faut faire en sorte qu'il n'y ait plus de non-réponses. La façon d'y parvenir en dégradant le moins possible l'efficacité du redressement consiste à « imputer » ces non-réponses, autrement dit à les remplacer par des valeurs prédites à partir des autres réponses des individus. Pour ce faire, nous avons utilisé la méthode la plus simple préconisée par Béatrice Neiter et Benoît Buisson (2010) : cette méthode consiste à rechercher une variable "auxiliaire" fortement liée à la variable à imputer, puis à imputer les valeurs manquantes par tirage aléatoire sans déformer la distribution de la variable "auxiliaire". Ainsi, nous avons imputé les réponses manquantes pour l'âge à partir de la situation professionnelle, pour la ZEAT et TUU à partir du type de logement, pour la PCS des personnes en emploi (celle des autres personnes étant entièrement renseignée) à partir du niveau de diplôme.

Calculer les coefficients de pondération et tester le redressement

Le calcul des coefficients est ensuite réalisé en opérant un calage sur les marges des variables retenues, avec la méthode logit. Nous avons choisi cette méthode pour les chances de convergence qu'elle offre, ainsi que pour la possibilité offerte de borner la recherche d'une solution en spécifiant un poids minimal et un poids maximal. De cette façon, et sous réserve de vérifier l'absence d'accumulation aux bornes, nous contrôlons l'évolution du rapport de poids.

En réalité, pour l'enquête Vico, du fait de la très forte surreprésentation des femmes, nous avons opéré deux calages distincts (un pour les femmes et un autre pour les hommes), comme dans le cadre d'un échantillon stratifié non-proportionnel : chacun des deux sous-échantillons est calé sur les marges des différentes variables retenues dans la sous-population de référence du sexe correspondant. Cette technique permet de tenir compte des variations des caractéristiques sociodémographiques en fonction du sexe, tout en garantissant la convergence de l'algorithme. Ensuite, les coefficients ainsi calculés séparément pour les hommes et les femmes sont corrigés par un simple rapport pour tenir compte de la sous-représentation des hommes et la surreprésentation des femmes dans l'échantillon : les coefficients issus du calage des hommes seront multipliés par 1,8 et ceux issus du calage des femmes par 0,71.

Le redressement ainsi obtenu a une efficacité de 66 % pour les femmes et de 67 % pour les hommes, ce qui donne pour l'ensemble une efficacité de 54 %. La réunion des deux calages et l'application des coefficients multiplicateurs augmentent l'amplitude des poids, amplifient la déformation de l'échantillon et font donc mécaniquement baisser le critère d'efficacité. Les coefficients vont d'un minimum de 0,19 à un maximum de 6,63, soit un rapport de 1 à 35,2. Mais si l'on considère indépendamment chaque strate (hommes/femmes), qui correspond à

10 https://www.insee.fr/fr/metadonnees/source/operation/s1449/presentation (consultée le 20/09/2022). Les données issues de l'enquête Emploi offrent deux avantages par rapport aux données issues du recensement (Fichiers détail anonymisés, INSEE). Premièrement, les fichiers détail de l'enquête Emploi sont disponibles plus rapidement, en juin N+1, alors qu'il faut attendre 3 ans pour les résultats du recensement car ils sont établis à partir de 5 années de recensement. Deuxièmement, si les données de l'enquête Emploi ne peuvent pas être exploitées à un niveau géographique fin (commune), à la différence des données du recensement, les variables sont plus nombreuses et souvent plus détaillées (dernier diplôme obtenu par exemple).

notre plan de redressement, les rapports de poids sont bien plus proches des bornes conseillées (de 1 à 13,2 pour les hommes, de 1 à 14,2 pour les femmes). Les fortes distorsions observées dans notre échantillon expliquent ces dispersions élevées.

La qualité de ce redressement peut enfin être appréciée en examinant des « tris à plat » redressés d'un certain nombre d'observations pour voir si on « retrouve » des distributions observées dans la population de référence ou dans des enquêtes similaires réalisées aux mêmes dates. On distingue trois types d'observations :

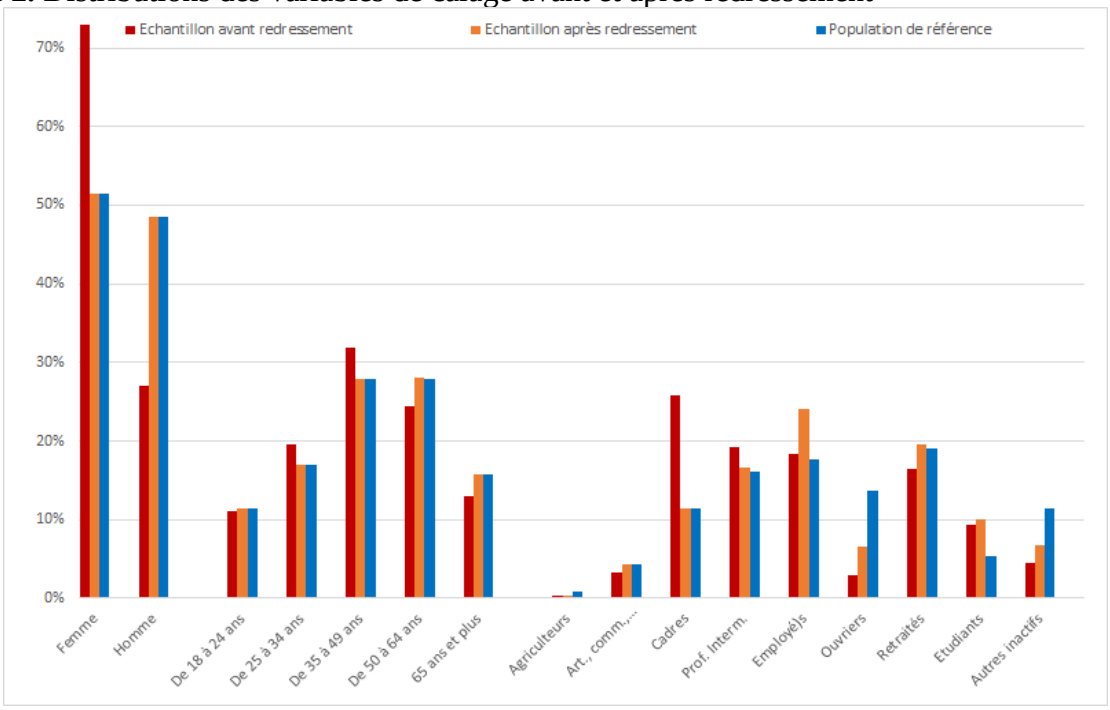
- Celles correspondant à des variables qui ont été utilisées pour le calage, et qui devraient être totalement ou assez fortement corrigées ;
- Celles qui n'ont pas été incluses, mais qui sont observables à la fois dans l'échantillon et dans la population de référence ;
- Et enfin celles qui ne sont pas observables à la fois dans l'échantillon et dans la population de référence, mais qui sont susceptibles de souffrir quand même de biais significatifs.

Il en ressort que si les cinq variables utilisées pour le calage sont très bien redressées (voir Graphique 2), et que c'est également le cas pour plusieurs variables observables décrivant les conditions de logement par exemple, il n'en va pas tout-à-fait de même pour d'autres variables d'intérêt, que celles-ci soient ou non observables dans la population de référence. Par exemple, selon les résultats de l'enquête Coconel, 7 % des Français·es n'ont pas résidé dans leur logement habituel pendant le confinement (Lambert et al., 2020). Dans l'enquête Vico, cette proportion passe de 8,6 % en pourcentages bruts à 8,1 % après redressement, sans qu'il soit toutefois possible de garantir laquelle des deux enquêtes prédit le mieux la proportion réelle dans la population de référence, qui reste inobservable.

L'échantillon redressé reste surtout biaisé sur un certain nombre de caractéristiques sociodémographiques assez facilement observables, comme en particulier le diplôme. Sur cette variable importante, le redressement a un impact positif mais insuffisant, dans la mesure où on continue de compter trop de diplômé·es du supérieur dans l'échantillon quelle que soit la classe d'âge ou le sexe, par rapport à la distribution théorique dans la population de référence, même si l'impact de la pondération est plus fort sur les tranches d'âge les plus actives (25-64 ans). Et au total, alors qu'il y a 30 % de diplômé·es du supérieur dans la population des 18-74 ans, on en trouve 69 % dans l'échantillon Vico avant redressement et encore 61 % après redressement (Graphique 3).

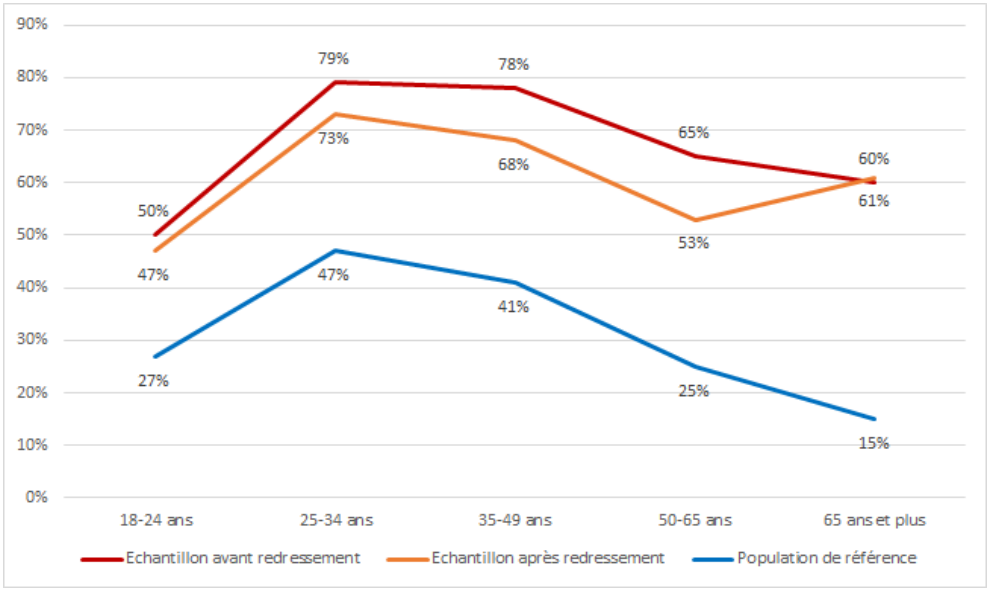
Une solution aurait été d'inclure le niveau de diplôme dans les variables de calage. Nous avons évidemment essayé cette approche, mais les coefficients qui en résultaient avaient une dispersion bien trop forte, dégradant ainsi l'efficacité du redressement et donc la significativité des futurs résultats. En bornant le calage, pour fixer l'amplitude maximale, c'est cette fois la convergence de l'algorithme qui pose problème. Nous retrouvons ici une limite conjointe du redressement et du mode de passation de l'enquête : lorsque les distorsions sont trop fortes, le redressement est médiocre ou impossible. Dans le cas de l'enquête Vico, le redressement ne permet pas de corriger correctement le taux d'abstention au premier tour des élections municipales de 2020. Alors qu'au niveau national, l'abstention était de 55 %, les données Vico, une fois redressées, l'estiment à 44%.

Figure 2. Distributions des variables de calage avant et après redressement



Sources : INSEE, Enquête Emploi, 2018 ; CNRS, Enquête Vico, avril-mai 2020.
Champ : Personnes résidant habituellement en France et âgées de 18 ans et plus (N = 16 224).
Lecture : La proportion de femmes est de 73 % dans l'échantillon avant redressement, 52 % après redressement, et elle est également de 52 % dans la population de référence.

Figure 3. Proportion de répondant·es ayant un diplôme d'études supérieures, avant et après redressement



Sources : INSEE, Enquête Emploi, 2018 ; CNRS, Enquête Vico, avril-mai 2020.
Champ : Personnes résidant habituellement en France et âgées de 18 ans et plus (N = 16 224).
Lecture : La proportion de répondant·es de 18-24 ans ayant un diplôme d'études supérieures est de 50 % dans l'échantillon avant redressement et de 47 % après redressement, alors qu'elle est de 27 % parmi les 18-24 ans dans la population de référence.

Quand utiliser les coefficients de pondération ? Une réponse (provisoire) de sociologues à un (épineux) problème de statistique

La question de la possibilité technique d'utiliser des coefficients de pondération dans les traitements statistiques est une question épineuse, à laquelle cet article n'entend pas apporter de réponse définitive. Les apports des statisticien·nes sur ce point peuvent être rangés dans deux catégories. D'une part, des travaux théoriques évaluent l'opportunité de la pondération d'un traitement inférentiel compte-tenu de ces hypothèses sous-jacentes[11]. D'autres travaux évaluent la possibilité opérationnelle d'introduire des poids dans les traitements statistiques[12]. Deux cas doivent ici globalement être distingués, selon que l'indicateur en question est linéaire, comme la moyenne, et l'introduction de poids est triviale, ou que l'indicateur n'est pas linéaire, comme la variance, auquel cas le problème est plus difficile à résoudre. Or, la réalisation de tests statistiques implique presque systématiquement le calcul à la fois de la moyenne et de la variance.

Nous entendons proposer ici une lecture sociologique des enjeux de l'utilisation, ou non, des pondérations. Plutôt que de considérer que l'usage de coefficients de pondération est « possible » ou « impossible » statistiquement, nous défendons une posture sociologique en affirmant que ce choix relève d'une « stratégie » d'administration de la preuve, qui est prise entre deux feux. En utilisant les pondérations, on se trouve en position de généraliser ses résultats, pour autant qu'on prenne toutes les précautions nécessaires au regard de la représentativité visée par le redressement. Pour autant, nous avons longuement vu tout au long de cet article que le redressement est tout sauf une technique « magique » permettant une généralisation automatique. Là comme ailleurs, la raison sociologique doit orienter la raison statistique. En particulier, il conviendra de vérifier l'effet des individus dont le poids s'écarte fortement du poids moyen. Dans des situations comme celles de l'enquête Vico, où le mode de passation du questionnaire induit presque mécaniquement une dispersion forte des poids, cette précaution devient impérative. En bref, la généralisation a un coût méthodologique dont il faut prendre la mesure, et dont il faut restituer les effets au moment de l'administration de la preuve. Cela étant dit, ne pas utiliser les pondérations est aussi un choix qui a un coût méthodologique élevé. Certes, puisque tous les individus « pèsent » le même poids, il n'y a aucun risque que l'un d'entre eux biaise un traitement et la « maîtrise des opérations » est complète ; mais le ou la chercheur·e reste alors « prisonnier·e » des biais de son échantillon.

L'alternative entre utiliser ou ne pas utiliser les coefficients de pondération se présente donc comme un choix cornélien, et il serait illusoire de chercher à aboutir à une règle définitive. Un premier signal doit être ici la qualité du redressement. Un redressement de mauvaise qualité peut signifier qu'il est impossible, avec ces données-là, de parvenir à réaliser cette opération complexe d'extrapolation. Répétons-le, ce type de situation doit être traitée au cas par cas par l'enquêteur ou l'enquêtrice, pas seulement en fonction de critères techniques. Pour guider la réflexion, nous proposons deux étapes qui peuvent servir de canevas :

1. Un test avec et sans redressement. Si le traitement donne un résultat identique avec et sans redressement, cela signifie que le coût de la généralisation est presque nul. Pourquoi ne pas le payer ? Dans le cas où le redressement n'apporte qu'une correction marginale à un échantillon déjà très bien construit, le·la praticien·ne se trouve généralement dans ce cas de figure.
2. Dans le cas où une différence est observable, il est nécessaire d'analyser le traitement pour différencier les individus en fonction de leur influence statistique.

11 Pour un exemple particulièrement didactique à propos de la régression, voir Davezies et D'Hautfoeuille, 2009.

12 On pourra ici se reporter à la bibliographie disponible en documentation des logiciels de statistique, dont la plupart prévoient sous le label "survey" des outils permettant de pondérer les indicateurs de statistiques descriptives.

Pour chaque technique, de tels outils existent et sont parfaitement reconnus : la valeur du khi-deux pour les cases tableaux croisés, la distance de Cook pour un modèle de régression ou encore les techniques de détection d'observations marginales (*outliers*) pour les analyses factorielles. En fonction du résultat donné par l'outil de détection adéquat, deux cas de figure se distinguent :

a. Si les individus ayant un coefficient de redressement largement supérieur à 1 sont ceux qui ont une forte influence sur le traitement, il apparaît plus sage de renoncer au redressement ou de montrer une grande prudence dans l'interprétation ;

b. Si les individus qui influencent fortement le traitement ne sont pas ceux qui ont un coefficient de redressement élevé, alors il est probablement possible de conserver le redressement. Cela, sous réserve que l'utilisation de la pondération ne dégrade pas significativement d'autres indicateurs de robustesse – par exemple, le coefficient de détermination d'un modèle de régression ou la qualité de représentation des profils dans une analyse factorielle.

Par conséquent, pour résumer tout cela, il nous semble que pour toutes les exploitations de ces données qui visent à estimer les taux de pénétration de comportements ou d'attitudes (analyse univariée), ou qui visent à estimer la significativité de liaisons entre variables (analyse bivariée), nous préconisons l'utilisation systématique de ces coefficients de pondération, mais il est recommandé d'examiner les tests de significativité des analyses bivariées et multivariées avec et sans pondération pour tester la robustesse de leurs résultats. Dans le cadre d'analyses multivariées, il conviendra de vérifier les effets des individus ayant les coefficients de pondération les plus élevés, par exemple en examinant leurs contributions aux axes des analyses géométriques des données, ou leurs distances de Cook dans les régressions multiples. Dans les analyses géométriques des données, les classifications et les régressions multiples, il est généralement recommandé de redresser, mais il peut être judicieux d'y renoncer dans certains cas particuliers où les distorsions apparaissent élevées (Afsa, 2016). Dans le cas des classifications, il reste toujours possible de pondérer *a posteriori*.

Conclusion

Les enquêtes en ligne - et l'enquête Vico n'y échappe donc pas - restent caractérisées par des biais d'échantillonnage qui sont d'abord des biais de couverture liés à l'accès à internet et aux outils numériques (Frippiat et Marquis, 2010). En 2019, l'INSEE indiquait que 10 % des foyers (parmi ceux comptant une personne de moins de 75 ans, soit un univers assez proche de celui de l'enquête Vico) n'avaient pas d'accès à internet, et que près de 25 % des individus ne l'utilisaient pas au quotidien (et ont ainsi peu de chances de répondre à une enquête en ligne de plus de 20 minutes). La diffusion d'un questionnaire en ligne couvre donc probablement au mieux 80 % de la population visée, et la population non-couverte possède des caractéristiques sociodémographiques probablement nettement différentes. L'échantillon redressé ne peut donc être considéré comme représentatif que de la population des 18-74 ans ayant accès à internet.

Il ne faut pas pour autant se limiter à cette affirmation statistique sans chercher à « ouvrir la boite noire » du raisonnement statistique. Tout au long des différentes étapes distinguées dans cet article, nous avons cherché à illustrer combien le chemin vers un échantillon convenablement redressé est davantage fait d'hésitations que d'une application mécanique d'une série d'étapes standardisées. Notre retour d'expérience peut avoir, à ce titre, vocation à servir de guide pour se poser les bonnes questions.

Références

Afsa C (2016) « Le modèle Logit : Théorie et Application », *Documents de travail*, Insee, https://www.insee.fr/fr/statistiques/2022139 (consulté le 27/12/2022).

Bès MP, Bidart C, Defossez A et al (2020) « La vie en confinement : objectifs et premiers résultats », *La vie en confinement : études et résultats*, https://vico.hypotheses.org/17 (consulté le 27/12/2022).

Davezies L, D'Hautfoeuille X (2009) « Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à des données d'enquête », *Documents de travail*, G 2009/06, Insee, Direction des études et synthèses économiques, https://www.insee.fr/fr/statistiques/1380863 (consulté le 27/12/2022).

Deville JC, Särndal CE (1992) « Calibration estimators in survey sampling », *Journal of the American statistical Association*, 87, 418: 376-382.

Deville JC, Särndal CE, Sautory O (1993) « Generalized raking procedures in survey sampling », *Journal of the American statistical Association*, 88, 423: 1013-1020.

Dussaix AM, Grosbras JM (1993) *Les sondages : principes et méthodes*, Paris : Presses universitaires de France.

Fischer C, Bayham L (2019) « Mode and Interviewer Effects in Egocentric Network Research », *Field Methods*, 31: 195-213.

Frippiat D, Marquis N (2010) « Les enquêtes par Internet en sciences sociales : un état des lieux », *Population*, 65, 2 : 309-338.

Gerville-Réache L, Couallier V, Paris N (2011) « Echantillon représentatif (d'une population finie) : définition statistique et propriétés », https://hal.archives-ouvertes.fr/hal-00655566/document (consulté le 27/12/2022).

Guilbert P (2001) « La pratique du redressement », in Lejeune M (dir.), *Traitements des fichiers d'enquêtes : redressements, injections de réponses, fusions*, Grenoble : Presses universitaires de Grenoble.

Lambert A, Cayouette-Remblière J, Guéraut E et al (2020) « Logement, travail, voisinage et conditions de vie : ce que le confinement a changé pour les Français », *Population & Sociétés*, 579, https://www.ined.fr/fr/actualites/presse/coronavirus-logement-travail-voisinage-conditions-de-vie/ (consulté le 27/12/2022).

Lavallée P, Beaumont JF (2016) « Weighting: Principles and practicalities », *C Wolf, D Joye, TW Smith, YC Fu (Herausgeber): The SAGE Handbook of Survey Methodology, SAGE, London*, 460-476.

Lejeune M (2021) *La singulière fabrique des sondages d'opinion*. Paris : L'Harmattan.

Lemel Y (1976) « Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondage », 273-282.

Mariot N, Mercklé P, Perdoncin A (dir.) (2021) Personne ne bouge. Une enquête sur le confinement du printemps 2020. UGA Éditions, ⟨10.4000/books.ugaeditions.18372⟩. ⟨hal-03509636⟩

Martin O (1999) *Raison statistique et raison sociologique chez Maurice Halbwachs*, Éditions Sciences Humaines, https://www.cairn.info/revue-histoire-des-sciences-humaines-1999-1-page-69.htm (consulté le 27/12/2022).

Neiter B, Buisson B (2010) « Comment redresser une enquête thématique ? », *Documents de travail*, Insee, https://genes.bibli.fr/doc_num.php?explnum_id=10244 (consulté le 27/12/2022).

Quivy R, Van Campenhoudt L (2018) *Manuel de recherche en sciences sociales*, Dunod.

Sautory O (1991) « La macro SAS : CALMAR (redressement d'un échantillon par calage sur marges) », *Document de travail de la Direction des Statistiques Démographiques et Sociales no. F9310*.

Selz M (dir.) (2013) *La représentativité en statistique*, Ined Editions.

Annexe

Liste des titres qui ont relayé l'annonce de notre questionnaire :
La Provence, Corse Matin, La Montagne, Le Populaire du Centre, La République du Centre, Le Berry républicain, L'Yonne républicaine, L'Écho républicain, Le Journal du Centre, L'Éveil de la Haute-Loire, Sud-Ouest, La Marseillaise, Le Télégramme de Brest, Le Courrier picard, L'Est républicain et Le Républicain lorrain, La Voix du Nord, L'Est Éclair, Libération Champagne, La Dépêche du Midi et enfin Ouest-France.

Un courrier a été adressé aux rédactions des quotidiens régionaux afin de leur présenter notre enquête. Ce courrier d'une page indiquait à la fois l'objet de notre enquête, l'étude des liens sociaux et des solidarités pendant le confinement, et insistait sur le fait que "nous aussi", en tant que chercheur.es, nous nous retrouvions en situation exceptionnelle, coupé.es de nos outils de travail et dans l'impossibilité de réaliser une enquête de terrain à l'aide d'enquêteurs.trices. Nous avons également insisté sur le fait que la presse locale, grâce à sa proximité avec le terrain dans ce contexte d'isolement collectif, nous semblait être un excellent vecteur de diffusion de notre enquête.
La dernière partie du courrier était consacrée à deux points :
- La mise en place du site web qui permettait d'obtenir des informations plus précises sur le projet (composition de l'équipe et finalité de l'enquête).
- Des précisions sur l'exploitation des données et la déclaration de traitement de données personnelles établie auprès du Service Protection des Données du Cnrs.

Déclaration de conflits d'intérêts
Les auteurs déclarent n'avoir aucun conflit d'intérêt potentiel pour tout ce qui concerne le déroulement de la recherche, les droits d'auteur et/ou la publication de cet article.

Financement
Cet article s'appuie sur la première vague d'enquête du projet « La vie en confinement ». Ce programme a bénéficié d'un soutien financier de l'Agence nationale française de la recherche (appel Flash Covid-19).

Remerciements
Nous remercions l'ensemble des collègues qui se sont investi·es dans l'aventure du projet Vico, toutes les personnes qui ont accepté de répondre à notre enquête et également aux titres de la presse quotidienne et régionale qui ont relayé notre questionnaire.