PLAN DE SONDAGE DU BAROMÈTRE DE SANTÉ PUBLIQUE FRANCE : CONCILIER ESTIMATIONS RÉGIONALES, NATIONALE ET SUR DES SOUS-POPULATIONS

Noémie Soullier[1], Jean-Baptiste Richard[2] & Abdelkrim Zeghnoun[3]

[1] *Santé publique France, France, noemie.soullier@santepubliquefrance.fr*
[2] *Santé publique France, France, jean-baptiste.richard@santepubliquefrance.fr*
[3] *Santé publique France, France, abdelkrim.zeghnoun@santepubliquefrance.fr*

Résumé. Le Baromètre de Santé publique France est une enquête transversale répétée tous les deux ans, interrogeant les personnes âgées de 18 à 79 ans résidant en France hexagonale, Guadeloupe, Guyane, Martinique et à La Réunion et vivant en logement ordinaire. Elle interroge sur les opinions, habitudes et connaissances en lien avec la santé et aborde de nombreux thèmes. Elle a pour objectif de produire des estimations au niveau régional et de chaque département et région d'outre-mer, et au niveau de la France hexagonale globalement. Elle vise également à apporter un éclairage sur les inégalités sociales, en documentant les indicateurs dans des sous-populations d'intérêt comme les jeunes ou les personnes avec un faible niveau de vie. Le plan de sondage du Baromètre de Santé publique France a été construit afin de répondre à l'ensemble de ces objectifs, tout en respectant des contraintes de coûts. Aussi, un plan de sondage stratifié à allocations inégales a été défini. Il permet de sur-représenter les domaines d'intérêt pour les estimations dans lesquels un taux de réponse plus faible est attendu, afin de disposer d'un effectif suffisant pour les analyses. Il garantit également que la précision des estimations soit suffisante à la fois au niveau régional et au niveau France hexagonale.

Mots-clés. Plan de sondage, stratification, estimations régionales

Abstract. The French Health Barometer is a cross-sectional survey repeated every two years, interviewing people aged 18 to 79 residing in mainland France, Guadeloupe, Guyana, Martinique and Reunion and living in ordinary housing. It investigates opinions, behaviors and knowledge related to health and addresses many topics. Its objective is to produce estimates at the regional level and for each overseas department and region, and at the level of mainland France as a whole. It also aims to shed light on social inequalities, by documenting indicators in subpopulations of interest such as young people or people with a low standard of living. The French Health Barometer survey plan was designed to meet all of these goals, while containing costs. Thus, a stratified sampling design with unequal probabilities was defined. It makes it possible to over-represent the domains of interest for the estimates in which a lower response rate is expected, in order to have sufficient numbers for the analyses. It also guarantees that the precision of the estimates is sufficient both at the regional level and at the mainland France level.

Keywords. Sampling design, stratification, regional estimation

1 Présentation de l'enquête

Le Baromètre de Santé publique France a pour objectifs de :

- Suivre les comportements de santé de la population résidant en France et leurs déterminants, avec une attention particulière accordée aux comportements modifiables et aux analyses des inégalités sociales et territoriales, notamment en produisant des estimations au niveau régional ;
- Documenter des enjeux de santé publique ou orienter une décision politique, notamment via l'interrogation des connaissances et des opinions de la population en vue de la mise en place ou de l'évaluation d'actions de prévention ou de politiques publiques ;
- Enrichir les connaissances sur l'état de santé de la population résidant en France, pour des indicateurs non couverts par d'autres enquêtes (telles que l'enquête santé européenne EHIS menée par la Drees) ou par d'autres systèmes de surveillance (dont le Système National des Données de Santé).

Le traitement est fondé sur la réalisation d'un intérêt public relatif à la mise en œuvre de la mission de surveillance de la santé conférée à Santé publique France par l'article L. 1413-1 du code de la santé publique.

L'étude, en constituant un observatoire des comportements des personnes résidant en France, vise à aider à la définition, à l'orientation et à l'évaluation des politiques publiques de santé, de prévention et d'information de la population. Par sa régularité (biennale), l'étude participe également à la surveillance et à la veille sanitaire. Enfin, l'étude participe à la recherche, aux études, à l'évaluation et à l'innovation dans les domaines de la santé, notamment en produisant des résultats sur les inégalités sociales et territoriales.

1.1 Champ et unités enquêtées

La zone d'étude est la France hexagonale et les départements et régions d'Outre-Mer (DROM) hors Mayotte (c'est-à-dire la Martinique, la Guadeloupe, la Guyane et La Réunion). Pour ces territoires, la base de sondage (fichier Fideli) assure une bonne couverture de la population, ce qui n'est pas le cas pour Mayotte. Des enquêtes spécifiques doivent être menées à Mayotte, comme cela avait été le cas avec l'enquête Unono Wa Maore (1).

La population d'étude est constituée des adultes âgés de 18 à 79 ans résidant dans un logement ordinaire. Ce champ d'âge permet de couvrir l'ensemble des populations d'intérêt des actions de prévention menées par Santé publique France (SpFrance) et permet ainsi de développer un questionnaire et des objectifs communs. Par ailleurs, les âges en dehors du champ du Baromètre de SpFrance sont couverts par d'autres enquêtes dédiées (l'enquête EnClass de l'OFDT et l'enquête Enabee de Santé publique France pour les jeunes, les enquêtes Autonomie de la Drees pour les plus âgés qui couvrent également le champ des institutions, l'enquête EHIS de la Drees à la fois pour les mineurs et les personnes âgées), avec une méthodologie plus adaptée (autorisation parentale pour les mineurs, recours au face-à-face et au proxy pour les personnes les plus âgées). Enfin, la base de sondage utilisée est moins performante en deçà de 18 ans et au-delà de 80 ans, car la proportion de personnes à charge pour lesquelles les noms et prénoms sont moins souvent disponibles est plus importante à ces âges. Le champ d'âge a également été harmonisé avec le champ couvert par l'enquête de santé et de biosurveillance de Santé publique France (Albane). Enfin, l'enquête pilote a montré que la population des 80-85 ans était beaucoup moins joignable et participait beaucoup moins bien que les personnes âgées de 70-79 ans, ce qui justifie ce seuil.

Critères d'inclusion :

- Résider en France hexagonale, à la Martinique, en Guadeloupe, en Guyane ou à La Réunion au moment de l'enquête ;
- Etre âgé de 18 à 79 ans au 1er janvier de l'année de l'enquête ;
- Résider en logement ordinaire (maison, appartement ou habitation mobile individuelle).

Critères d'exclusion :

- Résider en communauté (EHPAD, maison de retraite, prison, caserne, foyer, hôtel) au moment de l'enquête.

Le tirage au sort est un tirage d'individus ; il est réalisé par l'Insee dans une base d'individus issue des fichiers fiscaux et plus précisément du dispositif Fidéli à partir de spécifications établies par Santé publique France. Cette base de sondage permet de réaliser un tirage d'individus, de disposer de leurs coordonnées (postales et/ou téléphoniques et/ou adresses e-mail), et également de disposer d'informations auxiliaires permettant de corriger la non-réponse totale. Le millésime le plus récent du fichier est utilisé, soit le millésime de l'année n-2 pour l'enquête collectée l'année n. Un seul individu par logement est enquêté. Le plan de sondage est stratifié afin de surreprésenter dans l'échantillon tiré au sort les régions les moins peuplées et les populations les moins enclines à répondre (jeunes et déciles de niveau de vie les plus faibles), dans le but d'assurer la production d'estimations sur ces domaines.

Concrètement, dans la base de sondage, le champ de l'enquête Baromètre de Santé publique France 2024 correspond à l'intersection des conditions suivantes :

- individu résidant en France hexagonale, Guadeloupe, Martinique, Guyane ou à La Réunion ;
- individu âgé de 18 à 79 ans (bornes incluses) au 1er janvier 2024 ;
- individu non décédé dans le dernier millésime de la base de sondage ;
- individu résidant dans un logement ordinaire ;
- individus résidant dans ce logement la majorité ou la moitié du temps (i.e. individus avec un poids positif dans Fidéli, poids=1 ou 0,5) ;
- résidences principales avec des occupants dans le champ Fidéli.

Tous les individus répondant aux critères ci-dessus sont inclus, quelles que soient les informations de contact disponibles (civilité, nom/prénom, coordonnées téléphoniques, adresse mail…).

1.2 Paramètres d'intérêt de l'enquête

Les questions portent sur les opinions, habitudes et connaissances en lien avec la santé des personnes au moment de l'enquête ou sur une période passée récente (jusqu'à 5 ans avant). Certaines questions portent également sur une projection dans l'avenir, jusqu'à un horizon de 5 ans.

Les indicateurs principaux sont des proportions estimées au niveau national et régional, dans l'ensemble de la population et par catégorie de la population. Des analyses d'évolution sont également menées. Des analyses d'associations sont réalisées, à l'aide des méthodes appropriées à l'indicateur étudié.

Le questionnaire permettra de recueillir les données déclaratives suivantes :

- Des déterminants des opinions et habitudes de santé ou caractéristiques individuelles et contextuelles (variables d'ajustement) pouvant influencer les opinions, attitudes, habitudes et l'état de santé : caractéristiques socio-démographiques et socio-économiques, individuelles (profession et catégorie sociale, pays de naissance, niveau de diplôme, sexe, âge…) et contextuelles (conditions de logement, conditions de travail, composition du foyer…).
- Des données relatives à l'état de santé : santé perçue (mini-module européen de trois questions), maladies chroniques (diabète, hypertension), santé mentale (symptomatologie dépressive mesurée à l'aide de l'échelle CIDI-SF, anxiété mesurée par l'échelle CIDI-TAG, conduites suicidaires, bien-être mesuré par l'échelle de Cantril), maladies vectorielles et sexuellement transmissibles, chutes et accidents.
- Les habitudes de santé : consommations de tabac, d'alcool et de drogues, pratiques des jeux d'argent et de hasard, temps de sommeil, dépistage, pratiques vaccinales, pratiques alimentaires, activité physique et sédentarité, mesures de contrôle et d'hygiène.
- Les opinions et connaissances en matière de santé : sentiment d'information et perception du risque, exposition à des campagnes de prévention et modification des habitudes en lien avec l'exposition, connaissance des recommandations existantes.

Les analyses consisteront à produire :

- des analyses descriptives des indicateurs d'intérêt (proportions et intervalles de confiance pour les variables qualitatives, moments de la distribution pour les variables qualitatives) au niveau national, régional et par catégorie de population (sexe, âge, position sociale…) ;
- des analyses multivariées des indicateurs d'intérêt, permettant de décrire, toutes choses égales par ailleurs, les associations entre les habitudes/opinions d'intérêt et les caractéristiques socio-économiques, l'état de santé et les déterminants de ces habitudes/opinions en lien avec la thématique étudiée.

1.3 Protocole de contact et de collecte

La collecte des données s'effectue au premier semestre de l'année. Elle démarre mi-février, afin que les indicateurs de consommation d'alcool au cours des 7 derniers jours ne soient pas impactés par le défi de janvier ou « janvier sobre » (« Dry January ») invitant à ne pas boire d'alcool au mois de janvier. Elle démarre après le carnaval dans les DROM.

La collecte est réalisée par un prestataire de Santé publique France.

Le protocole de collecte est concurrentiel différé et inclut les modes de réponse internet et téléphone. Il consiste dans un premier temps à proposer uniquement le questionnaire par internet (pendant 5 semaines en Hexagone et 4 semaines dans les DROM). Dans un second temps, le questionnaire peut être administré par téléphone par un enquêteur. Le questionnaire par internet est réalisé en une seule séquence et disponible pendant toute la durée du terrain.

Les personnes reçoivent une invitation par courrier postal et par e-mail (lorsque disponible). Pendant la phase internet, les personnes sont relancées une fois par courrier, 6 fois par e-mail et une fois par téléphone. Un second courrier de relance annonce le démarrage de la phase téléphonique, et est accompagné d'un e-mail de relance. Jusqu'à 4 e-mails de relance supplémentaires sont envoyés, ainsi qu'un dernier courrier de relance. Le terrain dure environ 15 semaines.

2 Contraintes et hypothèses du plan de sondage

L'ensemble du territoire faisant partie du champ de l'enquête est mobilisé pour le tirage. Il s'agit donc d'un tirage à 1 degré. Le plan de sondage, décrit ci-après, présente un tirage aléatoire stratifié à allocations inégales.

2.1 Contraintes du plan de sondage

Le plan de sondage a été défini pour répondre aux contraintes suivantes :

- Contraintes budgétaires

Le budget de l'enquête n'est pas uniquement lié au nombre de répondants, mais également au nombre d'individus échantillonnés, ainsi qu'à la répartition des répondants entre les modes et au moment de réponse. Plus les personnes échantillonnées répondent tôt et par internet, moins les coûts sont élevés. Cependant, le poste qui pèse le plus sur le coût de la collecte reste les entretiens par téléphone : avec les hypothèses prises pour 2024, ils représentent 35 % du coût de collecte de l'enquête. L'envoi des courriers, dépendant à la fois du nombre d'individus échantillonnés, du nombre de répondants et du moment où ils répondent, est également un poste important et représente 21 % du coût de collecte. Afin de refléter ces éléments, les contraintes budgétaires sous-tendant l'échantillonnage sont alors exprimées en nombre de personnes échantillonnées combiné à un nombre d'entretiens téléphoniques, plutôt qu'en nombre de répondants :

- 65 000 personnes échantillonnées, dont 8 700 répondants par téléphone en Hexagone,
- 25 000 personnes échantillonnées, dont 3 000 répondants par téléphone dans les DROM.

- Contrainte de précision pour diffusion d'estimations au niveau régional

La contrainte de précision (i) est fixée par un coefficient de variation (CV) inférieur à 16,5 % pour l'estimation d'une prévalence p supérieure ou égale à 15 % sur un domaine équivalent à un quintile des répondants (par exemple les 18-29 ans) au niveau régional.

- Contraintes de précision pour diffusion d'estimations sur la France hexagonale

Deux contraintes de précision pour des variables d'intérêt estimées au niveau de la France hexagonale ont été étudiées :

(ii) Un coefficient de variation de 5 % maximum pour estimer la prévalence des pensées suicidaires, soit une prévalence de 4 % (2) ;

(iii) Une largeur de l'intervalle de confiance (IC) à 95 % de 1,5 % pour estimer la prévalence du tabagisme quotidien, soit une prévalence de 24,5 % parmi les 18-75 ans (3), afin de pouvoir détecter une différence de l'ordre de 1,5 % entre deux éditions de l'enquête ;

(iv) Une contrainte plus générale de dispersion des poids de sondage, visant à limiter l'augmentation de variance due au plan de sondage pour les estimations de l'ensemble des variables d'intérêt, est fixée à un effet plan des poids de sondage[1] de 1,1 maximum.

- Sur-représentation de certains groupes de populations

Afin de compenser par anticipation des taux de réponse attendus plus faibles, certains

[1] Cet effet plan inclut à la fois la dispersion introduite par la sur-représentation au sein de chaque région, et la dispersion introduite par des allocations inégales entre régions.

groupes de populations sont surreprésentés dans l'échantillon ; cela permet d'obtenir un effectif de répondants suffisant pour permettre des analyses sur ces domaines. Ces groupes de population sont les suivants : les jeunes de 18-29 ans, les niveaux de vie les plus faibles, les régions Guyane, Martinique, Guadeloupe, Corse, Hauts-de-France et Grand Est.

2.2 Stratification du plan de sondage

Les contraintes de diffusion régionale et de surreprésentation nécessitent la mise en œuvre d'un plan de sondage stratifié. La stratification est réalisée selon 3 dimensions :

- la région de résidence au sens de la nouvelle région administrative (soit 13 régions hexagonales et 4 régions outre-mer) ;
- l'âge en 2 classes : 18-29 ans / 30-79 ans ;
- le décile de niveau de vie en 2 classes : décile manquant et 1er à 3ème déciles (mq, D1-D3) / 4ème à 10ème déciles (D4-D10).

La stratification selon la région permet de satisfaire des contraintes de précision pour les estimations régionales et de sur-représenter certaines régions. Les stratifications sur l'âge et sur le niveau de vie permettent de sur-représenter les plus jeunes et les niveaux de vie les plus faibles. Les déciles de niveau de vie manquants ont été regroupés avec les faibles niveaux de vie car leurs taux de réponse à l'enquête pilote sont similaires.

La stratification repose donc sur 68 strates au total (17 strates régionales x 2 strates d'âge x 2 strates de niveau de vie). Les strates régionales (âge x niveau de vie) sont définies ainsi par la suite :

1. 18-29 ans - Faible niveau de vie (mq-D1-D3)
2. 30 ans et plus - Faible niveau de vie (mq-D1-D3)
3. 18-29 ans - Bon niveau de vie (D4-D10)
4. 30 ans et plus - Bon niveau de vie (D4-D10)

2.3 Coefficients de sur-représentation

Afin de compenser des taux de réponse plus faibles dans certaines régions et ainsi atteindre l'objectif de précision des estimations régionales pour chaque région, celles-ci sont sur-représentées par l'inverse du taux de réponse supposé, soit un coefficient de sur-représentation d'environ :

- 1,43 pour la Guyane (taux de réponse supposé = 70 % du taux de réponse moyen),
- 1,18 pour la Martinique, la Guadeloupe et la Corse (taux de réponse supposé = 85 % du taux de réponse moyen),
- 1,11 pour les régions Hauts-de-France et Grand Est (taux de réponse supposé = 90 % du taux de réponse moyen).

Afin de compenser des taux de réponse plus faibles observés dans certaines sous-populations d'intérêt lors de l'enquête pilote, et ainsi de disposer d'effectifs suffisants dans ces sous-populations pour les analyses, les strates correspondant aux personnes âgées de 18 à 29 ans, et aux déciles de niveau de vie les plus faibles sont sur-représentées. Afin de limiter la dispersion des taux de sondage au niveau d'une région donnée, cette sur-représentation est contrôlée et sera réalisée selon les coefficients présentés dans le tableau 1 ci-dessous (identiques pour chaque région) :

Tableau 1. Coefficients de sur-représentation des sous-strates régionales

Sous-strate régionale	Coefficient de sur-représentation
1 - 18-29 ans - Faible niveau de vie	1,2
2 - 30 ans et plus - Faible niveau de vie	1,1
3 - 18-29 ans - Bon niveau de vie	1,1
4 - 30 ans et plus - Bon niveau de vie	1,0

2.4 Hypothèses concernant les taux de réponse

De manière globale, on suppose que les taux de réponse observés lors de l'enquête pilote vont augmenter de 5 points de pourcentage, soit 13,5 %, cette augmentation étant liée au caractère obligatoire demandé pour l'enquête en 2024 (4).

Les taux de réponse estimés par sous-strate régionale varient de 22 % dans la strate 1 à 52 % dans la strate 4. Les taux de réponse supposés par strate sont le produit du coefficient appliqué à la région et du taux de réponse estimé dans chaque sous-strate régionale. Ils sont utilisés par la suite pour calculer le nombre d'individus à tirer au sort dans chaque strate *h*, en divisant le nombre de répondants souhaités par le taux de réponse estimé.

$$n_{tas,h} = \frac{n_{rep,h}}{TR_h} . \quad (1)$$

Le taux de réponse moyen par région dépendra également de la répartition de la population régionale dans chaque sous-strate, avec en particulier des différences entre la France hexagonale et les DROM (voir tableau 2). Il varie de 22 % en Guyane à 41 % en Pays de la Loire.

Tableau 2. Répartition moyenne de la population par sous-strates régionales en France hexagonale et dans les DROM

Sous-strate régionale	Répartition moyenne en France hexagonale	Répartition moyenne dans les DROM
1 - 18-29 ans - Faible niveau de vie	10 %	15 %
2 - 30 ans et plus - Faible niveau de vie	29 %	49 %
3 - 18-29 ans - Bon niveau de vie	8 %	5 %
4 - 30 ans et plus - Bon niveau de vie	53 %	31 %

2.5 Hypothèses concernant l'effet plan

L'évaluation de la dispersion des poids de sondage s'appuie sur l'effet plan des poids de sondage, calculé selon la formule de Kish (5) :

$$deff_{sond} = \frac{n * \sum_{i=1}^{n} \omega_i^2}{\left(\sum_{i=1}^{n} \omega_i\right)^2}$$, où ω_i est le poids de sondage. (2)

Au niveau France hexagonale, nous supposons que limiter l'effet plan des poids de sondage à 1,1 permettra de limiter l'effet plan, après traitements post-collecte (correction de la non-réponse, calage), à 1,4.

Au niveau régional, les coefficients de surreprésentation par sous-strate régionale introduisent une dispersion limitée des poids de sondage au sein de chaque région (effet plan = 1,004) par rapport à un sondage aléatoire simple. On estime qu'après traitements post-collecte (correction de la non-réponse et calage), la dispersion des poids au sein d'une région $deff_{reg}$, sera au maximum de 1,2 dans chaque région.

3 Résolution des contraintes

La formule de calcul de la taille d'échantillon minimale pour satisfaire la contrainte de précision est la suivante :

$$\Longrightarrow \quad CV = \frac{StdDev}{p} = \frac{\sqrt{\frac{deff\ *p*(1-p)}{n}}}{p} \tag{3}$$

$$n_p = arrondi\left(\frac{deff\ *(1-p)}{CV^2*p}\right) \tag{4}$$

La contrainte (i), sous l'hypothèse d'un effet plan régional estimé à 1,2, nécessite un effectif minimum de 250 répondants par quintile soit au minimum 1 250 répondants par région.

La contrainte (ii), sous l'hypothèse d'un effet plan global estimé à 1,4, nécessite un effectif de 13 500 répondants minimum en Hexagone.

La contrainte (iii), sous l'hypothèse d'un effet plan global estimé à 1,4, nécessite un effectif de 17 700 répondants minimum en Hexagone (et est donc plus restrictive que la contrainte (ii)).

La contrainte (iv) nécessite de déterminer l'allocation entre régions (de France hexagonale) permettant de limiter l'effet plan des poids de sondage à 1,1 et de satisfaire les contraintes précédentes : au moins 1 250 répondants par région, un nombre total de répondants supérieur ou égal à 17 700 (contrainte de précision (iii)) et un nombre de personnes échantillonnées inférieur ou égal à 65 000 (contrainte budgétaire). L'allocation recherchée est celle qui permet de satisfaire ces contraintes et qui disperse le moins les poids de sondage (ou à dispersion donnée, qui donne la plus petite taille d'échantillon).

3.1 Résultats de l'allocation uniforme

Pour répondre à la contrainte (i), il pourrait être envisagé de tirer au sort un échantillon qui, en compensant les taux de réponse estimés, permette d'obtenir exactement 1 250 répondants par région (allocation minimale uniforme en nombre de répondants). Cependant, une telle allocation :

- disperse fortement les poids de sondage[2] ($deff_{sond}$ = 1,36, voir tableau 3), ne satisfaisant pas la contrainte (iv) ;
- ne respecte pas la contrainte (iii), le nombre de répondants étant alors de 16 300 environ en Hexagone.

3.2 Résultats de l'allocation proportionnelle

Inversement, pour répondre au mieux à la contrainte (iv), une allocation proportionnelle peut être envisagée, cette allocation minimisant la dispersion des poids de sondage. L'allocation proportionnelle est calculée pour satisfaire la contrainte (i) dans toutes les régions hors Corse[3], l'allocation pour la Corse étant fixée à celle de l'allocation uniforme (1 250 répondants). Cette allocation proportionnelle permet de minimiser la dispersion des poids de sondage ($deff_{sond}$ = 1,05) mais ne permet pas de respecter la contrainte (iii) puisque le nombre de personnes échantillonnées pour la France hexagonale serait alors de 83 100 environ.

[2] Les poids de sondage des régions les plus peuplées sont beaucoup plus élevés que ceux des régions les moins peuplées : environ 3 000 pour l'Ile-de-France vs 600 en Centre-Val-de-Loire (et 60 en Corse).

[3] L'atteinte de l'objectif de 1 250 répondants en Corse implique un taux de sondage important, qui répercuté à l'échelle de la France métropolitaine impliquerait de tirer au sort environ 830 000 individus (pour environ 325 000 répondants), ce qui n'est pas réaliste.

3.3 Mise en œuvre d'une allocation mixte

Une solution de compromis entre l'allocation uniforme et l'allocation proportionnelle est donc nécessaire. Cette solution correspond à une allocation mixte de forme additive[4] entre l'allocation uniforme et l'allocation proportionnelle pour chaque région (6-8) :

$$n_{add,reg} = \alpha n_{prop,reg} + (1-\alpha) n_{uni,reg} \text{ avec } \alpha \in [0\,;1]. \quad (5)$$

Ainsi si α=0, alors on retrouve l'allocation uniforme ; et si α=1, on retrouve l'allocation proportionnelle.

Ici, la valeur de α qui respecte l'ensemble des contraintes est α=0,42. Cette valeur de α correspond à un effet plan égal à 1,1, permettant ainsi de respecter toutes les contraintes en minimisant la taille de l'échantillon (et donc le coût).

Tableau 3. Diagnostic des poids de sondage pour différentes valeurs de α, résultats pour la France hexagonale

	α=0 (uniforme)	**α=0,42**	**α=0,5**	**α=1 (proportionnel)**
Effectif tiré au sort (échantillon)	42 200	59 400	62 650	83 100
Effectif répondant total	16 300	23 050	24 300	32 350
dont répondants par téléphone	5 700	8 050	8 500	11 300
Effectif minimum répondant par région	1 250	1 250	1 250	1 250
Effectif maximum répondant par région	1 250	3 260	3 650	6 050
Poids de sondage minimum	50	50	50	50
Poids de sondage maximum	3 150	1 210	1 080	660
Effet plan	1,36	1,10	1,08	1,05
Rapport max / min des poids de sondage	60	23	21	12

NB : le poids de sondage minimum correspond à la région Corse, le poids de sondage maximum correspond à la région Ile-de-France.

La taille de l'échantillon à tirer au sort est d'environ 60 000 individus en France hexagonale. L'échantillon tiré au sort par région varie de 3 100 individus en Centre-Val-de-Loire à 8 150 individus en Ile-de-France.

Dans les DROM, seule la contrainte (i) s'applique. La résolution de cette contrainte par une allocation uniforme pour chaque DROM donne un échantillon de 18 000 individus tirés au sort dans les DROM, ce qui respecte la contrainte budgétaire. L'échantillon tiré au sort varie de 3 600 individus pour La Réunion à 5 800 pour la Guyane.

4 Description du tirage au sort

Le tirage au sort est un tirage systématique trié dans l'ordre par les variables suivantes :

- la présence d'un mail ou d'un numéro de téléphone (indicatrice) ;
- le sexe (homme, femme, manquant) ;
- l'âge en classes (18-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-74, 75-79) ;
- le décile de niveau de vie du logement ;
- l'identifiant du logement ;
- l'identifiant de l'individu.

Les tris sur la présence d'un mail ou d'un numéro de téléphone, ainsi que sur le sexe, ont pour objectif de se prémunir contre des tirages atypiques et ainsi d'assurer que la proportion de

[4] Une allocation mixte de forme puissance (allocation de Bankier) a également été étudiée, mais donnait de moins bons résultats en termes de dispersion des poids.

personnes sans mail ni téléphone, et la proportion de personnes sans sexe renseigné ne dépasse pas la proportion présente dans la base de sondage. Par ailleurs, le sexe est corrélé aux variables d'intérêt et est un domaine d'analyse.

Les tris sur l'âge et les déciles de niveau de vie permettent d'assurer une bonne répartition sur ces variables fortement corrélées aux variables d'intérêt, avec des modalités plus fines que les regroupements définissant les strates. L'âge a été placé en premier dans le tri car le nombre de modalités est moins important (fige moins le tri) et que cette variable est plus associée à la consommation de tabac que le décile de niveau de vie.

Le tri sur l'identifiant du logement pourra permettre a minima de regrouper les personnes qui auraient les mêmes caractéristiques sur toutes les variables de tris précédentes, mais n'évitera pas totalement le tirage de plusieurs individus au sein d'un même logement. En effet, la stratification faisant intervenir un critère individuel, le tirage ne peut garantir qu'un seul individu par logement soit tiré au sort. Or, l'enquête interroge un seul individu par logement. Aussi, après échange avec la division Sondages de l'Insee, les allocations ont été augmentées pour tenir compte d'une proportion de doublons d'individus au sein d'un même logement estimée à 0,17 % (observé dans d'autres enquêtes), et un seul individu par logement sera sélectionné aléatoirement, postérieurement au tirage le cas échéant (les doublons exclus seront considérés comme des non-répondants).

Compte tenu du calendrier du terrain d'enquête et du protocole multimode concurrentiel différé, il n'a pas été prévu d'échantillon de réserve. En effet, les éléments permettant de définir la nécessité d'y recourir pour palier un taux de réponse plus faible qu'attendu interviendraient trop tardivement dans l'avancée du terrain (a minima après l'ouverture du terrain téléphonique). A l'inverse, un effectif de répondants supérieur à l'attendu, que pourrait permettre d'éviter la mise en place d'échantillons de réserve, n'est pas problématique pour l'exploitation des données et aura un impact limité sur les coûts, compte tenu du protocole de collecte visant à maximiser la réponse par internet.

Bibliographie

1. Ruello M, Richard JB. Enquête de santé à Mayotte 2019 - Unono Wa Maore. Méthode. Saint-Maurice; 2022.
2. Léon C, du Roscoät E. Prévalence et évolution des pensées suicidaires en France métropolitaine en 2020 – Résultats du Baromètre santé. In: suicide Ond, editor. Suicide : mesurer l'impact de la crise sanitaire liée au Covid-19 - Effets contrastés au sein de la population et mal-être chez les jeunes. 5e rapport2022.
3. Pasquereau A, Andler R, Guignard R, Soullier N, Beck F, Nguyen-Thanh V. Prévalence du tabagisme et du vapotage en France métropolitaine en 2022 parmi les 18-75 ans. Bulletin Epidémiologique Hebdomadaire. 2023;9-10:152-8.
4. Didier E, Le Gléau J-P, Godinot A, Padieu R, Royer J-F, Bodin J-L, et al. L'obligation de réponse dans la statistique publique. Statistique et société. 2016;4(2).
5. Kish L. Survey Sampling. New York: John Wiley & Sons, Inc.; 1965.
6. Rebecq A, Merly-Alpa T. Optimisation d'une répartition mixte. Techniques d'enquête. 2018;44(2).
7. Koubi M, Mathern S. La nouvelle méthode d'échantillonnage de l'enquête trimestrielle ACEMO depuis 2006 - Amélioration de l'allocation de Neyman. Direction de l'animation de la recherche, des études et des statistiques (DARES); 2009.
8. Chiodini PM, Manzi G, Martelli BM, Verrecchia F. Divide, Allocate et Impera: Comparing Allocation Strategies via Simulation. Department of Economics, Management and Quantitative Methods at Università degli Studi di Milano; 2017.