

## COMBINAISON ET CORRECTION DE LA NON-RÉPONSE D'UN ÉCHANTILLON MULTIMODE ET MONOMODE

*Khaled Larbi (\*)*

*(\*) Insee, Direction de la méthodologie et de la coordination statistique et internationale*

*khaled.larbi@insee.fr*

**Mots-clés** : Multimode, Correction de la non-réponse, partage des poids

**Domaines** : Traitement de la non-réponse totale, Théorie des sondages après collecte

---

### Résumé

Les enquêtes statistiques utilisent de plus en plus des modes de collecte multiples afin d'améliorer les taux de réponse. L'approche dite multimode séquentielle consiste à proposer d'abord un mode peu coûteux, comme Internet, puis à relancer les non-répondants via un mode plus intermédiaire, tel que le téléphone ou le face-à-face. Cette stratégie permet d'équilibrer les coûts et la qualité des données.

L'existence d'un mode de collecte bon marché offre aussi la possibilité de tirer un échantillon supplémentaire monomode, collecté uniquement par Internet. Sa grande taille permettrait d'obtenir des estimations plus précises, notamment à des échelles locales. Toutefois, les taux de réponse faibles en mode Internet font craindre une sélection des répondants sur des caractéristiques inobservées, rendant les corrections classiques de la non-réponse insuffisantes et risquant de biaiser les estimations.

Ces travaux s'intéressent à la combinaison de deux échantillons : 1. un échantillon multimode, combinant Internet et téléphone, supposé peu affecté par le biais de non-réponse ; 2. un échantillon monomode Internet, éventuellement surreprésenté dans certains domaines.

Le premier fournit des estimations fiables, le second accroît la taille totale et la précision locale. L'enjeu est donc de savoir comment les combiner pour obtenir des estimateurs à la fois efficaces et robustes aux biais de sélection. Deux questions guident l'étude : - Peut-on corriger la non-réponse et combiner les deux échantillons pour produire de meilleurs estimateurs que ceux issus du seul multimode ? - Quels gains attendre pour les domaines surreprésentés dans le monomode ?

Plusieurs stratégies de combinaison et de correction sont comparées à l'aide de simulations, inspirées de l'enquête Logement 2023. Trois situations sont examinées : 1. non-réponse ignorable dans les deux échantillons ; 2. sélection sur inobservable uniquement dans le mode Internet ; 3. sélection sur inobservable dans tous les modes.

L'objectif est d'évaluer dans quelles conditions l'association d'un échantillon de qualité (multimode) et d'un échantillon volumineux mais biaisé (monomode) permet d'améliorer la précision globale et locale des estimations, tout en maîtrisant le risque de biais.

### Abstract

Recent surveys increasingly use mixed-mode data collection to improve response rates while controlling costs. In a sequential multimode design, respondents are first invited to answer online, and nonrespondents are later contacted by phone or face-to-face. This approach combines efficiency and data quality.

An additional Internet-only sample can be collected at low cost to increase sample size and enable more detailed local analyses. However, this monomode sample may suffer from selection bias due to lower response rates and unobserved differences among respondents.

This study explores methods to combine a reliable multimode sample with a larger but potentially biased Internet sample, aiming to improve estimation efficiency and domain-level precision. Using simulation studies inspired by the 2023 French Housing Survey, several strategies for nonresponse correction and sample integration are compared under different response mechanisms, from ignorable to nonignorable nonresponse.

The goal is to assess whether combining these two samples can yield more accurate and robust estimates despite potential selection issues.

---

## 1. L'enquête nationale sur le logement

L'enquête nationale sur le logement (ENL), menée régulièrement depuis 1955 par l'Insee, constitue la source de référence pour l'analyse des conditions de logement des ménages en France. Elle vise à décrire le parc de logements et les conditions d'habitation des ménages, à mesurer leurs dépenses liées au logement et à évaluer les effets des politiques publiques dans ce domaine. Elle complète les informations issues du recensement en apportant des données financières détaillées (loyers, charges, plans de financement, revenus) absentes de celui-ci.

L'enquête recueille des informations à la fois sur les logements et sur les ménages qui les occupent. Elle décrit le parc des résidences principales selon le type d'habitat, le statut d'occupation ou la localisation, et précise les caractéristiques physiques et la qualité du logement (taille, confort sanitaire, chauffage, dépendances, état, environnement, sécurité). Elle s'intéresse aussi aux modalités juridiques d'occupation, aux difficultés d'accès au logement, à la solvabilité des ménages, aux aides perçues, aux dépenses énergétiques et aux travaux réalisés. Les caractéristiques socio-économiques des occupants (taille et composition du ménage, revenus, situation professionnelle) sont également recueillies, ainsi que leur opinion sur leur logement, leurs projets résidentiels et leur mobilité récente. Ces données permettent notamment de calculer les loyers imputés et de produire des indicateurs de référence sur la structure du parc et les taux d'effort des ménages.

L'édition 2023-2024 introduit le tirage d'un échantillon  $S_{\text{multimode}}$ , collecté par multimode combinant le face-à-face, le téléphone et Internet, afin d'améliorer les taux de réponse tout en réduisant la mobilisation du réseau d'enquêteurs. Un deuxième échantillon  $S_{\text{monomode}}$  tiré selon un plan à un degré est soumis à un protocole de collecte Internet. Le questionnaire, d'une durée totale d'environ une heure, est séquencé en trois parties pour limiter les abandons, notamment dans les modes à distance. Un programme de tests mené jusqu'en 2023 a permis d'ajuster le protocole et d'assurer la comparabilité avec les éditions précédentes. De plus, dans cette nouvelle édition, l'échantillon monomode  $S_{\text{monomode}}$  sur-représente la région francilienne dans une visée méthodologique.

Monomode Idf	Monomode hors Idf	Multimode Idf	Multimode hors Idf
24781	25219	4046	23254

L'enquête nationale sur le logement occupe une place centrale dans le dispositif d'observation statistique du logement. Elle se distingue par l'articulation entre les caractéristiques des logements et celles des ménages qui les occupent, ainsi que par la taille de son échantillon, qui permet une description fine de multiples sous-populations : propriétaires, locataires du parc social, accédants récents ou ménages en situation de mal-logement. L'enquête constitue ainsi une source de référence pour de nombreuses analyses structurelles et chronologiques, ainsi que pour la modélisation des comportements résidentiels. Son importance stratégique se traduit par un soutien financier régulier de la part de ses partenaires publics et par son rôle clé dans la production d'une vision complète et cohérente du logement en France.

## 2. Combinaison des échantillons et correction de la non-réponse

Dans la suite, nous supposons disposer de deux échantillons d'individus répondants tirés dans une population ( $U$ ) :

- $S_{\text{multimode, rep}}$ , constitué de répondants issus d'un protocole multimode ;
- $S_{\text{monomode, rep}}$ , constitué de répondants issus d'un protocole monomode Internet.

Nous cherchons à combiner ces deux échantillons de manière à obtenir des estimations aussi efficaces que possible au sens de l'erreur quadratique moyenne.

Dans notre analyse, nous nous restreignons à l'étude d'estimateurs de totaux d'une variable d'intérêt  $\{y_k\}$ , notée  $t_y = \sum_{k \in U} y_k$ .

Deux grandes familles de méthodes sont comparées ici :

- **Combinaison puis correction de la non-réponse** : les échantillons  $S_{\text{multimode, rep}}$  et  $S_{\text{monomode, rep}}$  sont d'abord combinés pour former un échantillon unique  $S$ , muni d'une pondération construite à partir des pondérations initiales des deux échantillons, avant d'appliquer une correction de la non-réponse totale ;
- **Correction de la non-réponse puis combinaison** : les pondérations des échantillons  $S_{\text{multimode, rep}}$  et  $S_{\text{monomode, rep}}$  sont d'abord corrigées de la non-réponse afin d'obtenir des estimations approximativement sans biais des totaux, puis les deux échantillons sont combinés.

### 2.1. Méthode 1, 2, 3 : combinaison des échantillons et correction de la non-réponse

#### 2.1.1. Combinaisons

Dans la suite, nous supposons disposer de deux échantillons d'individus répondants tirés dans la population  $U$  :

- $S_{\text{multimode, rep}}$ , constitué de répondants issus d'un protocole multimode ;
- $S_{\text{monomode, rep}}$ , constitué de répondants issus d'un protocole monomode Internet.

Nous cherchons à combiner ces deux échantillons de manière à obtenir des estimations aussi efficaces que possible au sens de l'erreur quadratique moyenne.

Dans notre analyse, nous nous restreignons à l'étude d'estimateurs de totaux d'une variable d'intérêt  $\{y_k\}$ , notée  $t_y = \sum_{k \in U} y_k$ .

Deux grandes familles de méthodes sont comparées ici :

- **Combinaison puis correction de la non-réponse** : les échantillons  $S_{\text{multimode, rep}}$  et  $S_{\text{monomode, rep}}$  sont d'abord combinés pour former un échantillon unique  $S$ , muni d'une pondération construite à partir des pondérations initiales des deux échantillons, avant d'appliquer une correction de la non-réponse ;
- **Correction de la non-réponse puis combinaison** : les pondérations des échantillons  $S_{\text{multimode, rep}}$  et  $S_{\text{monomode, rep}}$  sont d'abord corrigées de la non-réponse afin d'obtenir des estimations approximativement sans biais des totaux, puis les deux échantillons sont combinés.

Pour ces méthodes, nous considérons deux échantillons tirés dans la population  $U$  :

- $S_{\text{multimode}}$ , tiré selon un plan  $p_{\text{multimode}}$  à deux degrés et muni des poids de tirage  $d_k^{\text{multimode}}$ . Cet échantillon peut être partitionné selon le mode de réponse des individus :

$$S_{\text{multimode}} = S_{\text{multimode, rep, int}} \cup S_{\text{multimode, rep, tel|int}} \cup S_{\text{multimode, non-rep}}$$

- $S_{\text{monomode}}$ , tiré selon un plan  $p_{\text{monomode}}$  à un degré et muni des poids de tirage  $d_k^{\text{monomode}}$ .

Les poids sont définis de manière à vérifier :

$$\mathbb{E} \left( \sum_{k \in S_{\text{multimode}}} d_k^{\text{multimode}} y_k \right) = \mathbb{E} \left( \sum_{k \in S_{\text{monomode}}} d_k^{\text{monomode}} y_k \right) = t_y$$

Ces deux échantillons peuvent être combinés pour former un échantillon unique, dont la pondération assure une estimation sans biais des totaux sous le plan de sondage.

Une approche naturelle consiste à utiliser une combinaison convexe des pondérations  $d_k^{\text{multimode}}$  et  $d_k^{\text{monomode}}$ .

Une alternative, étudiée dans cet article, consiste à conserver l'échantillon multimode  $S_{\text{multimode}}$  et à ne retenir que les répondants de l'échantillon monomode  $S_{\text{monomode}}$ .

Un ajustement des pondérations initiales est alors nécessaire : intuitivement,

$$\sum_{k \in S_{\text{multimode, rep, int}}} d_k + \sum_{k \in S_{\text{monomode, rep}}} d_k$$

ne fournit pas une estimation du même paramètre que

$$\sum_{k \in S_{\text{multimode, rep, int}}} d_k.$$

À la place, nous considérons la pondération suivante :

$$\tilde{d}_k(\lambda) = \begin{cases} \lambda_k d_k, & \text{si } k \in S_{\text{multimode, rep, int}} \cup S_{\text{monomode, rep}} \\ d_k, & \text{si } k \in S_{\text{multimode, rep, tel}} \end{cases}$$

Plusieurs choix de  $\lambda$  sont possibles :

- $\lambda_1 = \frac{\sum_{k \in S_{\text{rep}}^{\text{monomode}}} d_k}{\sum_{k \in S_{\text{rep}}^{\text{monomode}} \cup S_{\text{rep, int}}^{\text{multimode}}} d_k}$  : permet d'assurer que la somme des poids de tirage sur les individus répondants sur internet (monomode ou multimode) soit égale à la somme des poids de tirage sur les individus répondants en multimode internet,
- $\lambda_2 = \frac{1}{2}$  : permet d'être cohérent avec une approche par partage des poids non pondérée,
- $\lambda_3$  est obtenu en effectuant le calage de  $S_{\text{rep}}^{\text{monomode}} \cup S_{\text{multimode, rep, int}}$  sur  $S_{\text{multimode, rep, int}}$  en mobilisant des variables auxiliaires  $\{x_k\}$ . Plus formellement,  $\lambda_{3k} = \frac{w_k}{d_k}$  avec  $w_k$ , les poids obtenus après le calage décrit ci-avant.

### 2.1.2. Correction de la non-réponse

L'échantillon  $S_{\text{monomode, rep}} \cup S_{\text{multimode, rep, int}}$  muni de la pondération  $\tilde{d}_k(\lambda)$  est ensuite corrigé de la non-réponse totale : un modèle  $\hat{m}$  est appris permettant d'obtenir un estimateur de la probabilité de réponse  $\hat{p}_k$ . La pondération corrigée de la non-réponse  $\tilde{d}_k^{\text{CNR}}$  est obtenue en inflatant les poids  $\tilde{d}_k$  par l'inverse de la probabilité de réponse estimée :

$$\tilde{d}_k^{\text{CNR}}(\lambda) = \frac{\tilde{d}_k(\lambda)}{\hat{p}_k}$$

## 2.2. Méthode 4 et 5 : correction de la non-réponse et combinaison

### 2.2.1. Correction de la non-réponse

Contrairement aux méthodes 1, 2 et 3, la correction de la non-réponse est réalisée en amont sur chaque échantillon de répondants  $S_{\text{monomode, rep}}$  et  $S_{\text{multimode, rep}}$ . Des estimateurs des probabilités de réponse  $\hat{p}_k^{\text{monomode}}$  et  $\hat{p}_k^{\text{multimode}}$  sont obtenus à partir de modèles  $\hat{m}^{\text{multimode}}$  et  $\hat{m}^{\text{monomode}}$  à l'aide d'informations auxiliaires disponibles sur  $S_{\text{monomode}}$  et  $S_{\text{multimode}}$ .

Les pondérations corrigées de la non-réponse sont données par  $d_k^{\text{multimode, CNR}} = \frac{d_k^{\text{multimode, CNR}}}{\hat{p}_k^{\text{multimode}}}$  et  $d_k^{\text{monomode, CNR}} = \frac{d_k^{\text{monomode, CNR}}}{\hat{p}_k^{\text{monomode}}}$  et

Les modèles  $\hat{m}_{\text{multimode}}$  et  $\hat{m}_{\text{monomode}}$  peuvent être appris de manière :

- indépendante :  $\hat{m}_{\text{multimode}}$  en utilisant  $S_{\text{multimode, rep}}$  et  $\hat{m}_{\text{monomode}}$  avec  $S_{\text{monomode, rep}}$ . Par exemple, un modèle logistique est appris en utilisant comme variable d'intérêt  $R_k$  l'indicatrice de réponse et un modèle logistique est appris en faisant de même avec  $S_{\text{multimode, rep}}$ . Dans la suite de l'article, cette méthode est décrite comme **la méthode 4**.
- interdépendante :  $\hat{m}_{\text{multimode}}$  et  $\hat{m}_{\text{monomode}}$  sont appris en mobilisant l'information des deux échantillons. Par exemple, il est possible de reprendre les travaux de O. Guin, A. Leduc, L. Kozlowski, et N. Paliot [1] :
  - la probabilité de répondre au téléphone conditionnellement au fait de ne pas répondre par internet est modélisée et permet d'obtenir un estimateur  $\hat{p}_k^{\text{tel|non int}}$  en utilisant l'échantillon  $(S_{\text{multimode, rep, tel|int}} \cup S_{\text{multimode, non-rep}})$  et en supposant que nous disposons de toutes les variables décrivant la non-réponse (ainsi que de la bonne forme fonctionnelle).
  - la probabilité de répondre par internet est modélisée à l'aide de l'indicatrice de réponse sur Internet sur l'échantillon  $S_{\text{multimode, rep, int}} \cup S_{\text{multimode, rep, tel|int}}$ . La modélisation prend en compte une pondération : 1 si  $k \in S_{\text{multimode, rep, int}}$  et  $\frac{1}{\hat{p}_k^{\text{tel|non int}}}$  si  $k \in S_{\text{multimode, rep, tel|int}}$  afin que les individus de  $S_{\text{multimode, rep, tel|int}}$  soient *représentatifs* de l'échantillon  $S_{\text{multimode, rep, tel|int}} \cup S_{\text{multimode, non-rep}}$ . Un estimateur de la probabilité de répondre par internet  $\hat{p}_k^{\text{int}}$  est obtenu.

- la probabilité de répondre pour un individu  $k \in S_{\text{multimode}}$  est estimée par  $\hat{p}_k^{\text{multimode}} = \hat{p}_k^{\text{int}} + (1 - \hat{p}_k^{\text{int}})\hat{p}_k^{\text{tel non int}}$  et pour un individu  $k \in S_{\text{monomode}}$  par  $\hat{p}_k^{\text{monomode}} = \hat{p}_k^{\text{int}}$

Dans la suite, la méthode décrite ci-avant est nommée **méthode 5**.

## 2.2.2. Combinaison

À l'aide des probabilités de réponse estimées, il est possible de construire des pondérations corrigées de la non-réponse pour les échantillons  $S_{\text{multimode,rep}}$  et  $S_{\text{monomode,rep}}$ . Pour une variable d'intérêt  $\{y_k\}$ , deux estimateurs du total sur la population  $t_y$  peuvent être produits :  $\hat{t}_{y, \text{multimode}} = \sum_{k \in S_{\text{multimode}}} y_k \frac{d_k^{\text{multimode}}}{\hat{p}_k^{\text{multimode}}}$  et  $\hat{t}_{y, \text{monomode}} = \sum_{k \in S_{\text{monomode}}} y_k \frac{d_k^{\text{monomode}}}{\hat{p}_k^{\text{monomode}}}$ .

Si ces estimateurs sont sans biais alors toute combinaison convexe de ces deux estimateurs  $\hat{t}_{y, \mu} := \mu \hat{t}_{y, \text{multimode}} + (1 - \mu) \hat{t}_{y, \text{monomode}}$  avec  $\mu \in [0; 1]$  l'est aussi. Pour autant, la variance d'une combinaison convexe de ces estimateurs  $\hat{t}_{y, \mu}$  est différente de la variance de  $\hat{t}_{y, \text{multimode}}$  et  $\hat{t}_{y, \text{monomode}}$  : il est possible de trouver  $\mu^*$  telle que  $\mathbb{V}(\hat{t}_{y, \mu}) \leq \mathbb{V}(\hat{t}_{y, \text{multimode}})$  et  $\mathbb{V}(\hat{t}_{y, \mu}) \leq \mathbb{V}(\hat{t}_{y, \text{monomode}})$  (potentiellement avec au moins une inégalité stricte).

Le choix optimal de  $\mu^*$  dépend donc de la variance (et des covariances) sous le plan des deux échantillons. Le plan de l'enquête Logement est complexe : il est délicat d'obtenir l'expression close de  $\mu^*$ . À la place, nous nous proposons de comparer plusieurs valeurs de  $\mu$  :

- $\mu_1 = \frac{n_{\text{multimode}}}{n_{\text{multimode}} + n_{\text{monomode}}}$  - la taille relative de l'échantillon multimode (Remarque : il s'agit du  $\mu^*$  optimal si les échantillons  $S_{\text{multimode}}$  et  $S_{\text{monomode}}$  avaient été tirés selon des plans aléatoires simples sans remise).
- $\mu_2 = \frac{n_{\text{multimode, rep}}}{n_{\text{multimode, rep}} + n_{\text{monomode, rep}}}$  - la taille relative de l'échantillon des répondants multimode
- $\mu_3 = \frac{1}{2}$

## 3. Simulation

Afin d'étudier les différentes stratégies de redressements, des simulations sont réalisées. Ces simulations nécessitent de choisir certaines modélisations simplifiées pour les différentes étapes :

- simulation de la base de sondage et des variables d'intérêt,
- simulation du tirage,
- simulation du comportement de non-réponse.

### 3.1. Simulation de la base de sondage et des variables d'intérêt

Afin de reproduire les étapes de tirage d'échantillon, il est nécessaire de construire une base de sondage décrivant la population  $\mathcal{U}$ . Cette base de sondage contient différentes variables :

- une variable de strate géographique : cette variable permet de savoir si l'individu est dans la strate A (Île-de-France) ou dans la strate B. Cette stratification va être utilisée dans les plans de sondage.
- des variables auxiliaires observées (notées  $\{x_k\}$  où  $x_k \in \mathbb{R}^{d_x}$ ): ces variables sont liées aux variables d'intérêt et sont observées sur l'échantillon (répondants ou non),
- des variables auxiliaires inobservées (notées  $\{z_k\}$  où  $z_k \in \mathbb{R}^{d_z}$ ) : ces variables ont une influence sur les probabilités de réponse mais ne sont pas observées par l'utilisateur. Par

ailleurs, ces variables sont liées également aux variables d'intérêt  $\{y_k\}$  : certaines méthodes de correction de la non-réponse dans un cadre multimode proposée comme la méthode 5 utiliseront les variables  $\{y_k\}$  comme des proxys des variables  $\{z_k\}$ ,

- des variables d'intérêt (notées  $\{y_k\}$  où  $y_k \in \mathbb{R}^{d_y}$ ).

Les couples  $(x_k, y_k)$  sont générés selon une distribution multivariée dont les lois marginales et la matrice des corrélations sont déterminées en amont.

Le vecteur  $\{z_k\}$  est déterminé à l'aide d'un modèle linéaire :

$$z_k = y_k \beta + \Sigma \varepsilon_k$$

.

Le coefficient de détermination  $R^2$  issu de ce modèle permet de quantifier à quel point  $y_k$  est lié à  $z_k$ .

## 3.2. Simulation du tirage

Le tirage des échantillons monomode et multimode se fait en deux temps :

- tirage par SASSR stratifié d'un échantillon avec l'allocation  $n_{\text{Idf}} = n_{\text{monomode,Idf}} + n_{\text{multimode,Idf}} = 28827$  et  $n_{\text{hors Idf}} = n_{\text{monomode,hors Idf}} + n_{\text{multimode,hors Idf}} = 48473$ . On note  $S_{\text{Idf}}$ , l'échantillon tiré dans la strate Idf et  $S_{\text{hors Idf}}$ , l'échantillon tiré dans la strate hors Idf
- tirage des individus du lot multimode  $S_{\text{multimode}}$ :
  - tirage de l'échantillon des individus du lot multimode en Idf  $S_{\text{multimode, Idf}}$ : par SASSR de  $n_{\text{multimode,Idf}}$  dans  $S_{\text{Idf}}$
  - tirage de l'échantillon des individus du lot multimode hors Idf  $S_{\text{multimode, hors Idf}}$ : par SASSR de  $n_{\text{multimode,hors Idf}}$  dans  $S_{\text{hors Idf}}$
- tirage des individus du lot monomode  $S_{\text{monomode}}$ :
  - il s'agit des individus tirés dans l'échantillon et non tirés pour appartenir au lot multimode :  $S_{\text{monomode}} = S/S_{\text{multimode}}$

## 3.3. Simulation du comportement de non réponse

La formulation proposée ici s'inspire de travaux réalisés par P. Sillard [2]. Ces variables latentes peuvent être appréhendées comme des utilités à répondre. Dans la suite, nous considérons qu'il existe deux modes de collecte sans perte de généralité et nous considérons un échantillon  $S$  de taille  $n$ .

### 3.3.1. Définition des utilités

L'utilité à répondre d'un individu  $k$  à chaque mode est une fonction  $U : (x_k, z_k, \nu_k) \rightarrow (U_k^{\text{int}}, U_k^{\text{tel}})$  où  $(U_k^{\text{int}}, U_k^{\text{tel}}) \in \mathbb{R}^n \times \mathbb{R}^n$ .

Dans nos travaux, on suppose que l'utilité d'un individu  $k$  est une fonction linéaire en  $(x_k, z_k)$  auquel on ajoute un bruit. Plus formellement,  $U : (x_k, z_k, \nu_k) \rightarrow (U_k^{\text{int}}, U_k^{\text{tel}}) = (x_k^T \alpha^{\text{int}} + z_k^T \beta^{\text{int}} + \nu_k^{\text{int}}, x_k^T \alpha^{\text{tel}} + z_k^T \beta^{\text{tel}} + \nu_k^{\text{tel}})$

Quelques remarques : - pour un individu  $k$  donnée,  $\nu_k^{\text{int}}$  n'est pas forcément indépendant de  $\nu_k^{\text{tel}}$ . La dépendance entre ces deux bruits permet de contrôler la corrélation entre  $U_k^{\text{int}}$  et  $U_k^{\text{tel}}$  conditionnellement à  $(x_k, z_k)$ . Par exemple, si  $\nu_k^{\text{int}}$  et  $\nu_k^{\text{tel}}$  sont corrélés négativement, alors l'utilité de répondre par internet et l'utilité de répondre par téléphone seront liées négativement (conditionnellement à  $(x_k, z_k)$ ).

Par exemple, P. Sillard [2] considère la fonction d'utilité suivante :  $U : (X, Y, \nu) \rightarrow \left(0.3 + 0.8X - \frac{\bar{Y}-Y}{30} + \nu_1, 0.2 + 0.8X - \frac{\bar{Y}-Y}{30} + \nu_2\right)$   
 où  $(\nu_1, \nu_2) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.2 & 0.5 \\ 0.5 & 0.2 \end{pmatrix}\right)_{\otimes N}$

### 3.3.2. Définition de l'assignation à un mode

À partir de la définition des utilités, il est possible de définir des stratégies d'assignation[<sup>2</sup>] du mode de collecte : multimode concurrentiel, multimode séquentiel, monomode, etc.

D'après la formalisation proposée par PS, la stratégie d'assignation peut être vue comme une application  $M$  qui prend en entrant les utilités pour les différents modes et renvoie un mode[<sup>3</sup>] :  $M : (U^{\text{int}}, U^{\text{tel}}) \mapsto m \in \{\text{int}; \text{tel}\}$ .

#### Exemples :

- Lorsque l'échantillon est issu d'un lot monomode dont seul le mode internet est proposé :

$$M_{\text{monomode}} : (U^{\text{int}}, U^{\text{tel}}) \mapsto \text{int}$$

Autrement dit, peu importe l'utilité de l'individu à répondre au mode internet ou au mode téléphone, seul le mode 1 lui sera proposé.

- Lorsque le mode de collecte est multimode séquentiel (d'abord le mode 1, puis le mode 2), alors :

$$M_{\text{multimode seq}} : (U^{\text{int}}, U^{\text{tel}}) \mapsto \begin{cases} \text{int, si } U^{\text{int}} \geq 0, \\ \text{tel, sinon .} \end{cases}$$

Autrement dit, si l'individu a une utilité suffisamment importante pour répondre au mode internet, il répondra par ce mode, même si son utilité à répondre par le mode 2 est supérieure. Sinon, le mode 2 lui est assigné.

- Lorsque le mode de collecte est multimode concurrentiel (on suppose que l'utilisateur choisit son mode préféré), alors :

$$M_{\text{multimode conc}} : (U^{\text{int}}, U^{\text{tel}}) \mapsto \arg \max_{m \in \{\text{int}, \text{tel}\}} U_m$$

**Remarque** : l'assignation à un mode  $M$  permet de déterminer le mode de collecte par lequel un individu va répondre, s'il répond. Il est possible d'assigner un mode sans que l'individu ne réponde. Par exemple, dans le cas monomode 1, tous les individus sont assignés au mode 1, mais ne répondront pas si  $U_1 < 0$ . On ne peut pas explicitement parler de mode de collecte dans le cas d'un individu non-répondant.

### 3.3.3. Définition de l'utilité totale à répondre

L'utilité totale d'un individu à répondre est:

$$R^* : (U^{\text{int}}, U^{\text{tel}}, M) \rightarrow U^{\text{int}} \mathbb{1}_{\{M(U^{\text{int}}, U^{\text{tel}}) = \text{int}\}} + U^{\text{tel}} \mathbb{1}_{\{M(U^{\text{int}}, U^{\text{tel}}) = \text{tel}\}}$$

Il s'agit de l'utilité à répondre associée au mode assigné par la stratégie d'assignation  $M$ .

L'indicatrice de réponse  $R$  est définie par :

$$R : (U^{\text{int}}, U^{\text{tel}}, M) \rightarrow \mathbb{1}_{\{R^* \geq 0\}}$$

L'individu répondra si son utilité à répondre est suffisamment importante.

Quelques remarques :

- Prenons le cas d'une stratégie d'assignation monomode internet :  $M_{\text{monomode}} : (U^{\text{int}}, U^{\text{tel}}) \mapsto \text{int}$ . La probabilité de répondre d'un individu  $k$  est :

Si  $\nu_k^{\text{int}}$  suit une loi logistique alors  $\mathbb{P}(R_k = 1) = \frac{1}{1 + \exp(\mathbf{x}_k^T \alpha^{\text{int}} + \mathbf{z}_k^T \beta^{\text{int}})}$  : en plus de contrôler la dépendance entre utilité entre les modes, le choix de la distribution du bruit  $\nu_k = (\nu_k^{\text{int}}, \nu_k^{\text{tel}})$  permet de choisir la distribution de l'indicatrice de réponse (modèle logistique ici).

Dans la suite, nous générerons dans l'échantillon des répondants (conditionnellement à l'échantillon tiré) selon le modèle suivant :

$$U : (\mathbf{x}_k, \mathbf{z}_k, \nu_k) \rightarrow (U_k^{\text{int}}, U_k^{\text{tel}}) = (\mathbf{x}_k^T \alpha^{\text{int}} + \mathbf{z}_k^T \beta^{\text{int}} + \nu_k^{\text{int}}, \mathbf{x}_k^T \alpha^{\text{tel}} + \mathbf{z}_k^T \beta^{\text{tel}} + \nu_k^{\text{tel}})$$

avec  $\{(\nu_k^{\text{int}}, \nu_k^{\text{tel}})\}_k$  sont i.i.d et dont les marginales sont des lois logistiques (la structure de dépendance entre les deux composantes est donnée par une copule gaussienne dont le paramètre  $\rho$  est fixé en amont).

Afin de comparer les cinq méthodes présentées, nous nous proposons de considérer quatre scénarios :

Scé- na- rio	Endo- généité totale	Endogé- néité in- ternet	Commentaires
1	Non	Non	Les utilités ne dépendent pas des variables $z$
2	Non	Oui	La probabilité de réponse par Internet dépend des variables $\{x_k\}$ et $\{z_k\}$ mais la probabilité de réponse total dépend uniquement de $\{x_k\}$
3	Oui	Oui forte- ment	
4	Oui for- tement	Oui forte- ment	

## 3.4. Protocole de simulation

La qualité des différentes méthodes va être évaluée en comparant les biais et les erreurs quadratiques moyennes des estimateurs. Ces quantités vont être estimées en utilisant la méthode de Monte-Carlo.

Le protocole de simulation est décomposé en cinq étapes et ces étapes sont itérées indépendamment  $N_{\text{sim}} = 1000$  fois.

### 1. Tirage des échantillons

Deux échantillons,  $S_{\text{monomode}}$  et  $S_{\text{multimode}}$ , sont tirés selon le plan de sondage simplifié décrit précédemment.

### 2. Génération des échantillons de répondants

Les sous-échantillons de répondants,  $S_{\text{monomode, rep}}$  et  $S_{\text{multimode, rep}}$ , sont obtenus en appliquant le mécanisme de réponse défini plus haut. Pour chaque échantillon tiré  $S_{\text{monomode}}$  et  $S_{\text{multimode}}$ , les quatre scénarios décrits dans le tableau (ref) sont utilisés.

### 3. Application des méthodes d'estimation

À partir de ces échantillons, différentes méthodes sont mises en œuvre :

- **Méthodes 1 à 3** : combinaison des modes suivie d'une correction de la non-réponse ;
- **Méthodes 4 et 5** : correction de la non-réponse puis combinaison des modes.

### 4. Estimation du total

Pour une variable d'intérêt  $\{y_k\}_k$ , les estimateurs du total  $\hat{t}_y$  sont calculés selon chacune des méthodes.

Ainsi, une itération de simulation produit 5 (méthodes)  $\times$  4 (scénarios de non-réponse). À partir de ces  $N_{\text{sim}}$  simulations, pour une méthode  $m \in \{1, \dots, 5\}$  et un scénario  $s \in \{1, \dots, 4\}$  de non-réponse donnés, l'estimateur du biais de l'estimateur est défini par  $\widehat{\text{Biais}}(\hat{t}_y^{(m,s)}) = \frac{1}{N_{\text{sim}}} \sum_{l=1}^{N_{\text{sim}}} ((\hat{t}_y^{(m,s)})_l - \overline{\hat{t}_y^{(m,s)}})$  où  $\overline{\hat{t}_y^{(m,s)}} = \frac{1}{N_{\text{sim}}} \sum_{l=1}^{N_{\text{sim}}} (\hat{t}_y^{(m,s)})_l$  et l'erreur quadratique moyenne (EQM) de l'estimateur  $\hat{t}_y^{(m,s)}$  est définie par  $\widehat{\text{EQM}}(\hat{t}_y^{(m,s)}) = \frac{1}{N_{\text{sim}}} \sum_{l=1}^{N_{\text{sim}}} [(\hat{t}_y^{(m,s)})_l - \overline{\hat{t}_y^{(m,s)}}]^2$

## 4. Résultats

Les résultats sont présentés par scénario de non-réponse : les méthodes sont comparées à l'aune du biais et de l'erreur quadratique moyenne estimée. Dans les graphiques ci-après, les barres noires décrivent un intervalle de confiance lié à l'aléa inhérent aux simulations par la méthode de Monte-Carlo.

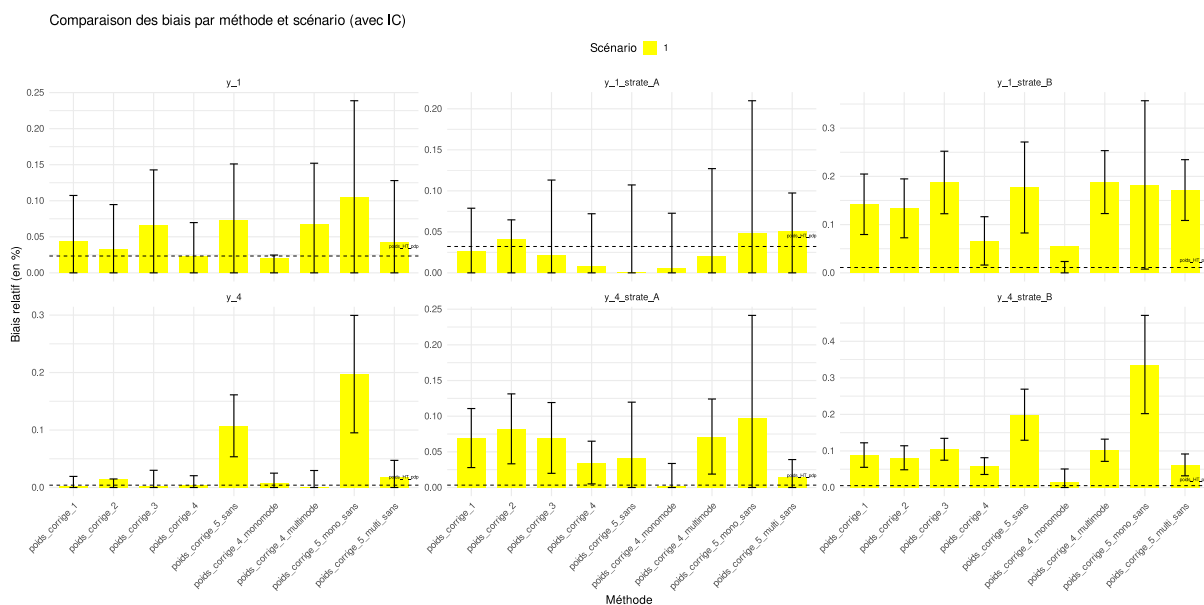
Les estimations de totaux de deux variables d'intérêt  $y_1$  et  $y_4$  sont comparés. Par ailleurs, nous considérons également la restriction de ces totaux à des domaines afin de comparer l'efficacité de ces méthodes à une échelle locale.

Les méthodes `poids_corrige_4_monomode` et `poids_corrige_4_multimode` (resp. `poids_corrige_5_mono_sans` et `poids_corrige_5_multi_sans`) désignent l'utilisation des méthodes 4 (resp méthode 5) sans combinaison : seul l'échantillon monomode ou multimode est utilisé avec une pondération corrigée de la non réponse.

## 4.1. Scénario 1 : dans le cas d'une non-réponse exogène (total et internet)

### 4.1.1. Biais

Étant donné que tous les variables nécessaires à l'estimation des probabilités de réponse sont observées : toutes les méthodes produisent des estimations sans biais.



### 4.1.2. Erreur quadratique moyenne

Il semblerait que :

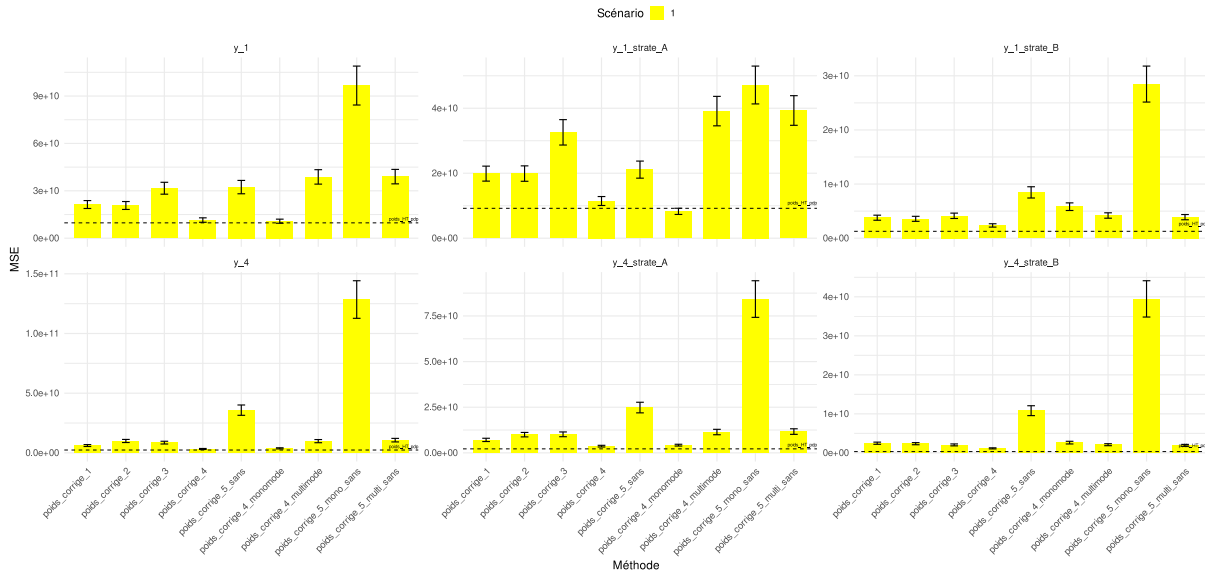
- les méthodes reposant sur la combinaison puis la correction de la non-réponse (méthodes 1, 2 et 3) soient moins efficaces ;
- la correction de la non-réponse indépendante (méthode 4) donne de meilleurs résultats que la méthode 5.

Cependant :

- des gains significatifs liés à la combinaison apparaissent lorsque la variable considérée est restreinte hors Île-de-France ;
- aucun gain significatif n'est observé lorsque la variable est restreinte à l'Île-de-France.

Ces résultats s'expliquent probablement par un choix non optimal des coefficients de combinaison  $\mu$ . Des travaux complémentaires sur la combinaison d'échantillons pourraient venir préciser ou nuancer cette conclusion.

Comparaison des mse par méthode et scénario (avec IC)

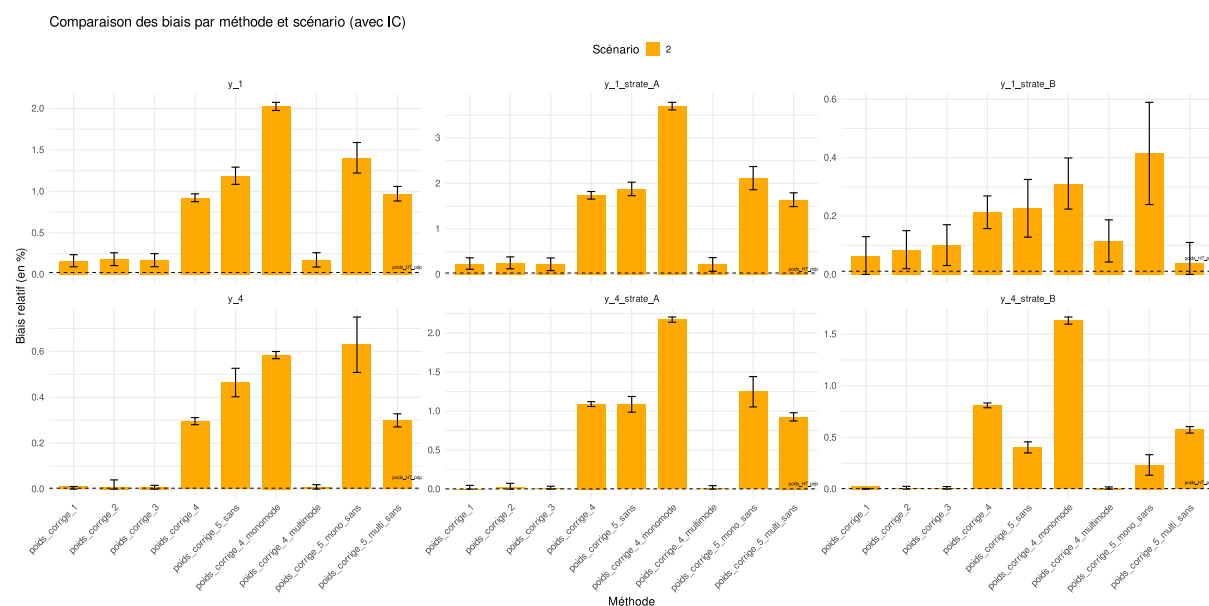


## 4.2. Scénario 2 : dans le cas d'une non-réponse endogène internet mais exogène total

Contrairement aux résultats observés sur le biais des différentes méthodes dans le scénario 1, les estimateurs apparaissent biaisés, à l'exception de la méthode 4 appliquée à l'échantillon multimode. Ce résultat est cohérent avec les hypothèses formulées : dans ce scénario, on suppose que toutes les variables nécessaires à la correction de la non-réponse totale sont disponibles dans l'échantillon multimode, tandis que certaines variables manquent pour corriger la non-réponse Internet.

Par ailleurs, certaines méthodes semblent moins sensibles au biais induit par le mécanisme de sélection sur Internet : les méthodes reposant sur la combinaison avant correction paraissent en atténuer l'effet.

## 4.2.1. Biais

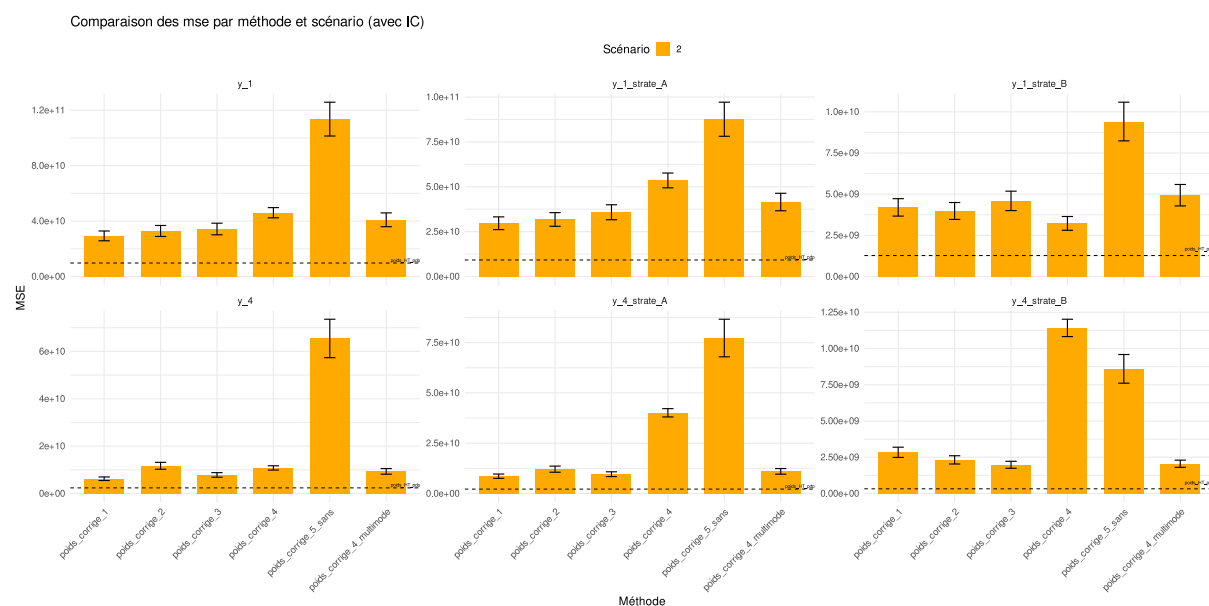


## 4.2.2. Erreur quadratique moyenne

Il semblerait que :

- les meilleures performances soient obtenues lorsque la combinaison est effectuée en amont (méthodes 1, 2 et 3) ou lorsque le multimode est utilisé avec une correction de la non-réponse classique (méthode 4 multi) ;
- lorsqu'on procède d'abord à la correction de la non-réponse puis à la combinaison, la combinaison n'apporte pas de gain significatif.

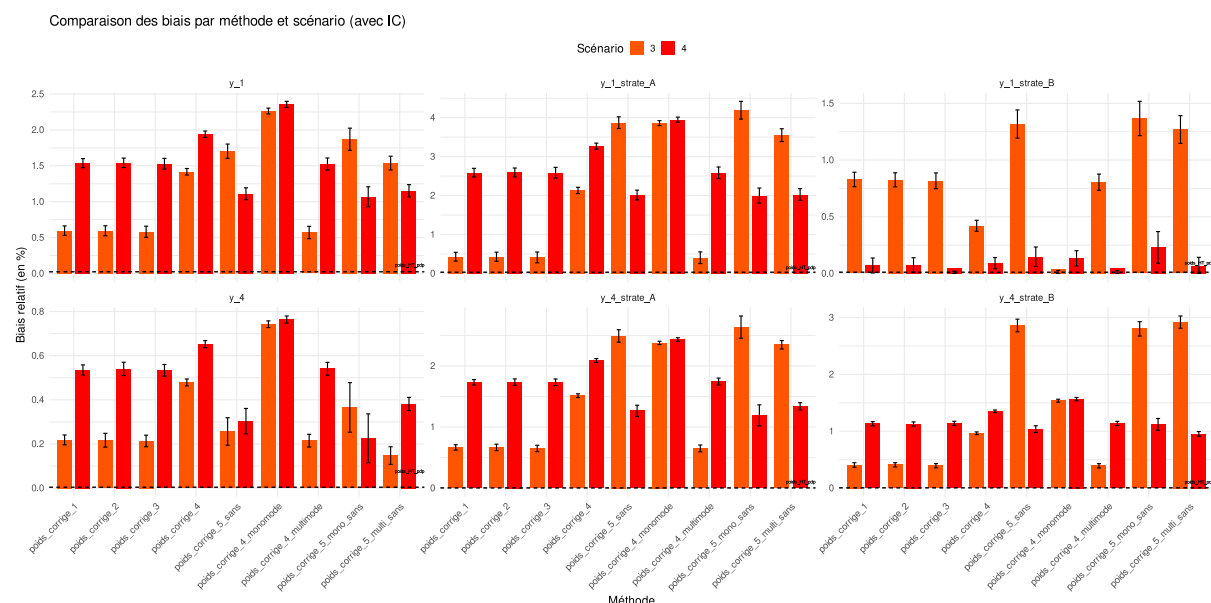
Enfin, il ne semble pas pertinent d'utiliser le monomode.



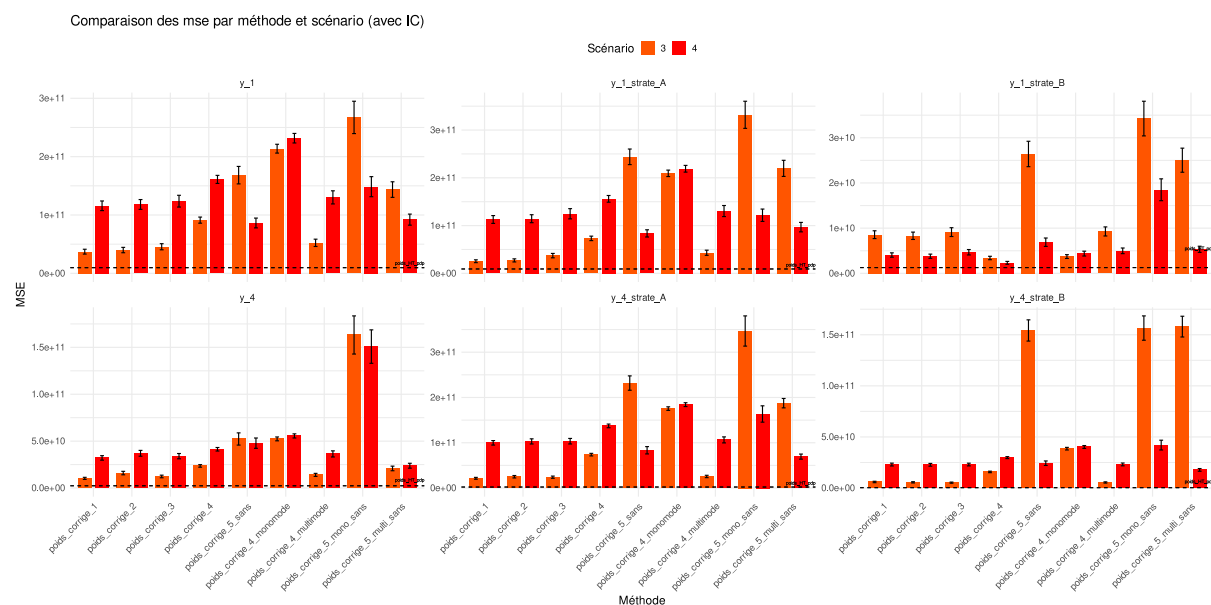
### 4.3. Scénario 3 et 4 : dans le cas d'une non-réponse endogène (total et internet)

Ces scénarios représentent des situations extrêmes, dans lesquelles certaines variables nécessaires à la correction de la non-réponse sont manquantes, tant pour le dispositif multimode que pour le mode Internet. L'ensemble des méthodes testées présente alors un biais notable. Toutefois, les approches consistant à combiner les échantillons avant d'appliquer la correction de la non-réponse semblent aboutir à des résultats moins dégradés. Néanmoins, il demeure difficile de tirer des conclusions générales à ce stade : des investigations complémentaires seraient nécessaires pour confirmer et préciser ces premières observations.

#### 4.3.1. Biais



#### 4.3.2. Erreur quadratique moyenne



## 5. Conclusions

Différentes stratégies de combinaison d'un échantillon multimode et d'un échantillon monomode Internet ont été comparées, dans un contexte où les mécanismes de non-réponse peuvent différer selon les modes de collecte. À partir de simulations inspirées du protocole de l'enquête Logement 2023, plusieurs méthodes de correction et de partage des poids ont été évaluées selon des scénarios plus ou moins favorables, allant d'une non-réponse ignorable à une sélection sur inobservables.

Les résultats montrent que, lorsque la non-réponse est ignorable, la combinaison des deux échantillons permet généralement d'améliorer la précision des estimations, notamment sur les domaines surreprésentés dans le monomode. En revanche, lorsque le biais de sélection est fort — en particulier lorsque certaines variables clés sont manquantes à la fois dans les dispositifs multimode et Internet —, toutes les méthodes présentent un biais significatif. Les approches consistant à combiner les échantillons avant de corriger la non-réponse semblent toutefois produire des résultats moins dégradés, suggérant un certain potentiel d'atténuation des biais.

Ces constats appellent à la prudence : les performances des méthodes dépendent étroitement du degré de non-ignorabilité et de la qualité des variables auxiliaires disponibles. Les travaux futurs pourraient approfondir ces analyses en testant d'autres configurations de plans de sondage, en intégrant des modèles de non-réponse plus complexes ou en mobilisant des sources de données externes pour enrichir les corrections.

Au-delà des résultats numériques, cette étude souligne l'intérêt d'une approche intégrée de la collecte et du traitement de la non-réponse, conciliant efficacité opérationnelle et rigueur méthodologique. Elle contribue ainsi à éclairer les conditions dans lesquelles les dispositifs mixtes de collecte peuvent être exploités de manière optimale pour garantir à la fois la précision et la fiabilité des estimations produites.

## 6. Annexe : choix des paramètres

### 6.1. Variables continues — Monomode

Nom variable	Distribution
x1	Uniforme (min = 1, max = 5)
x2	Normale (mean = 0, sd = 1)
x3	Exponentielle (rate = 0.5)
x4	Normale (mean = 3, sd = 1)
x5	Exponentielle (rate = 0.5)

### 6.2. Variables continues — Multimode

Nom variable	Distribution
x1	Uniforme (min = 6, max = 10)

Nom variable	Distribution
x2	Normale (mean = 0, sd = 1)
x3	Exponentielle (rate = 0.5)
x4	Normale (mean = 3, sd = 1)
x5	Exponentielle (rate = 0.5)

Nom variable	Distribution
y1	Exponentielle (rate = 0.5)
y2	Bêta (shape1 = 0.1, shape2 = 0.2)
y3	Normale (mean = 3, sd = 1)
y4	Normale (mean = 3, sd = 1)
y5	Normale (mean = 3, sd = 1)
y6	Normale (mean = 3, sd = 1)
y7	Normale (mean = 3, sd = 1)
y8	Normale (mean = 3, sd = 1)
y9	Normale (mean = 3, sd = 1)
y10	Normale (mean = 3, sd = 1)

## Bibliographie

- [1] O. Guin, A. Leduc, L. Kozlowski, et N. Paliod, « Correction de la non-réponse par repondération dans un contexte multimode séquentiel », in *Actes du 12ème Colloque Francophone sur les Sondages*, France, 2025.
- [2] P. Sillard, « Étude d'un design d'enquête multimode permettant d'identifier les effets de mode », sept. 2022.