L'utilisation de R-indicateurs pour la priorisation des enquêtes en cours de collecte

Thomas Merly-Alpa

INSEE, DMCSI, Division Sondages

01/04/2014

Sommaire

Problématique abordée

- Baisse locale des taux de réponse.
- Possibilité d'enquêter quelques FA supplémentaires.
- Comment les choisir ?
- Quel impact sur la précision des estimations ?

Les R-indicateurs

Les R-indicateurs

Définition

Définition

Le **R-indicateur** *est une mesure du manque d'association entre réponse et variables auxiliaires :*

$$R(\theta) = 1 - 2S(\theta)$$

avec :

$$S(\theta) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}$$

La propension à répondre d'un individu i est $\theta_i = \mathbb{P}[r_i = 1 \mid s_i = 1]$, où s_i indicatrice d'échantillonnage et r_i de réponse.

Valeurs

On a l'inégalité suivante.

$$S(\theta) \leq \sqrt{\bar{\theta}(1-\bar{\theta})} \leq \frac{1}{2}$$

Le R-indicateur est donc un indicateur compris entre 0 et 1.

- 1 signifie tous les θ_i égaux.
- 0 signifie une dispersion maximale.

À quoi sert le R-indicateur

Le R-indicateur permet :

- De mesurer l'évolution de la représentativité au sein d'une enquête.
- De comparer plusieurs protocoles d'une même enquête.
- De comparer plusieurs enquêtes sur différents sujets, de différents pays et même de différentes tailles.

Estimation

Dans les faits, on estime θ_i par $\hat{\theta}_i$ calculés par un modèle logistique fondé sur des variables qualitatives X. Un estimateur du R-indicateur est alors :

$$\hat{R}(\theta) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

avec :

$$\hat{\bar{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i\frac{s_i}{\pi_i}$$

Lien avec le biais

Le biais de l'indicateur d'Horvitz-Thompson $\hat{\bar{Y}}_{HT}$ vérifie :

$$\left|B(\hat{\bar{Y}}_{HT})\right| \leq \frac{1-\hat{R}(\theta)S(Y)}{2\bar{\theta}}$$

Donc on a pour tout γ :

$$\hat{R}(\theta) \geq 1 - 4\hat{\bar{\theta}}\gamma \Longrightarrow B(\hat{\bar{Y}}_{HT}) \leq \gamma$$

Un R-indicateur suffisamment grand donne une borne sur le biais.

Le R-indicateur partiel inconditionnel

Définition

Le **R-indicateur partiel inconditionnel** *mesure la distance à une réponse représentative pour une variable Z à H modalités :*

$$R_U(Z) = \sqrt{\sum_{h=1}^{H} \frac{N_h}{N-1}(\bar{\theta}_h - \bar{\theta})^2}$$

où N_h est l'effectif de la modalité h, $\bar{\theta}_h$ la proprension à répondre moyenne sur cette modalité.

On l'estime par :

$$\hat{R}_U(Z) = \sqrt{\sum_{h=1}^{H} \frac{\hat{N}_h}{N}(\hat{\bar{\theta}}_h - \hat{\bar{\theta}})^2}$$

Le R-indicateur partiel inconditionnel des modalités

Le R-indicateur partiel inconditionnel relatif à une modalité h de la variable Z est estimé par :

$$\hat{R}_U(Z,h) = \sqrt{\frac{\hat{N}_h}{N}}(\hat{\bar{\theta}}_h - \hat{\bar{\theta}})$$

Cet indicateur peut être positif ou négatif :

- Un R-indicateur partiel inconditionnel négatif signifie que la modalité est sous-représentée.
- Un R-indicateur partiel inconditionnel positif signifie que la modalité est sur-représentée.

Le R-indicateur partiel conditionnel

Definition

Le **R-indicateur partiel conditionnel** mesure la variance due à la variable Z dans chacun des J sous-groupes formés par le croisement des modalités de toutes les autres variables :

$$R_C(Z) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}(\theta_i - \bar{\theta}_j)^2}$$

On l'estime par :

$$\hat{R}_C(Z) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}\frac{s_i}{\pi_i}(\hat{\theta}_i - \hat{\bar{\theta}}_j)^2}$$

Le R-indicateur partiel conditionnel des modalités

Le R-indicateur partiel conditionnel relatif à une modalité h de la variable Z est estimé par :

$$\hat{R}_C(Z,h) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}\frac{s_i}{\pi_i}\mathbf{1}_{z_i=h}(\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

Cet indicateur est toujours positif.

Priorisation

On procède en deux étapes :

1. On choisit les variables ayant les R-indicateurs partiels les plus grands.
2. Parmi les variables sélectionnés, on priorise les groupes associés aux modalités ayant un R-indicateur partiel inconditionnel fortement négatif.

Un exemple simple

Cas d'école : enquête avec deux relances.
On simule un sondage simple dans une population de 10.000 personnes, répartie en hommes et femmes et en 3 groupes d'âge. On suppose que les hommes sont trois fois moins enclins à répondre, mais que l'âge n'a aucune influence.

Collecte sans priorisation

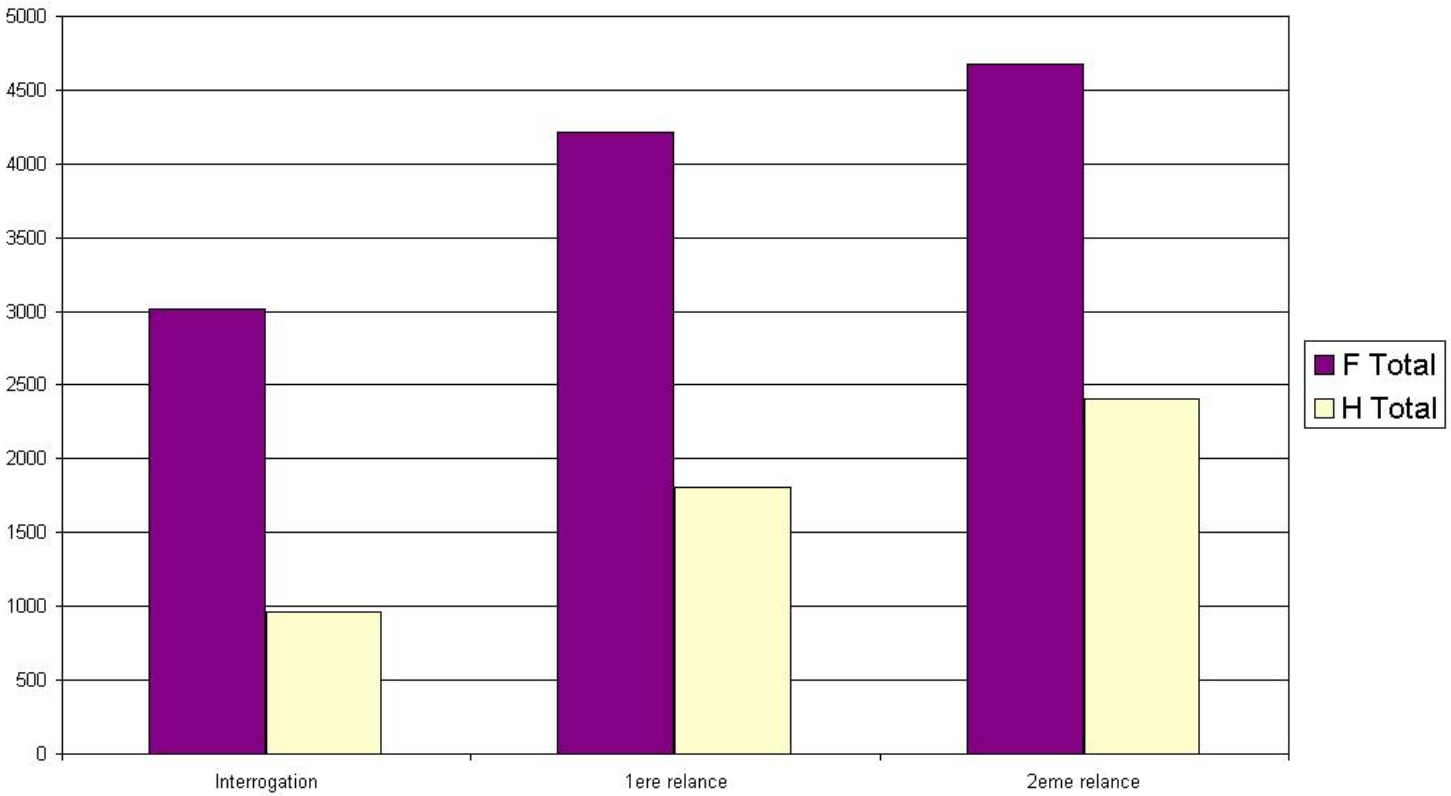


FIGURE: Répartition hommes/femmes des répondants

R-indicateurs partiels sans priorisation

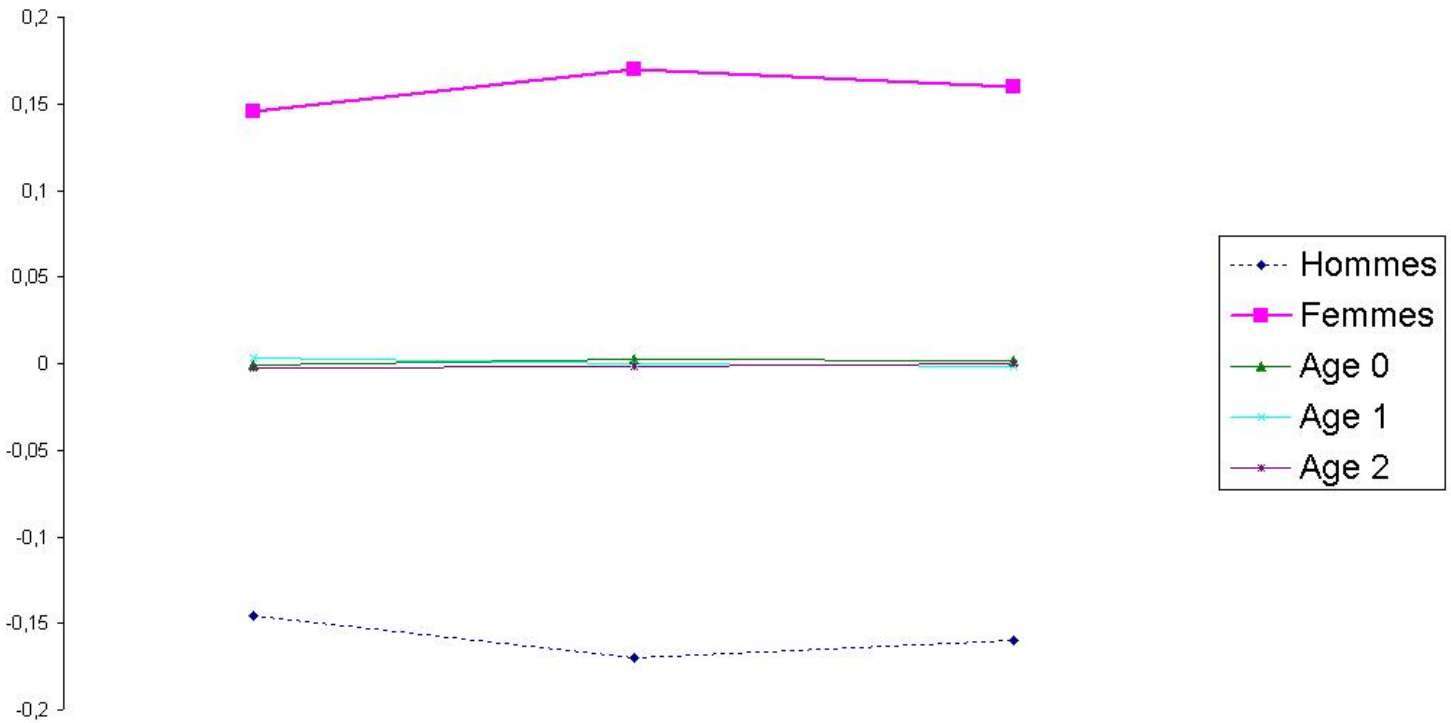


FIGURE: R-indicateurs partiels inconditionnels

Comment améliorer la représentativité ?

La représentativité évolue peu au cours de l'enquête :

R-indicateur	Début	Milieu	Fin
Sans priorisation	0.588	0.518	0.547

Les R-indicateurs partiels indiquent qu'il faudrait prioriser les hommes dès la première relance. Observons l'effet d'une telle priorisation.

Collecte avec priorisation

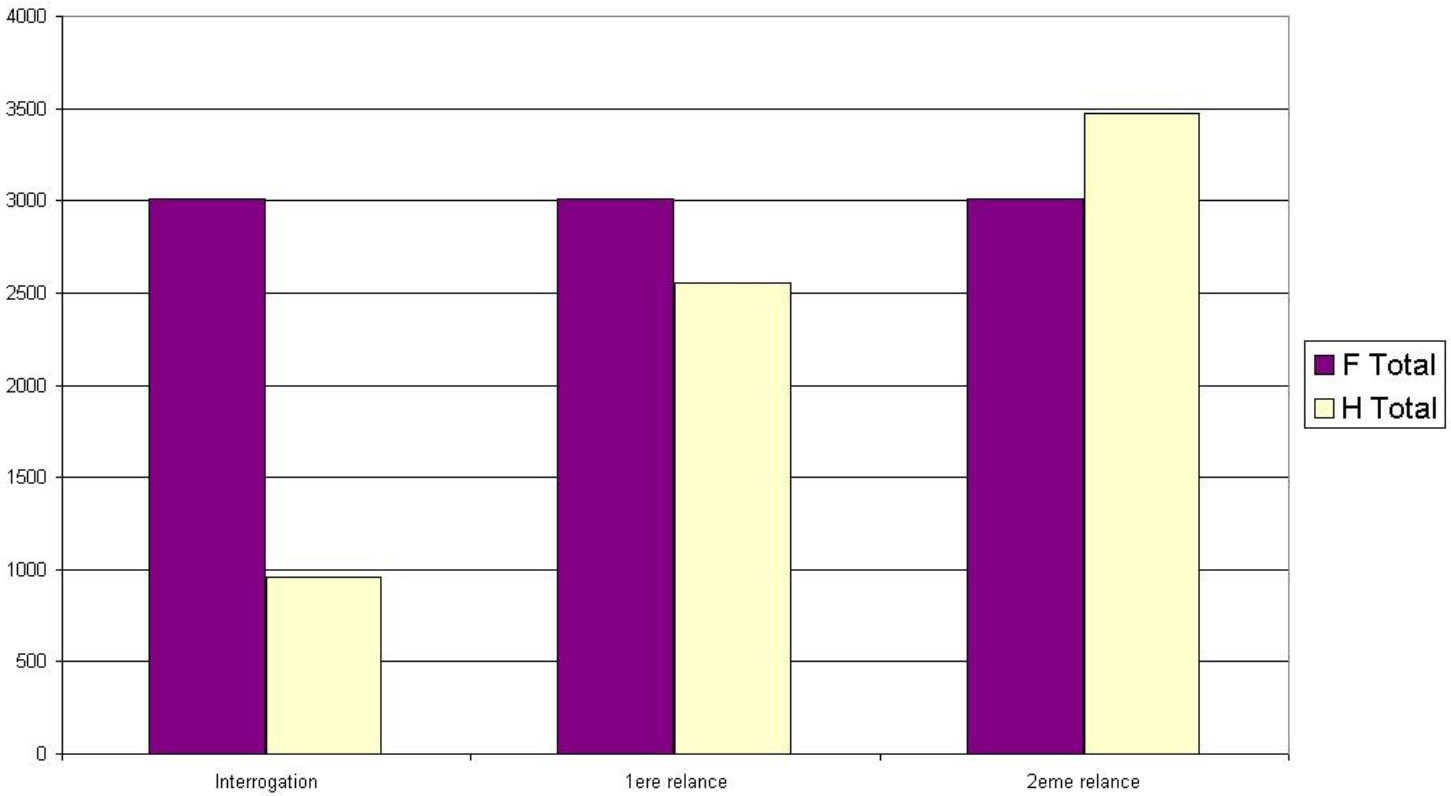
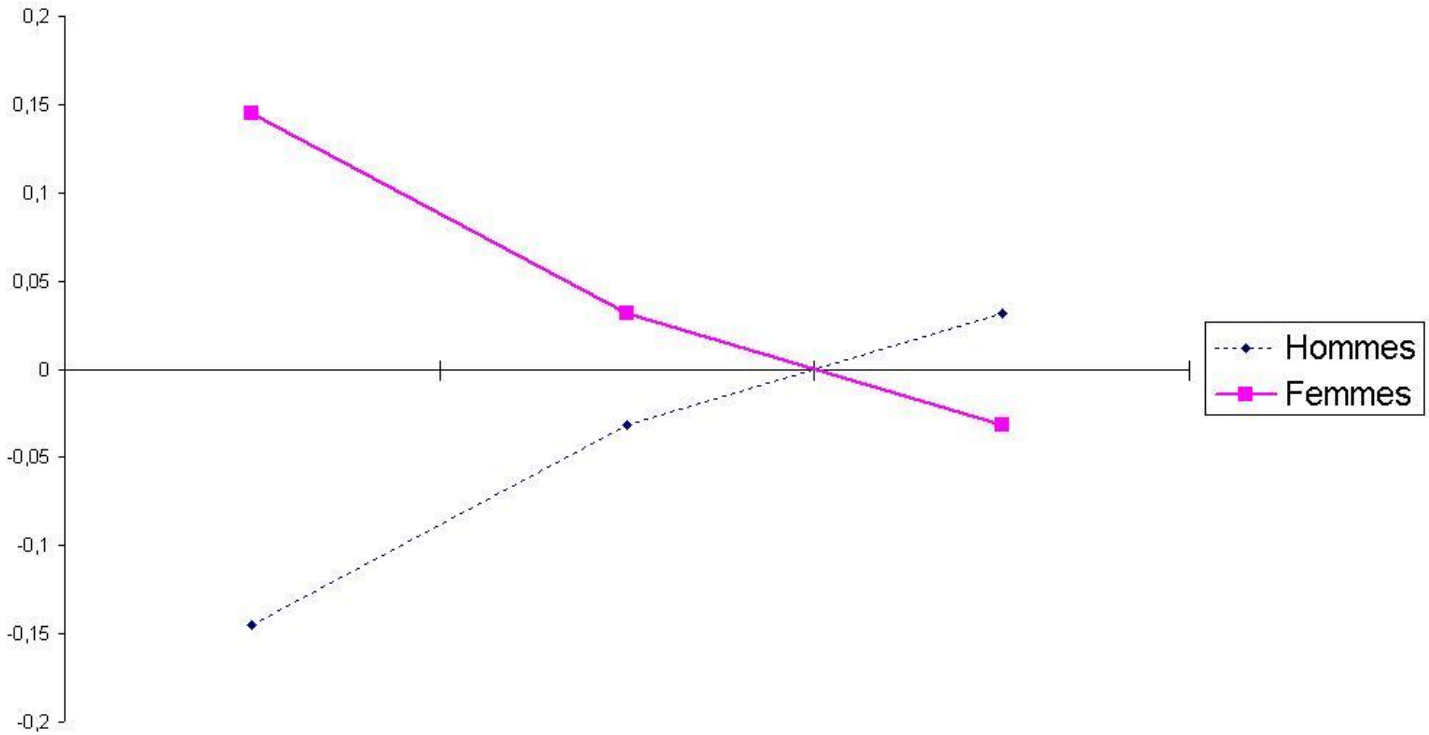


FIGURE: Répartition hommes/femmes des répondants

R-indicateurs avec priorisation

R-indicateur	Début	Milieu	Fin
Avec priorisation	0.588	0.909	0.906



Avantages

- Permet un traitement au cours de la collecte.
- Plus précis que les taux de réponse partiels.
- Facilite le calage.
- Pas lié à la variable d'intérêt...

Inconvénients

- Variance et biais du R-indicateur.
- Repose sur le modèle de non-réponse.
- Pas lié à la variable d'intérêt.

Application à l'enquête Patrimoine 2010

L'enquête Patrimoine 2010

L'enquête

L'enquête Patrimoine est une enquête répétée tous les 6 ans depuis 1986 qui vise à étudier le patrimoine moyen des Français, leur comportement vis à vis de ce patrimoine (transmissions, achats...) en lien avec leur situation personnelle et professionnelle. La dernière enquête date de 2010.
21000 ménages ont été enquêtés. Le taux de réponse a été de l'ordre de 68%.

Les échantillons

En plus de l'échantillon standard, un échantillon de hauts revenus a été utilisé. Chacun de ces échantillons est stratifié :

- Pour l'échantillon standard, en 6 strates : agriculteurs, indépendants, cadres, personnes ayant un revenu du patrimoine, personnes âgées, et le reste de la population.
- Pour l'échantillon non-standard, en 4 strates : riches urbains, personnes possédant un patrimoine élevé à dominante mobilière, ceux à dominante immobilière, et les autres.

Traitement des données

Le traitement des données post-collecte s'est fait en trois étapes :

- Calcul des probabilités de réponse avec un modèle logistique.
- Correction de la non-réponse par GRH (variante méthode des scores).
- Calage.

R-indicateurs totaux

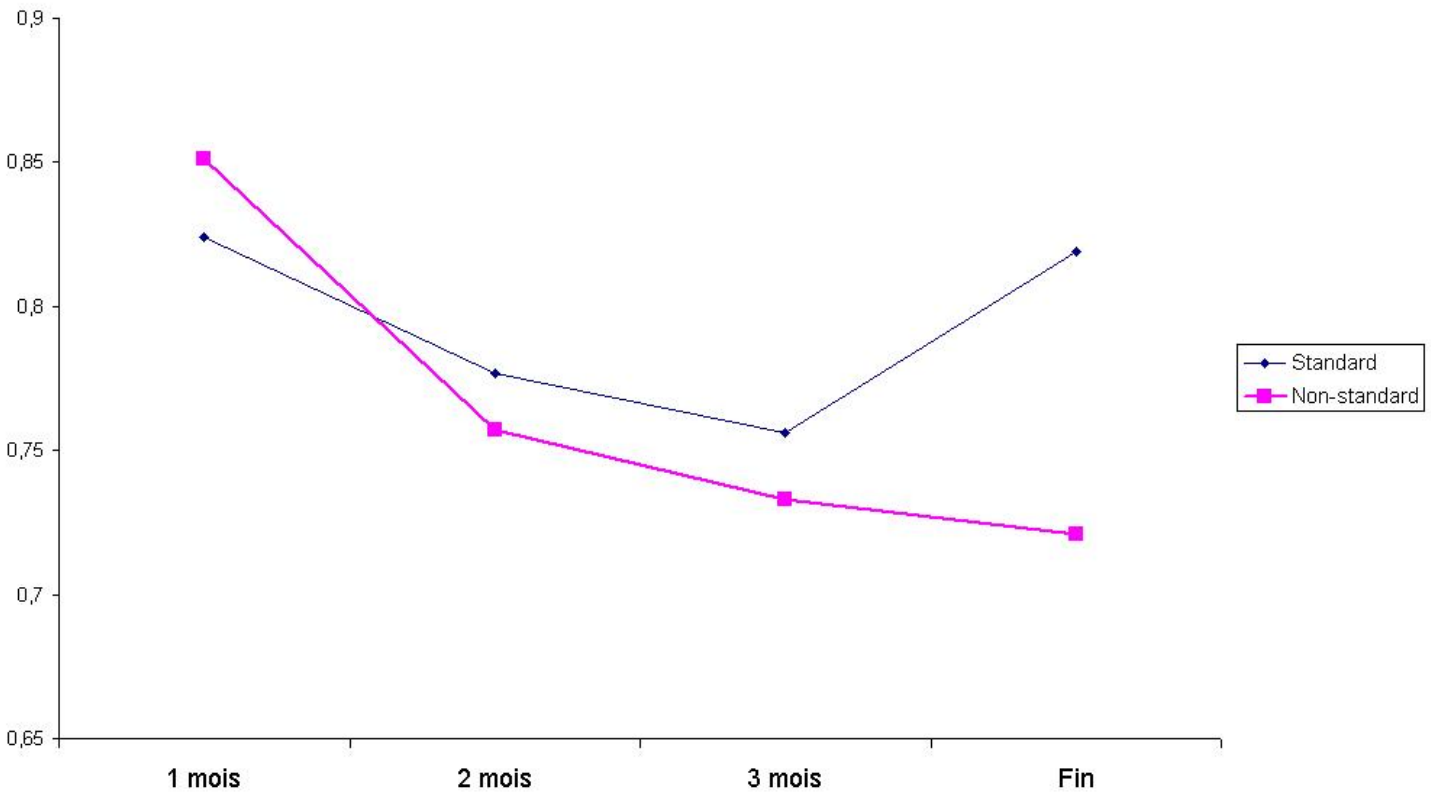


FIGURE: R-indicateurs totaux dans l'enquête Patrimoine 2010

R-indicateurs partiels inconditionnels

Variable	1 mois	2 mois	3 mois	Fin
Strate	3.52	6.85	8.65	3.05
Agriculteurs	-0.14	1.07	1.40	0.92
Personnes Âgées	-0.74	-3.03	-4.61	-1.55
Indépendants	-1.67	-2.55	-2.60	-0.89
Revenus pat.	-1.00	-1.49	-1.67	0.07
Cadres	-2.01	-3.01	-3.16	-1.50
Autres	2.00	4.34	5.67	1.73
Type de logement	1.07	1.95	1.93	3.17
Appartement	-0.82	-1.48	-1.47	-2.42
Maison	0.69	1.26	1.25	2.05

Au troisième mois

Standard	Inconditionnel	Conditionnel
Strate	8.67	1.88
Type de ménage	4.08	0.52
Gros Revenus Act.	2.40	0.18
Âge	3.12	0.25
HLM	1.67	0.26
Revenus Patrimoine	3.71	0.30
Surface	1.78	0.21

Non-standard	Inconditionnel	Conditionnel
Strate	7.39	0.275
Type de ménage	7.89	0.337
Très Gros Revenus Pat.	6.66	0.256

Taux de réponse CVS 2013

Enquête CVS (Cadre de Vie et Sécurité) 2013 :

- Baisse générale des taux de réponse.
- Quelques zones très affectées.

Comment compenser cet effet ? Application à des variables très dispersées (patrimoine brut...)

Simulations de la baisse des taux de réponse

Pour chaque ZAE, on calcule le ratio de sélection à partir des taux de réponse de la façon suivante :

$$p_k = \frac{2\,TR_{2013}}{TR_{2012} + TR_{2011}}$$

On simule 100 scénarios de diminution ; pour chacun d'eux on conserve p_k % des répondants à 3 mois. Les R-indicateurs moyens sont alors :

Standard	**Non Standard**
0.7610	0.6925

Dispersion des ratios de sélection

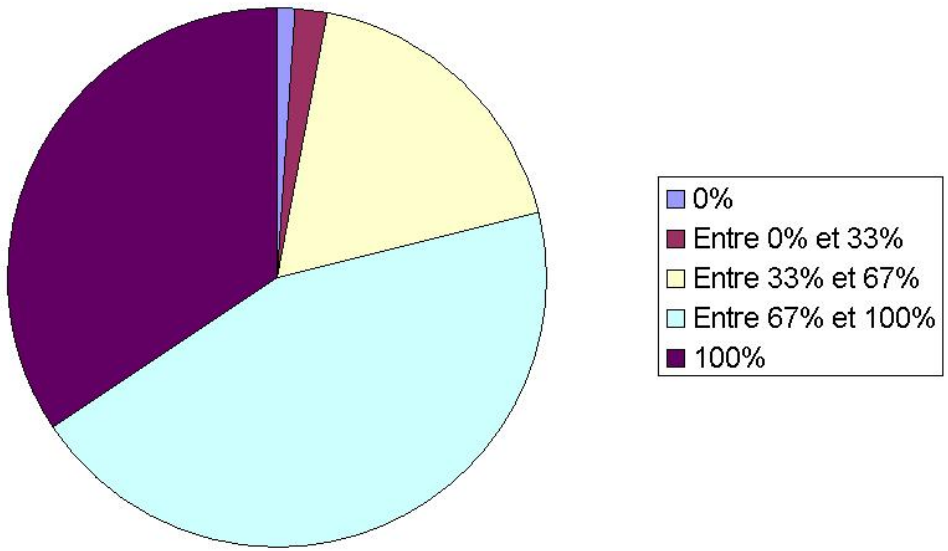


FIGURE: Dispersion des ratios de sélection des FAs dans les ZAE.

Groupes priorisés

En recalculant les R-indicateurs partiels sur l'échantillon réduit, on obtient les groupes suivants :

Échantillon standard	**Échantillon non-standard**
Personnes Âgées	Riches urbains
Cadres	Appartements
Indépendants à hauts revenus	Individus seuls
Individus seuls	
Appartements	

L'échantillon non-standard sera pris en premier pour la priorisation.

Rajout de FA

On considère deux méthodes de rajout de FA :

- La **sélection aléatoire**, qui consiste à choisir n FA supplémentaires dans l'échantillon non-standard, et s'il n'y en a pas un nombre suffisant de compléter aléatoirement dans l'échantillon standard.
- La **priorisation**, qui consiste à choisir parmi les groupes priorisés de l'échantillon non-standard n FA supplémentaires, et s'il n'y en a pas un nombre suffisant de compléter dans les groupes de l'échantillon standard.

Scénarios de priorisation

On considère plusieurs scénarios de rajout de n FA.

1. On peut réaliser ce rajout dans **la totalité des ZAE**.
2. On peut se limiter à celles dont la perte est d'au moins 33% des FA, ce qui fait se concentrer sur **environ 21% des ZAE**.
3. On peut aussi se limiter à celles qui ont été les plus atteintes et dont la perte est au moins de 66%, ce qui restreint le champ à **environ 3% des ZAE**.

Impact sur le R-indicateur

Nombre de FA	**Échantillon std**		**Échantillon non-std**	
Base	0.761		0.693	
Scénario 1	Priorisé	Aléatoire	Priorisé	Aléatoire
5	0.767	0.767	**0.702**	0.684
10	0.750	**0.770**	**0.703**	0.677
Scénario 2	Priorisé	Aléatoire	Priorisé	Aléatoire
5	**0.776**	0.771	**0.712**	0.697
10	**0.791**	0.781	**0.726**	0.703
Scénario 3	Priorisé	Aléatoire	Priorisé	Aléatoire
5	0.765	**0.767**	**0.711**	0.697
10	**0.773**	0.771	**0.720**	0.704

Impact sur le R-indicateur

Conclusions

- Dans l'échantillon non-standard, la priorisation est meilleure.
- Plus mitigé dans l'échantillon standard.
- Le scénario 1 n'est pas concluant ; on peut faire mieux avec moins.
- Le scénario 2 est le meilleur, indépendamment des contraintes de coût.

Précision des estimations

Nous nous intéressons aux variables d'intérêt suivantes : patrimoine brut moyen, net, financier, immobilier et professionnel.
La précision des mesures se décompose :

$$V_{\text{totale}} = V_{\text{Pat10}} + V_{\text{BaisseTR}}$$

On suppose V_{Pat10} fixe et on s'intéresse à V_{BaisseTR}.

Mise en oeuvre de la priorisation

On suit le scénario 2 et **on se concentre sur 20% des ZAE**. Dans les autres zones, on suppose que **la collecte continue à hauteur de 75%** (effort constant) :

- 25% des enquêteurs restants sont affectées à la priorisation des zones ciblées.
- Ils peuvent y réaliser 5 ou 10 FA (selon les difficultés de remplacement).

Hypothèse simplifiée de modélisation : ne prend pas en compte la géographie.

Nombre de répondants

Cette opération réduit le nombre de répondants :

n	Collecte terminée	Aléatoire	Priorisé
5	10600	10334	10305
10	10600	10529	10506

Impact sur la précision

On obtient les dispersions suivantes :

	Brut	Net	Fin	Imm	Prof
Collecte terminée	2175	2179	536	699	1904
Aléatoire, $n = 5$	2436	2270	539	766	1983
Priorisé, $n = 5$	2521	2337	552	741	2059
Aléatoire, $n = 10$	2327	2132	589	731	1870
Priorisé, $n = 10$	1967	1771	592	676	1504

Conclusion

- Prioriser selon les R-indicateurs semble utile.
- Utiliser des méthodes de priorisation dans toutes les zones est contreproductif.
- En ciblant les zones, on a moins de répondants mais une meilleure précision pour un même effort.

Pour finir...

Merci pour votre attention !