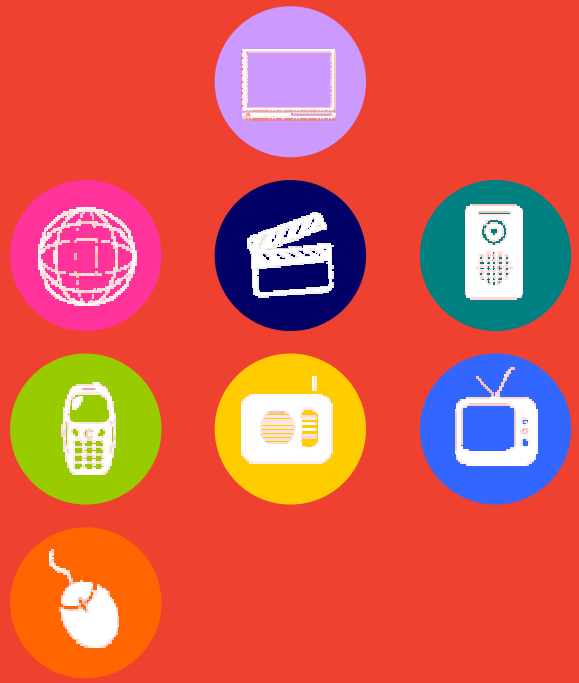
Calage sur composantes principales versus calage pénalisé

Application à la mesure d'audience hybride Internet



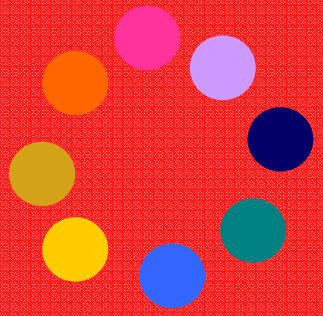
Colloque Francophone sur les sondages – Rennes 2012

mediametrie

Sommaire

mediametrie

1 – Problématique de la mesure hybride Internet Fixe

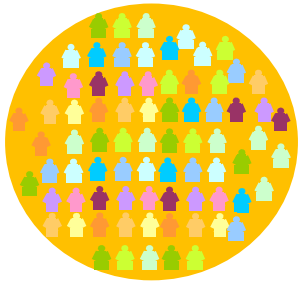


mediametrie

Problématique

Co existence de 2 mesures Internet fixe

➢**Une mesure fondée sur un panel de 25 000 individus**



> Installation d'un meter qui enregistre l'ensemble de l'activité Internet par individu

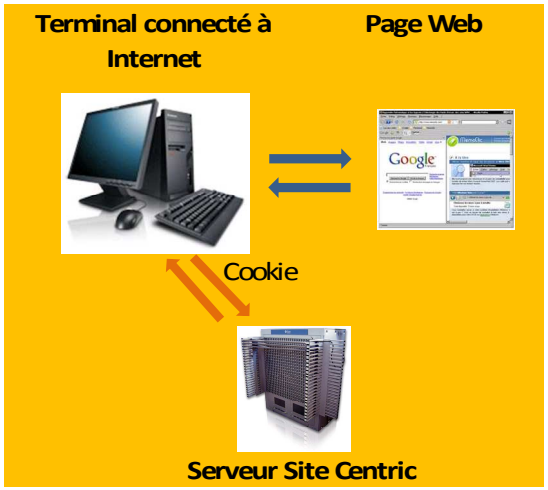
> Des résultats estimés et qualifiés pour l'ensemble des sites (visiteurs uniques, pages vues, temps passé, nombre de visites)

mediametrie

Problématique

Co existence de 2 mesures Internet fixe

➢**Une mesure fondée sur l'insertion d'un tag dans le code source**



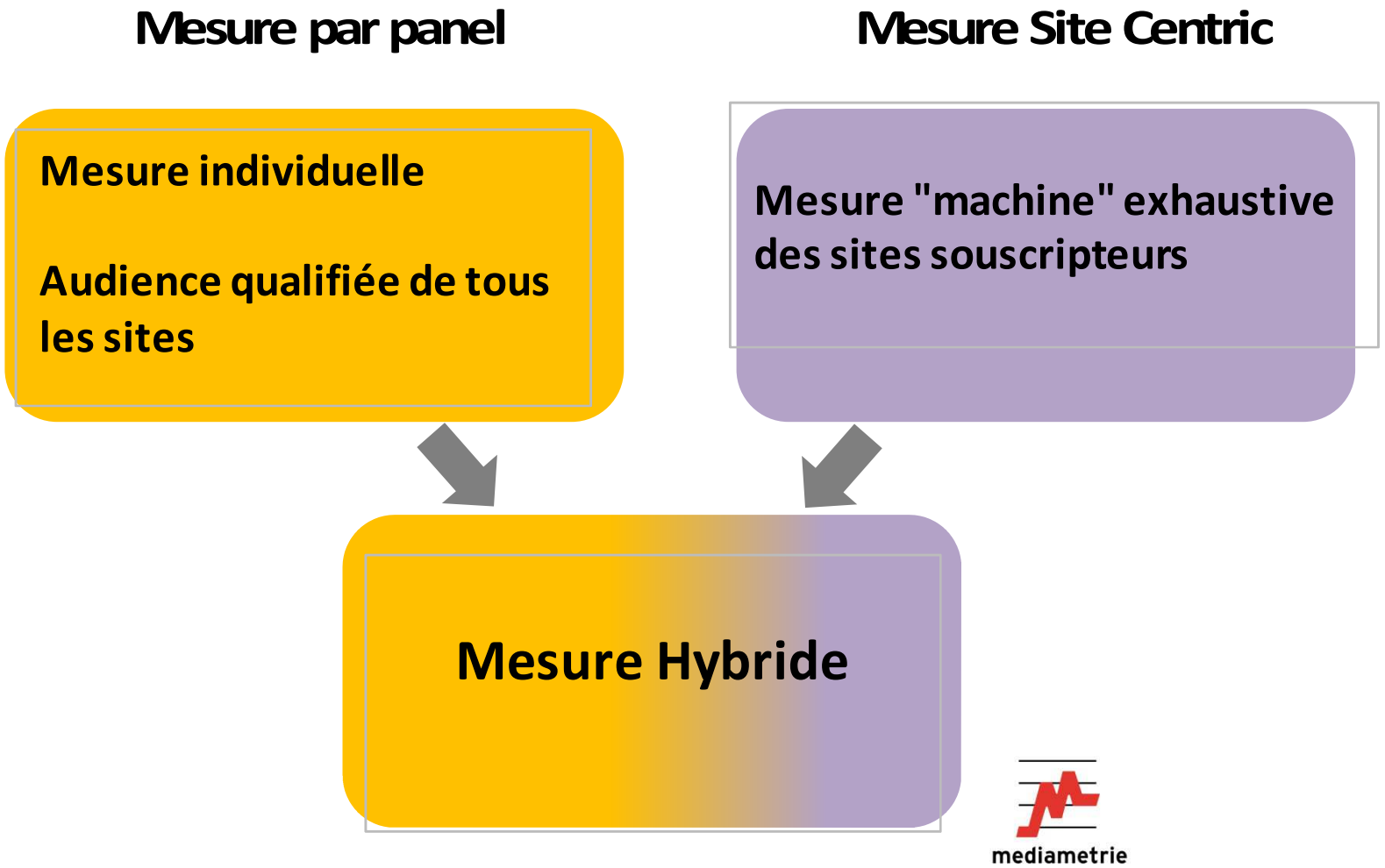
> Déploiement d'un tag sur chaque page

> Des résultats exhaustifs pour les sites souscripteurs

mediametrie

Problématique

Mesure hybride : prendre le meilleur des 2 mesures

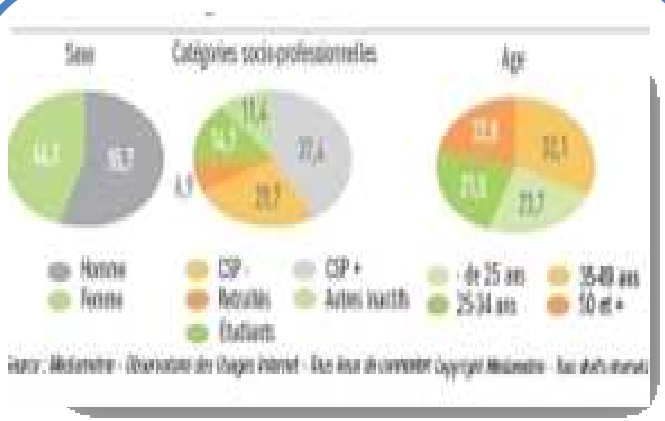


mediametrie

Problématique

Une approche par redressement

Objectifs Sociodémographiques



Le profil du panel est calé sur les résultats issus d'une enquête de cadrage (sexe, âge, CSP, etc.)

Objectifs Site Centric

	Visites	Pages
Site 1		
Site 2		
Site 3		
..		
Site N		

Les résultats Site Centric ne sont connus que pour environ 400 sites

Jeu de poids unique assurant les deux groupes de contraintes

(qualitatif pour les contraintes sociodémographique, quantitatifs pour les données Site Centric

> L'introduction de contraintes Site Centric modifie les poids de redressement de l'ensemble des individus : tous les sites bénéficient ainsi de l'apport de ces résultats complémentaires.

mediametrie

Problématique

Les données disponibles

> Données individuelles du panel

Variables catégorielles

Variables numériques (nombre de visites)

	Sexe	Age	…	CSP	Site 1	Site 2	…	Site M
Ind 1	H	19	…	Etudiant	0	12	…	143
Ind 2	F	46	…	CSP+	0	0	…	0
Ind 3	F	35	…	CSP-	0	1	…	0
…	…	…	…	…	…	…	…	…
Ind n	H	67	…	Retraité	3	2	…	0

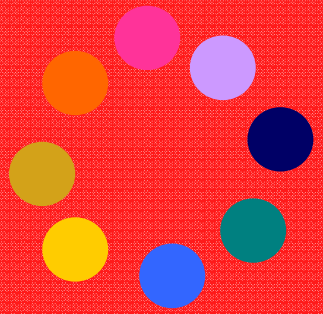
> Objectifs de redressement

36 Variables catégorielles		
Sexe	H	49%
	F	51%
Age	2 - 17 ans	23%
	18 - 24 ans	10%
	25 - 34 ans	14%
	35 - 49 ans	23%
	50 - 64 ans	20%
	65 ans et +	11%
…		

157 Variables numériques	
Site 1	13 620 324
Site 2	1 653 357
Site 3	11 963 627
Site 4	27 829 720
Site 5	1 013 487
Site 6	48 852 438
Site 7	2 013 454
…	
Site 157	35 852 123

mediametrie

2 - Présentation du calage pénalisé



mediametrie

Calage classique

Trouver des poids :

- qui se trouvent le plus proche possible (au sens d'une distance) des poids d'échantillonnage d_k
- qui estiment exactement le total des variables de calage

Avec la distance de chi-deux, les poids satisfont:

$$w_k = \arg\min \sum_s \frac{(w_k - d_k)^2}{d_k}$$

$$\hat{t}_{w,X_j} = t_{X_j},\ j = 1, \ldots, p$$

mediametrie

Le sur-calage

En présence d'un trop grand nombre de variables:

> Les poids de calage peuvent être négatifs ou trop grands

> Les rapports de poids satisfont difficilement des bornes

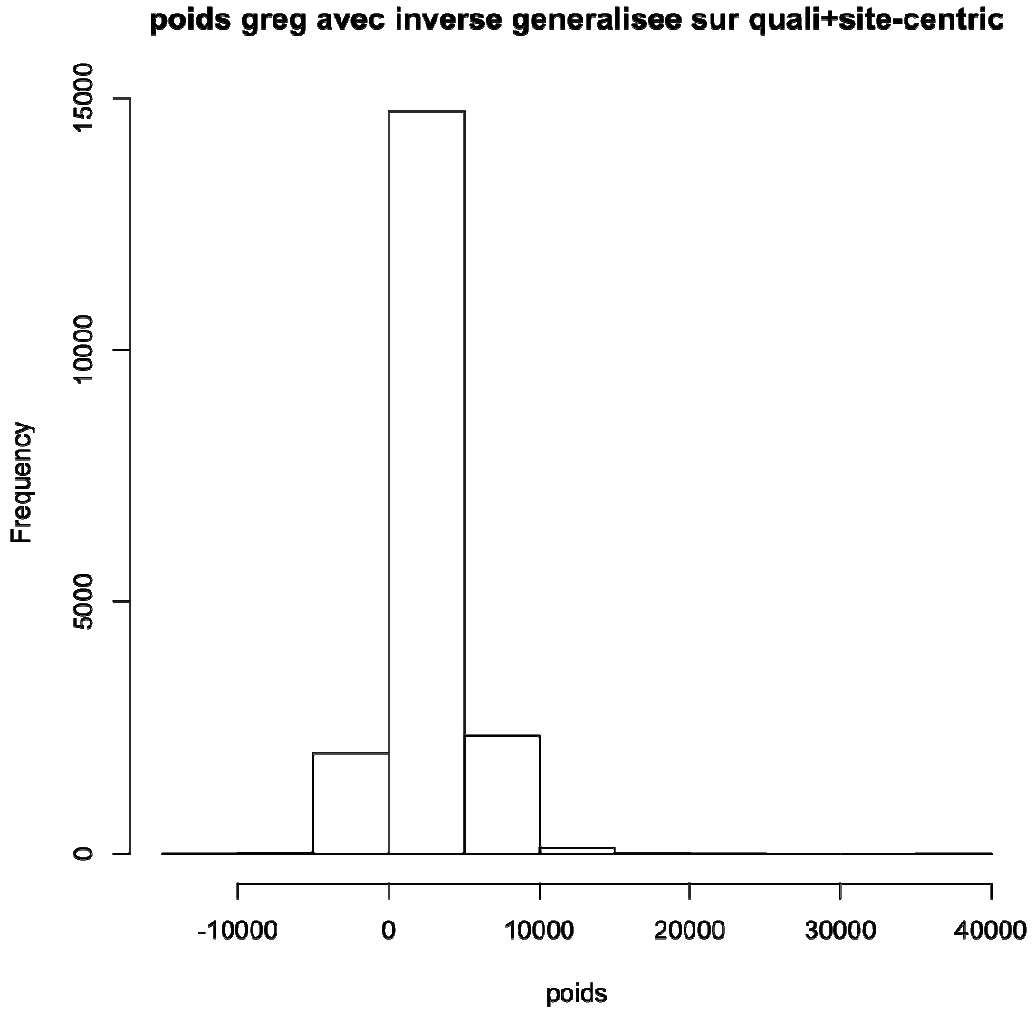
$$L \leq \frac{w_k}{d_k} \leq U$$

> Utiliser trop de variables de calage peut augmenter la variance (*Silva and Skinner, 1997*)

→ **Relaxer les contraintes de calage**

mediametrie

Le sur-calage



mediametrie

Calage pénalisé

(*Bardsley and Chambers, 1984; Rao and Singh, 2009; Beaumont and Bocci, 2008*)

On se donne une tolérance pour chaque contrainte:

$$\frac{\left| t_{w,X_j} - t_{X_j} \right|}{t_{X_j}} \leq \delta_j$$

Cela revient à chercher des poids qui satisfont

$$w = \arg\min \sum_s \frac{(w_k - d_k)^2}{d_k} + \frac{1}{\lambda} \sum_{j=1}^{p} C_j \left(t_{w,X_j} - t_{X_j} \right)^2$$

où C_j est le coût de ne pas satisfaire la j ème contrainte

mediametrie

Poids du calage pénalisé et matrice des coûts

Les poids du calage pénalisé sont donnés par :

$$w_k = d_k + d_k x'_k \left(\sum_s d_k x_k x'_k + \lambda C^{-1} \right)^{-1} (t_X - \hat{t}_{dX})$$

La matrice des coûts est définie par :

$$C = diag(C_1, \ldots, C_p)$$

$\lambda C_j^{-1} = 0$ alors la j-ème contrainte est enlevée;

$\lambda C_j^{-1} = \infty$ alors la j-ème contrainte est exacte;

mediametrie

Choix des paramètres de pénalisation

Tolérances « versus » coûts

• Il y a une relation entre les coûts et les tolérances: pour des tolérances choisies, les coûts peuvent être déterminés et vice-versa;
• Une proposition est de prendre comme coûts, les inverses des totaux des variables;
• On peut standardiser les variables et prendre pour tout j, $C_j = 1$

•**Paramètre** λ : plus difficile à déterminer

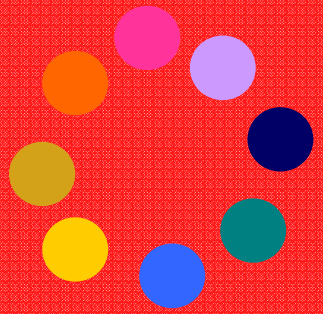
> La trace ridge (*Bardsley and Chambers, 1986*)

> L'algorithme de la bissection (*Beaumont and Bocci, 2008*)

> La validation croisée

mediametrie

3 - Présentation du calage sur composantes principales



mediametrie

Calage sur composantes principales

On réduit le nombre de variables auxiliaires de départ en gardant le maximum d'information et on fait un calage sur les nouvelles variables auxiliaires ainsi créées

On considère les premières **r** composantes principales correspondantes aux **r** plus grandes valeurs propres de la matrice

$$T_s = \sum_s d_k x_k x'_k$$

→ **On utilise comme variables de calage les r composantes principales**

mediametrie

Calage sur composantes principales

Les totaux des variables de départ ne sont pas estimés exactement; nous réalisons de cette façon un relâchement des équations de calage.

Le nombre de composantes principales est choisi selon des méthodes empiriques.

Cette méthode présente l'avantage que les nouvelles variables de calage peuvent être utilisées directement dans Calmar.

mediametrie

Calage partiel

Souvent, il y a des variables auxiliaires pour lesquelles on veut retrouver les totaux exacts à partir de l'échantillon (CSP, sexe, etc.)

On peut modifier l'approche proposée (*Bardsley and Chambers, 1986*) pour répondre à cette question

→ on partage les variables auxiliaires dans deux groupes:

- les variables pour lesquelles on veut avoir calage exact (en faible nombre)
- les autres variables

$$X = (\tilde{X}_1, \tilde{X}_2)$$

Il y a p_1 variables dans le premier bloc et $p - p_1 = p_2$ dans le deuxième

mediametrie

Calage partiel

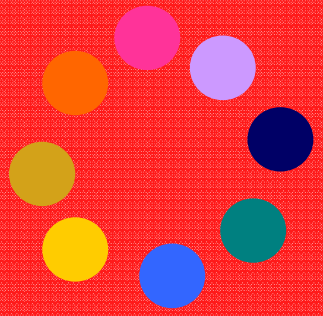
- **Pour le calage pénalisé ou ridge**: matrice des coûts C donnée par

$$C = diag\left(\underbrace{0}_{}, \underbrace{\cdots}_{p_1}, \underbrace{0}_{}, \underbrace{C_1}_{}, \underbrace{\cdots}_{p_2}, \underbrace{C_{p_2}}_{} \right)$$

- **Pour le calage sur CP**: variables de calage données par les $\tilde{X}_1$ et les r_2 composantes principales de $\tilde{X}_2$ et orthogonales à $\tilde{X}_1$

mediametrie

4 – Mise en œuvre et comparatif des performances



mediametrie

Mise en œuvre ACP

Deux tests sont réalisés :

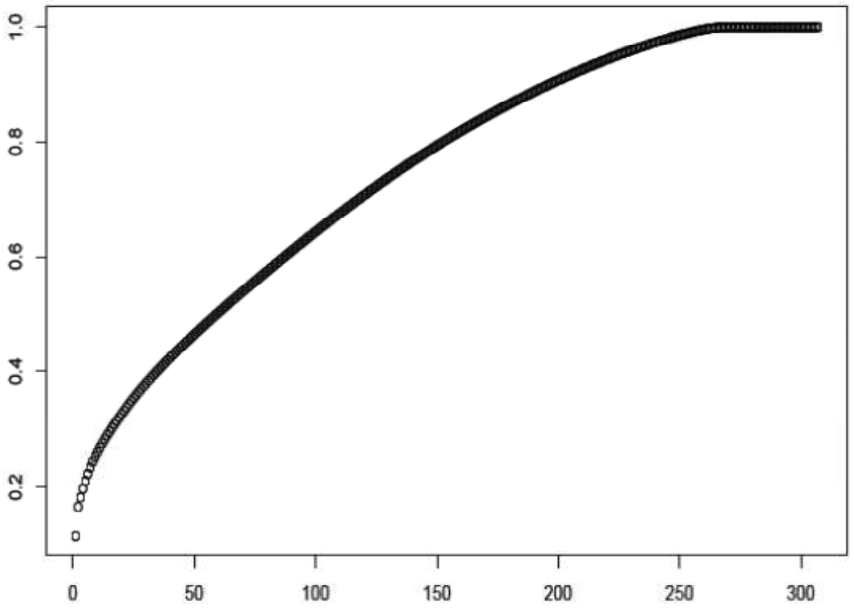
- Test 1 : on réalise l'ACP sur les variables sociodémographiques (données qualitatives) et les données Site Centric (données quantitatives)

- Test 2 : l'ACP n'est mise en œuvre que sur les données Site Centric

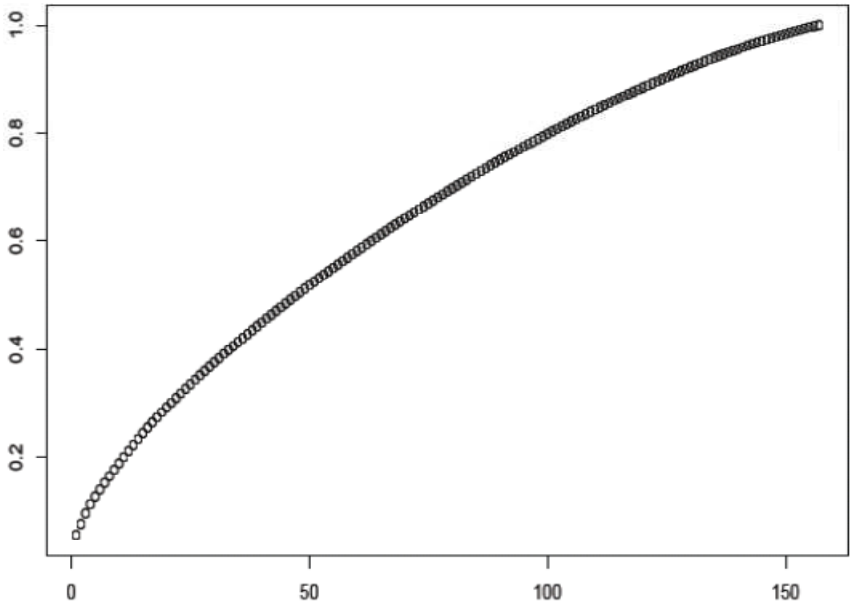
Mise en œuvre ACP

Pourcentage de la variance totale en fonction du nombre de valeurs propres :

Test 1 : quali + quanti



Test 2 : quanti



mediametrie

Mise en œuvre ACP

Test 1 : quali + quanti

Le nombre de valeurs propres introduit est testé itérativement.

→ A partir de 7 composantes principales, on obtient des poids de redressement négatifs avec la distance du Chi 2

→ Les résultats présentés ci-après tiennent ainsi compte des 6 premières composantes principales soit 22,3% de la variance totale.

mediametrie

Mise en œuvre Ridge

On fixe les bornes suivantes pour les rapports de poids :

L = 0.5

U = 5

Dans ce cas aussi, deux tests sont réalisés :

- **Test 1 : Ridge Brut**

On réalise le calage pénalisé sur les variables non standardisées.

→ Coûts = Inverse des totaux

- **Test 2 : Ridge standardisé**

Les variables quantitatives sont standardisées.

→ Coûts = 1

mediametrie

Comparatif des performances

Ecarts relatifs absolus - variables quantitatives

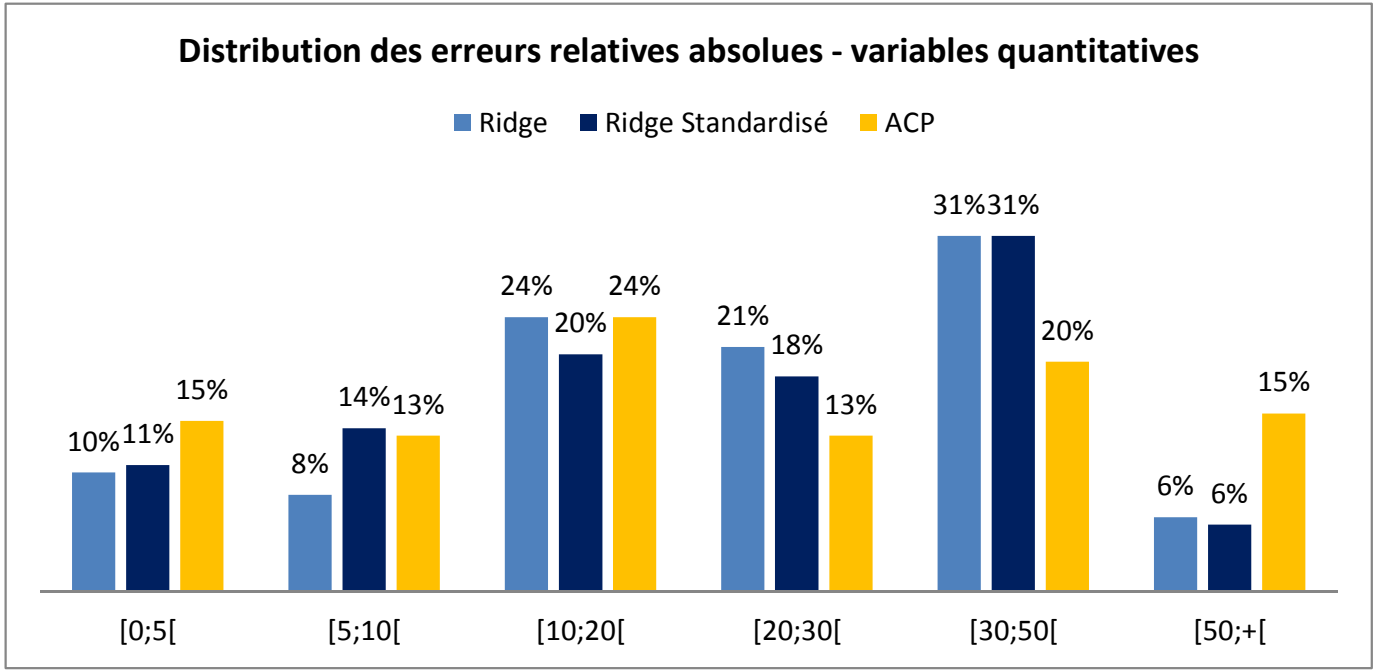
Variable	Min	Q1	Médiane	Moyenne	Q3	Max
Ridge	0,7%	12,1%	24,7%	**25,5%**	37,4%	75,6%
Ridge Standardisé	0,3%	11,6%	21,6%	**24,4%**	35,1%	86,6%
ACP	0,1%	9,3%	19,2%	**26,7%**	37,2%	115,0%

- Les performances des 3 méthodes sont comparables avec une erreur moyenne proche de 25% sur les variables quantitatives.

mediametrie

Comparatif des performances

Ecarts relatifs absolus - variables quantitatives



- La méthode ACP est celle qui maximise les écarts de moins de 10%. Elle a en contrepartie la plus grande part d'erreur de plus de 50%.

mediametrie

Comparatif des performances

Ecarts relatifs absolus - variables qualitatives

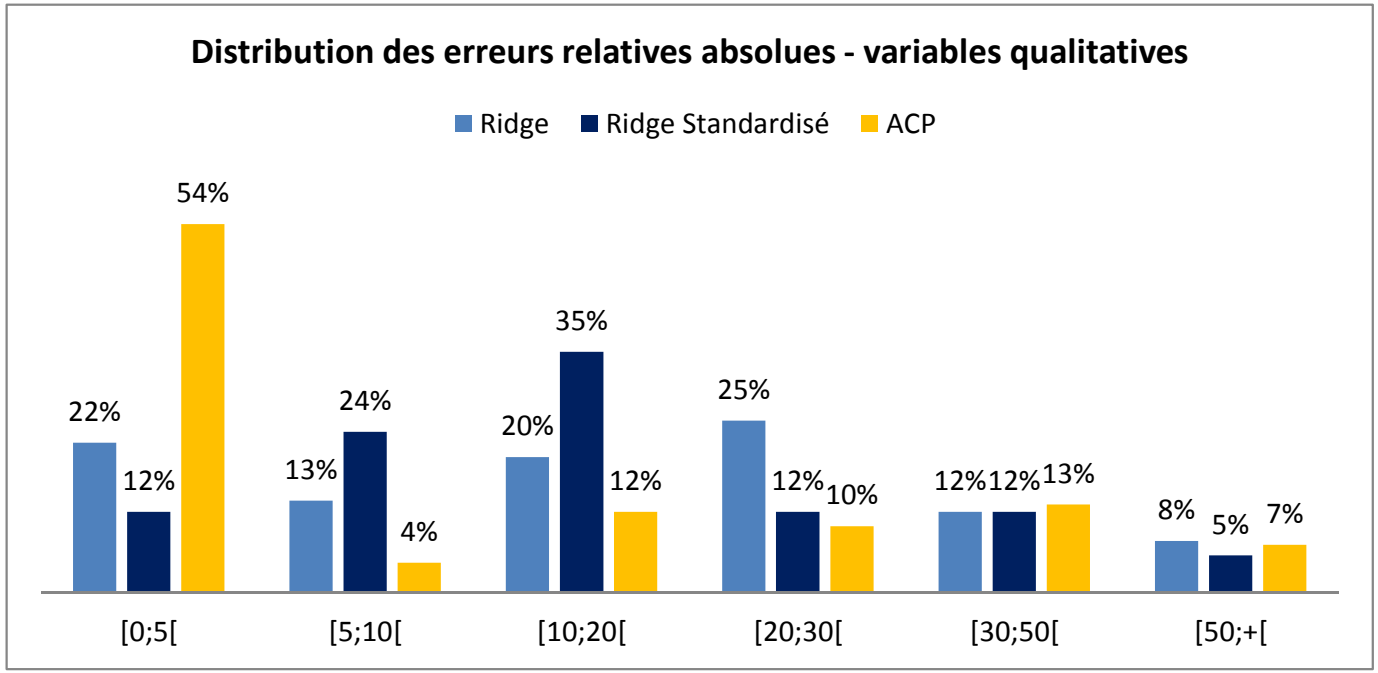
Variable	Min	Q1	Médiane	Moyenne	Q3	Max
Ridge	0,0%	4,8%	19,3%	**22,2%**	23,7%	252,3%
Ridge Standardisé	0,4%	6,2%	13,5%	**19,2%**	24,5%	206,2%
ACP	0,4%	3,3%	4,6%	**17,5%**	24,7%	196,8%

- La méthode ACP est celle qui minimise les erreurs sur les variables qualitatives avec une erreur moyenne de 17,5%.

mediametrie

Comparatif des performances

Ecarts relatifs absolus - variables qualitatives



- La méthode ACP est celle qui maximise les écarts de moins de 10%.

mediametrie

Comparatif des performances

Qualité des poids de redressement

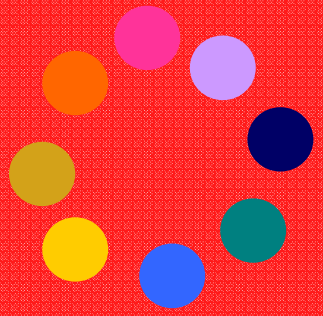
Méthode	n	Poids Min	Poids Max	Rapport de poids	Efficacité du redressement	VE
Ridge		1 303	3 965	3,04	99,89%	50 008 874
Ridge Standardisé	19197	1 302	5 408	4,15	98,76%	47 129 437
ACP		424	6 910	16,30	96,85%	50 000 000

- L'efficacité du redressement est comparable selon les 3 méthodes et proche de 100 : les poids ne sont guère éloignés des poids d'échantillonnage. A titre de comparatif, la méthode Calmar – sinus hyperbolique - obtient une efficacité bien moindre de 48%.

- L'étendue des poids est plus importante avec la méthode ACP à 16,3. Avec Calmar, le rapport de poids est beaucoup plus important à 40.

mediametrie

5 – Bilan et perspectives



mediametrie

Bilan

La mesure Hybride Internet Fixe repose sur l'insertion dans le redressement de :

- 36 variables sociodémographiques et comportementales classiques
- Plus de 150 variables quantitatives issues de la mesure Site Centric

Deux méthodes de redressement alternatives à la méthode Calmar actuellement utilisée ont été testées :

- La méthode Ridge (en standardisant ou non les variables) qui permet d'introduire une tolérance et/ou un coût sur l'atteinte des objectifs de redressement
- La méthode ACP qui permet de résumer les contraintes

Dans tous les cas, on constate un net gain dans la qualité des poids de redressement vs Calmar sinus hyperbolique lié cependant à une erreur relative moyenne élevée de 25% pour les variables quantitatives.

Perspectives

En l'état, les méthodes testées ne répondent pas suffisamment aux attentes de Médiamétrie.

Des travaux complémentaires vont être menés prochainement :

- sur la méthode ACP afin d'intégrer davantage d'information auxiliaire
- sur la méthode Ridge et la définition des tolérances et coûts introduits
- Autres méthodes

mediametrie