



Journées de méthodologie
statistique de l'Insee

2025

Étude de la combinaison des échantillons monomode
et multimode par simulation

Khaled Larbi

¹Insee, DMS

Novembre 2025



L'ENL 2023 et le multimode

- ▶ L'ENL est une enquête mobilisant plusieurs échantillons :
 - ▶ monomode : tirage à un degré avec une surreprésentation de l'Île-de-France de taille $n_{\text{monomode}} = 50\,000$
 - ▶ multimode séquentiel (Internet, téléphone et face-à-face) : tirage à deux degrés $n_{\text{multimode}} = 27\,300$

Table – Effectifs par croisement

Monomode IdF	Monomode RdF	Multimode IdF	Multimode RdF
24781	25219	4046	23254



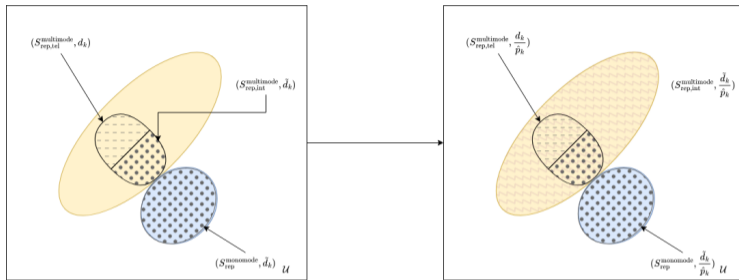
- ▶ Taux de réponse différents :
 - ▶ monomode : $\approx 30\%$
 - ▶ multimode : $\approx 70\%$
 - ▶ plus grand impact possible de l'endogénéité dans le cas monomode
- ▶ Comment corriger de la non-réponse sachant que nous réalisons une exploitation conjointe ?



- ▶ Deux possibilités :
 - ▶ combinaison des deux échantillons et traitement de la non-réponse commune :
methode_1, methode_2, methode_3
 - ▶ traitement de la non-réponse commune puis combinaison : methode_4,
methode_5



Combinaison avant CNR : méthodes 1 à 3



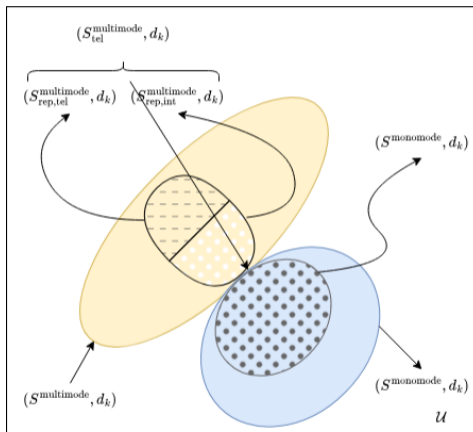
- ▶ On combine les répondants multimodes Internet et monomode Internet :

$$\text{pour tout } k \in S_{\text{rep}}^{\text{monomode}} \cup S_{\text{rep,int}}^{\text{multimode}}, \quad \tilde{d}_k = \lambda d_k$$

- ▶ puis CNR → les NR du monomode Internet ne sont pas pris en compte.

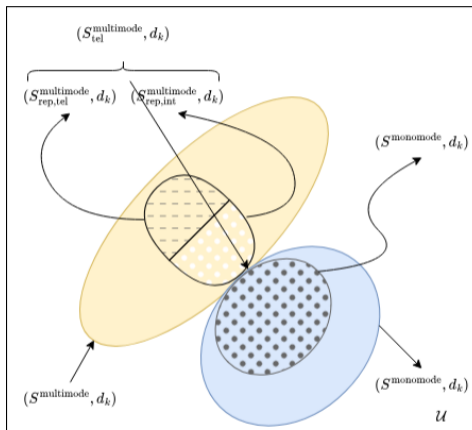


- ▶ methode_1 : $\lambda_1 = \frac{\sum_{k \in S_{\text{rep}}^{\text{monomode}}} d_k}{\sum_{k \in S_{\text{rep}}^{\text{monomode}} \cup S_{\text{rep, int}}^{\text{multimode}}} d_k}$
 - ▶ permet d'assurer que la somme des poids de tirage sur les individus répondants sur internet soit égale à la somme des poids de tirage sur les individus répondants en multimode internet
- ▶ methode_2 : $\lambda_2 = \frac{1}{2}$ (approche partage des poids non pondéré)
- ▶ methode_3 : généralisation de la méthode 1 en calant la pondération



- ▶ Estimation des probabilités dans les deux échantillons séparément
- ▶ Variable de réponse sans prise en compte du mode
- ▶ Partage des poids

$$\frac{1}{2}, \quad \frac{n_{\text{multimode}}}{n_{\text{multimode}} + n_{\text{monomode}}}, \quad \frac{n_{\text{multimode, rep}}}{n_{\text{multimode, rep}} + n_{\text{monomode, rep}}}$$

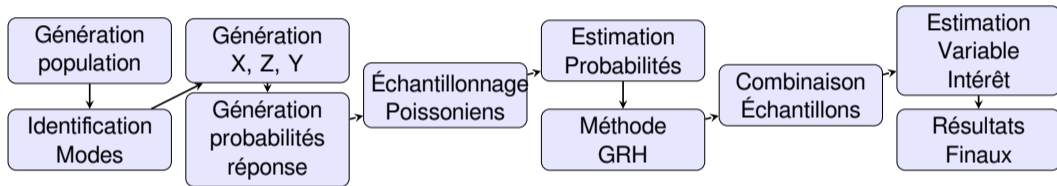


- ▶ Estimation des probabilités en utilisant la méthode additive avec \underline{y} sur $S^{\text{multimode}}$
- ▶ Application du modèle estimant les probabilités p_k^{Int} au monomode
- ▶ Partage des poids :

$$\frac{1}{2}, \quad \frac{n_{\text{multimode}}}{n_{\text{multimode}} + n_{\text{monomode}}}, \quad \frac{n_{\text{multimode, rep}}}{n_{\text{multimode, rep}} + n_{\text{monomode, rep}}}$$



- ▶ Comment comparer ces méthodes → Biais et variance (ou EQM) des estimations corrigés de la non-réponse pour différentes variables
- ▶ Calculs difficiles → utilisation de la méthode de Monte-Carlo afin d'estimer le biais et la variance
- ▶ Principe pour l'étude d'un estimateur $\hat{\theta}(S)$ de θ :
 - ▶ Tirage de $n_{\text{sim}} \approx 1000$ échantillons : $(S_1, \dots, S_{n_{\text{sim}}})$
 - ▶ Estimation du biais par $\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \hat{\theta}(S_j) - \theta$
 - ▶ Estimation de l'EQM par $\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} (\hat{\theta}(S_j) - \theta)^2$





- ▶ Création d'une base de sondage
- ▶ Utilisation d'informations de l'enquête ENL : stratification, taux de réponse, ...

Type de variable	Nombre	Observation	Particularités
Variables auxiliaires	$(n_X = 5)$	Observées pour tous	Différentes si l'individu est en IdF
Variables d'intérêt	$(n_Y = 10)$	Observées uniquement pour les répondants	Liées aux variables endogènes
Variables endogènes	$(n_Z = 2)$	Inobservées	Liées à la probabilité de réponse Internet



- ▶ Toutes les méthodes de CNR sont testées sous différents scénarios
- ▶ L'utilisateur ne dispose que des variables $\{\mathbf{x}_k\}$ pour corriger de la non-réponse mais la probabilité de répondre peut dépendre de $\{\mathbf{z}_k\}$
- ▶ Scénario de base : les probabilités de réponse ne dépendent que des variables observables $\{\mathbf{x}_k\}$
- ▶ Autres scénarios :
 - ▶ la probabilité de réponse totale est exogène mais pas les probabilités de réponse sur internet et par téléphone
 - ▶ la probabilité de réponse totale est endogène



Scénario	Endogénéité totale	Endogénéité internet
1	x	x
2	x	✓✓
3	✓	✓✓
4	✓✓	✓✓



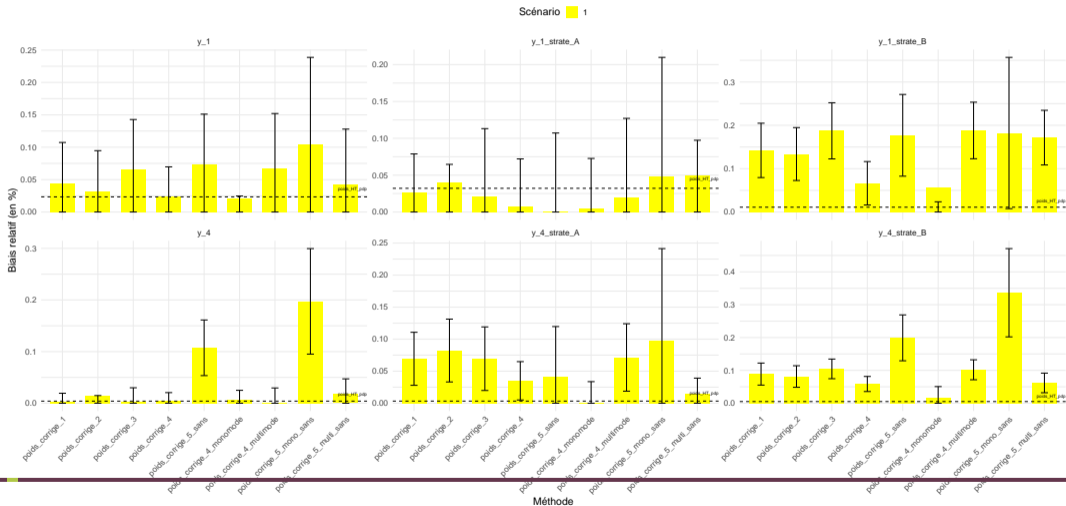
Il semblerait que :

- ▶ Meilleurs résultats : correction de la non-réponse indépendante « classique » dans $S^{\text{monomode,rep}}$ et dans $S^{\text{multimode,rep}}$
 - ▶ Pas de gain à combiner en amont (`methode_1`, `methode_2` et `methode_3`), ni à raffiner la CNR sur le monomode (`methode_5`)
- ▶ Une fois les échantillons corrigés de la NR, gain à la combinaison ?
 - ▶ Oui pour des estimations nationales (ou RdF)
 - ▶ Pas de gain significatif sur les estimations IdF



Résultat 1 : Biais

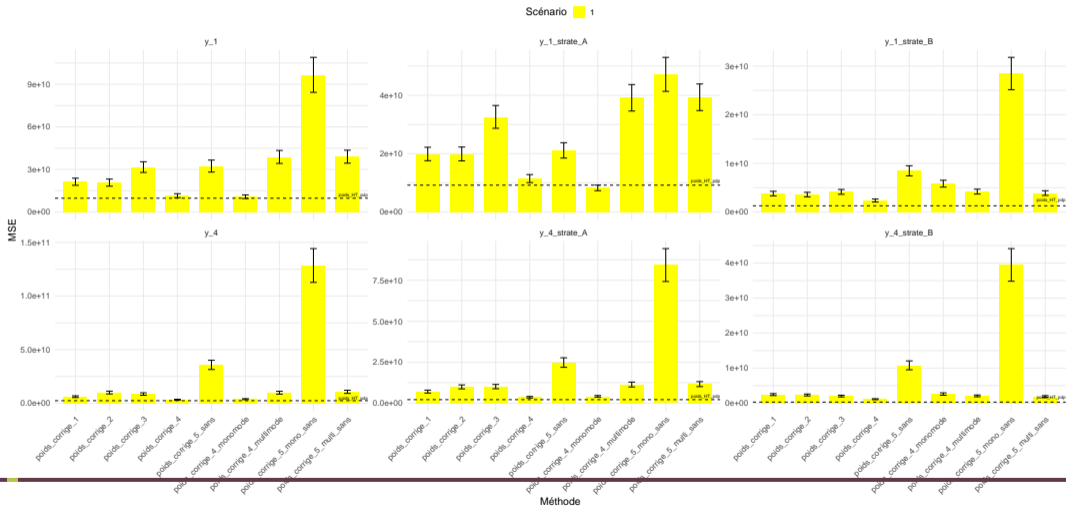
Comparaison des biais par méthode et scénario (avec IC)





Résultat 1 : Erreur quadratique moyenne

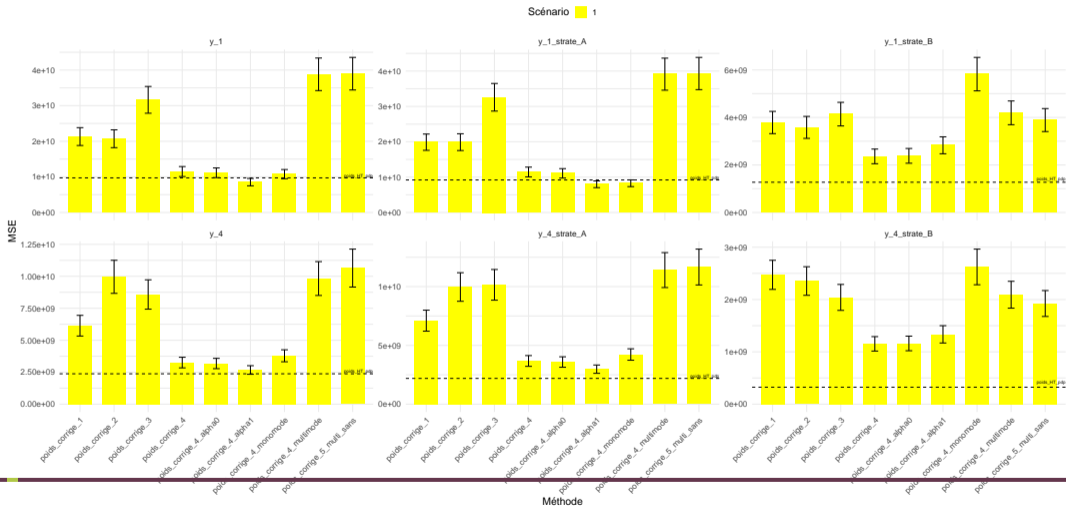
Comparaison des mse par méthode et scénario (avec IC)





Résultat 1 restreints : EQM

Comparaison des mse par méthode et scénario (avec IC)





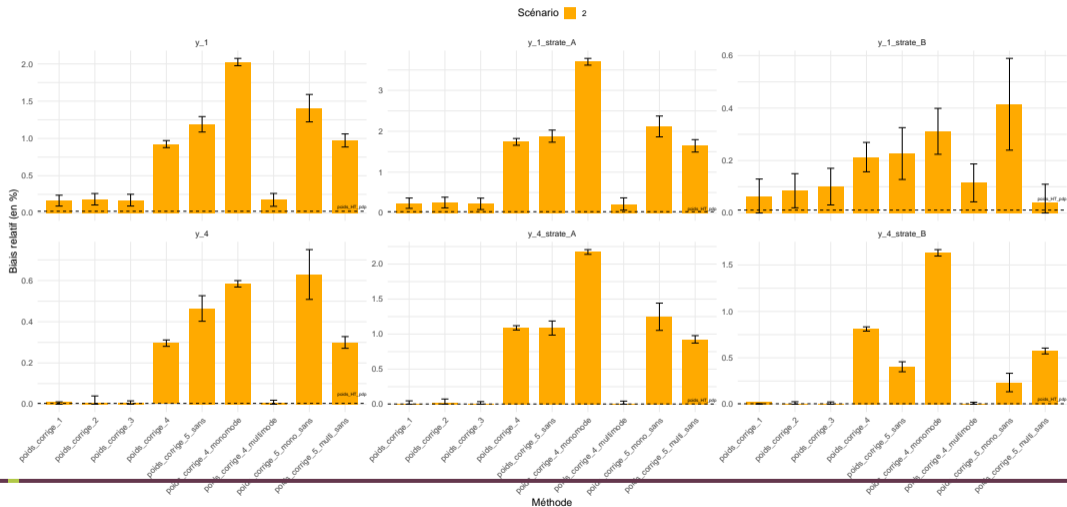
Il semblerait que :

- ▶ Meilleurs résultats : méthodes où on combine en amont (1,2,3) et l'utilisation du multimode avec CNR classique (4 multi)
- ▶ Si CNR puis combinaison : dégradation en termes de biais et d'EQM → il ne vaut mieux pas utiliser le monomode dans ces approches



Résultat 2 : Biais

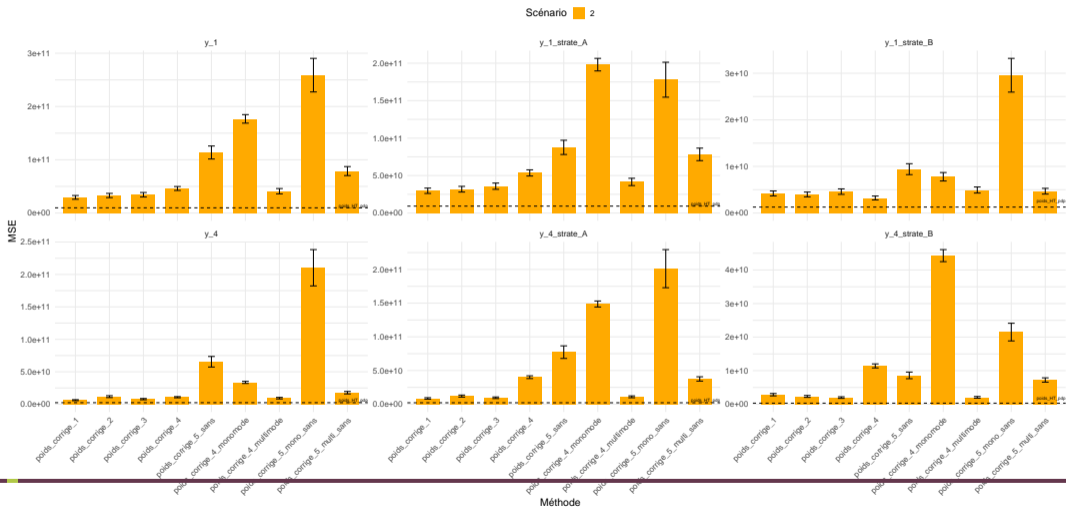
Comparaison des biais par méthode et scénario (avec IC)





Résultat 2 : EQM

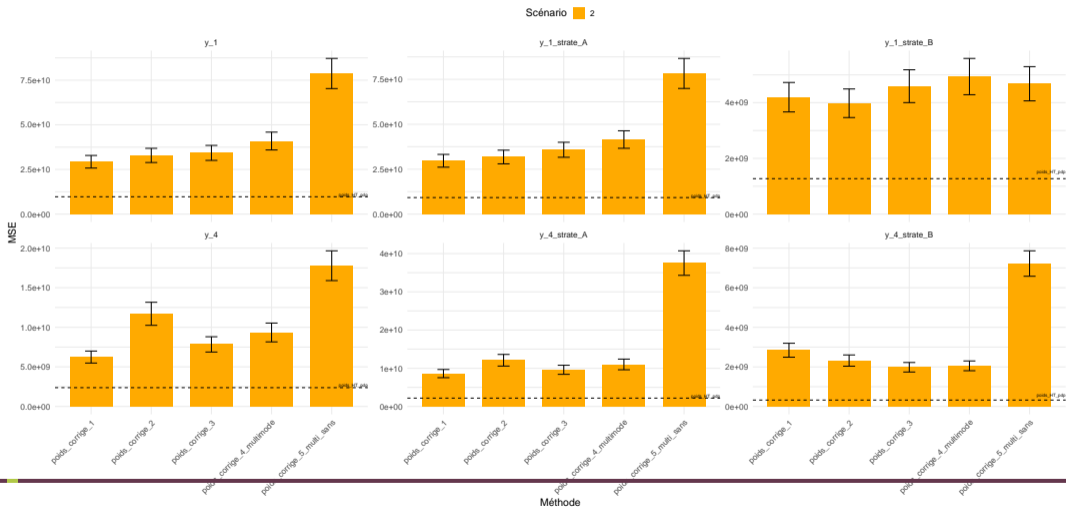
Comparaison des mse par méthode et scénario (avec IC)





Résultat 2 restreints : EQM

Comparaison des mse par méthode et scénario (avec IC)





Merci de votre attention !

