# Calage sur information auxiliaire incertaine : proposition d'algorithme de redressement ridge

Flavien Alleaume & Lorie Dudoignon

JdS 2013 – Toulouse - 29/05/2013

Mediametrie

# Plan

- Contexte
- Méthodes testées
- Résultats

# 1

# Contexte

# Contexte

**Panel Multi-Ecrans (PaME) :**

- Démarrage des recrutements au printemps 2012.
- Près de 3 000 foyers ayant accès à Internet et équipés TV.
- Mesure de la consommation TV et Internet à domicile :
  - TV
  - Ordinateur
  - Tablette et Smartphone

**Objectif :**

- Analyse des comportements bi-médias TV-Internet
- Complément aux dispositifs de référence :
- Médiamat pour TV
  - Mesure de la consommation TV à domicile / 5 000 foyers équipés TV
- Panel MNR pour Internet
  - Mesure de la consommation Internet / 20 000 individus internautes

# Contexte

**Besoin du marché :**

- Cohérence des résultats avec la référence Médiamat

**Redressement :**

- Critères socio-démographiques
- Niveau d'audience TV (durée/jour) - Médiamat
  - Total
  - Par chaîne ?
  - Par tranche horaire ?

Idée =

Autoriser un relâchement des contraintes, en particulier, pour les niveaux d'audience TV intégrés dans le redressement

# 2

# Méthodes

# Méthodes

## Calage sur marge

$$\text{Min} \sum_{k \in S} G(w_k, d_k) \quad s.c. \sum_{k \in S} w_k \mathbf{x}_k = T$$

- $d_k$ et $w_k$ les poids initial et final de l'individu k
- G fonction de distance
- $x_k$ le vecteur des variables auxiliaires de l'individu k
- T le vecteur des marges

## Avec la distance du $\chi^2$

- Solution formelle = estimateur par la régression – Särndal, 1980

$$w_k = d_k + d_k \mathbf{x}'_k \left( \sum_{k \epsilon S} d_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left( T - \sum_{k \epsilon S} d_k \mathbf{x}_k \right)$$

# Méthodes

## Distance du $\chi^2$ + contrôle de l'étendue des poids

- Shrinkage-Minimization – Singh & Mohl , 1996
  - Algorithme itératif
  - Initialisation :
    - Le jeu de poids $w_k^0$ est obtenu par l'estimateur de régression généralisé
  - Itération $\nu$+1
    - Rétrécissement des poids pour respecter des bornes fixées L et U :

$$w_k^{\nu *} = \begin{cases} L' d_k & si\ w_k^\nu < L'' d_k \\ U' d_k & si\ w_k^\nu > U'' d_k \\ w_k^\nu & sinon \end{cases} \quad \text{et} \quad \begin{cases} L' = \alpha L + (1-\alpha), U' = \alpha U + (1-\alpha) \\ L'' = \eta L + (1-\eta), U'' = \eta U + (1-\eta) \end{cases} \quad \text{où} \quad 0 < \alpha \leq \eta \leq 1$$

    - Calcul des poids par l'estimateur de la régression généralisé :

$$w_k^{\nu+1} = w_k^{\nu *} + w_k^{\nu *} \mathbf{x}'_k \left( \sum_{k \epsilon S} w_k^{\nu *} \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left( T - \sum_{k \epsilon S} w_k^{\nu *} \mathbf{x}_k \right)$$

À partir des poids rétrécis de l'itération précédente

# Méthodes

## Autres fonctions de distance

- Méthode Scaled Modified Chi-Square – Singh & Mohl , 1996
- Généralisation à d'autres distances – Beaumont & Bocci, 2008
  - Algorithme itératif
  - Initialisation :
    - Le jeu de poids $w_k^0$ est obtenu par l'estimateur de régression généralisé
  - Itération $\nu+1$
    - Calcul du facteur de cadrage qui permet d'introduire la distance G souhaitée

$$q_k^\nu = \frac{2(w_k^\nu - d_k)}{d_k \frac{dG(w_k^\nu, d_k)}{dw_k^\nu}}$$

    - Calcul des poids par l'estimateur de la régression généralisé :

$$w_k^{\nu+1} = d_k + d_k q_k^\nu \mathbf{x}'_k \left( \sum_{k \epsilon S} d_k q_k^\nu \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left( T - \sum_{k \epsilon S} d_k \mathbf{x}_k \right)$$

Avec une distance du $\chi^2$ modifiée par le facteur de cadrage

À partir des poids initiaux

# Méthodes

## Calage sur marges version Ridge

$$Min\left(\sum_{k\in S} G(w_k, d_k) + \lambda \underbrace{\left(\sum_{k\in S} w_k \boldsymbol{x}_k - T\right)' C \left(\sum_{k\in S} w_k \boldsymbol{x}_k - T\right)}_{\mathrm{D}}\right) \quad où \quad C = \begin{pmatrix} c^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c^j \end{pmatrix}$$

- D → relâchement des contraintes en autorisant un écart aux marges
- C la matrice de coût :
  - $c^i \to \infty$, la ième contrainte est obligatoire
  - $c^i = 0$, la ième contrainte est supprimée

## Avec la distance du $\chi 2$

- Solution formelle = Bardsley et Chambers, 1984

$$\omega_k(\lambda) = d_k + d_k \mathbf{x}'_k \left(\sum_{k\in S} d_k \mathbf{x}_k \mathbf{x}'_k + \lambda C^{-1}\right)^{-1} \left(T - \sum_{k\in S} d_k \mathbf{x}_k\right)$$

  - qui dépend de λ et de C

# Méthodes

## Shrinkage-Minimization version Ridge

- Rao & Singh, 2009
  - On conserve le principe de rétrécissement des poids à chaque itération
  - Le calcul des poids est modifié :

$$\omega_k^{\nu+1} = \omega_k^{\nu*} + \omega_k^{\nu*}\mathbf{x}'_k\left(\sum_{k\in S}\omega_k^{\nu*}\mathbf{x}_k\mathbf{x}'_k + C_\nu^{-1}\right)^{-1}\left(T - \sum_{k\in S}\omega_k^{\nu*}\mathbf{x}_k\right)$$

  – A partir des poids rétrécis de l'itération précédente.

  – La matrice des coûts est calculée à chaque itération $\nu$, de sorte que la tolérance (les écarts autorisés vis-à-vis des marges) augmente progressivement à chaque itération.

# Méthodes

## Méthode Scaled Modified Chi-Square version Ridge

- Beaumont & Bocci, 2008
  - Pour λ et C fixés, seule l'étape de calcul des poids est modifiée :

$$w_k^{\nu+1} = d_k + d_k q_k^{\nu} \mathbf{x}'_k \left( \sum_{k \epsilon S} d_k q_k^{\nu} \mathbf{x}_k \mathbf{x}'_k + \lambda C^{-1} \right)^{-1} \left( T - \sum_{k \epsilon S} d_k \mathbf{x}_k \right)$$

- Comment déterminer λ et C ?
  - Pour C, comme le conseillent les auteurs nous avons choisi de prendre l'inverse des totaux des variables
    - alternative proposée = C fixé par l'expérience
  - Pour λ, nous avons choisi d'utiliser l'algorithme de la bissection

# Méthodes

## Méthode Scaled Modified Chi-Square version Ridge

- Difficulté rencontrée :
  - Convergence beaucoup trop lente de l'algorithme
  - Car on revient toujours au poids initial

- Adaptation proposée :
  - Calcul des poids de l'itération ν+1 à partir des poids de l'itération ν (idem Shrinkage-Minimization)

$$w_k^{\nu+1} = w_k^{\nu} + w_k^{\nu} q_k^{\nu} \mathbf{x'}_k \left( \sum_{k \epsilon S} w_k^{\nu} q_k^{\nu} \mathbf{x}_k \mathbf{x'}_k + \lambda C^{-1} \right)^{-1} \left( T - \sum_{k \epsilon S} w_k^{\nu} \mathbf{x}_k \right)$$

  - Pas de preuve théorique
  - Résultats meilleurs qu'avec les 2 algorithmes précédents

# Méthodes

## Autres méthodes testées pour comparaison :

- Macro CALMAR – Sautory, 1994

- Iterative proportional fitting – Deming & Stephan, 1940
  - Avec bornages des poids
  - Relâchement des contraintes paramétrable
    - Seuil de convergence défini comme l'écart autorisé à chaque marge

# 3

# Résultats

# Résultats

## Cas testés :

- SD + DEI TTV → 1 critère audience
- SD + DEI TTV et TOP 4 Chaînes → 5 critères audience
- SD + DEI TTV et 17 Chaînes gratuites → 18 critères audience
- SD + DEI TTV et 17 Chaînes x 10 Tranches Horaires → 171 critères audience

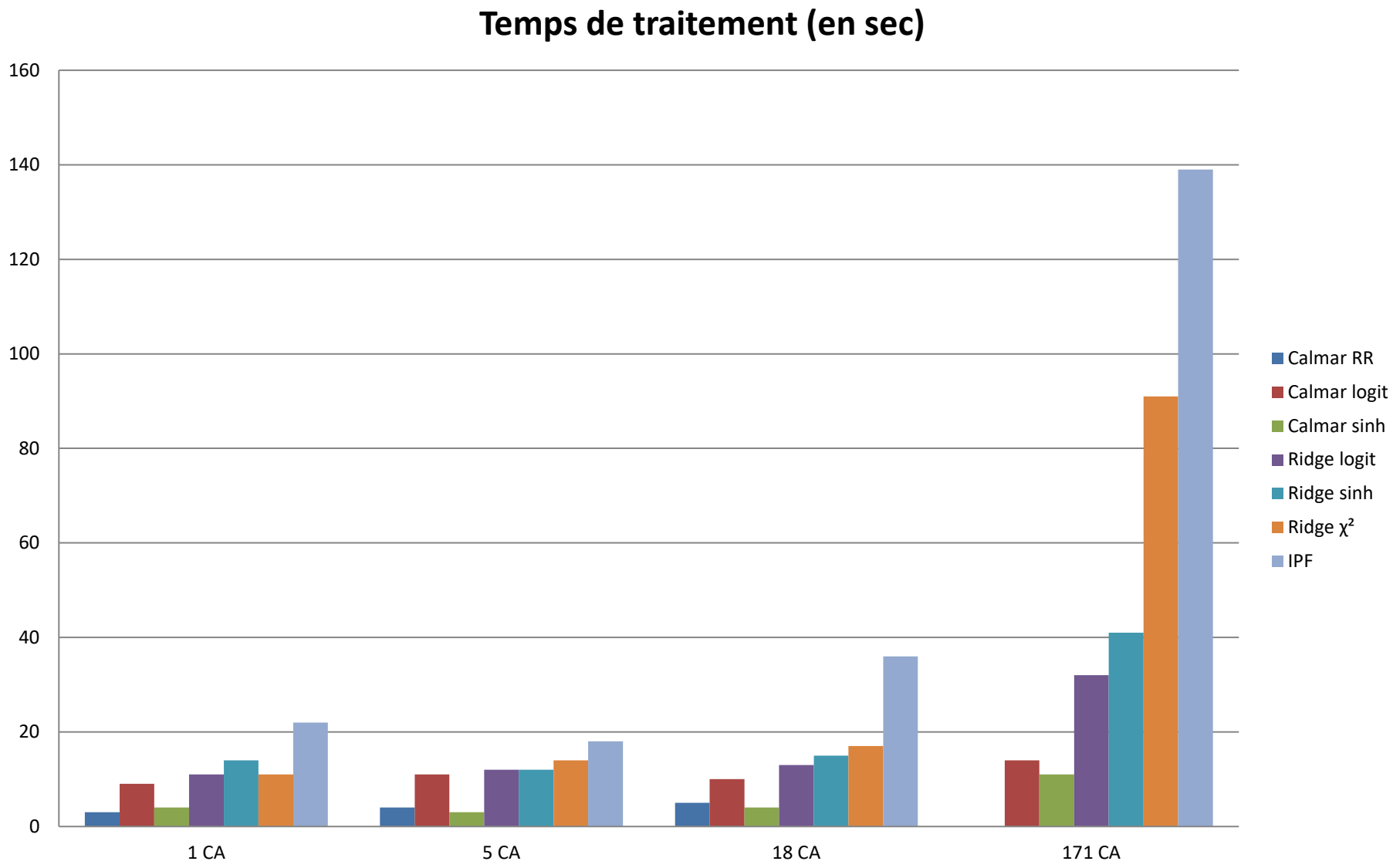
## Caractéristiques :

- Semaine du 8 au 14 avril 2013 (sans vacance scolaire ni jour férié)
- 5 800 individus environ

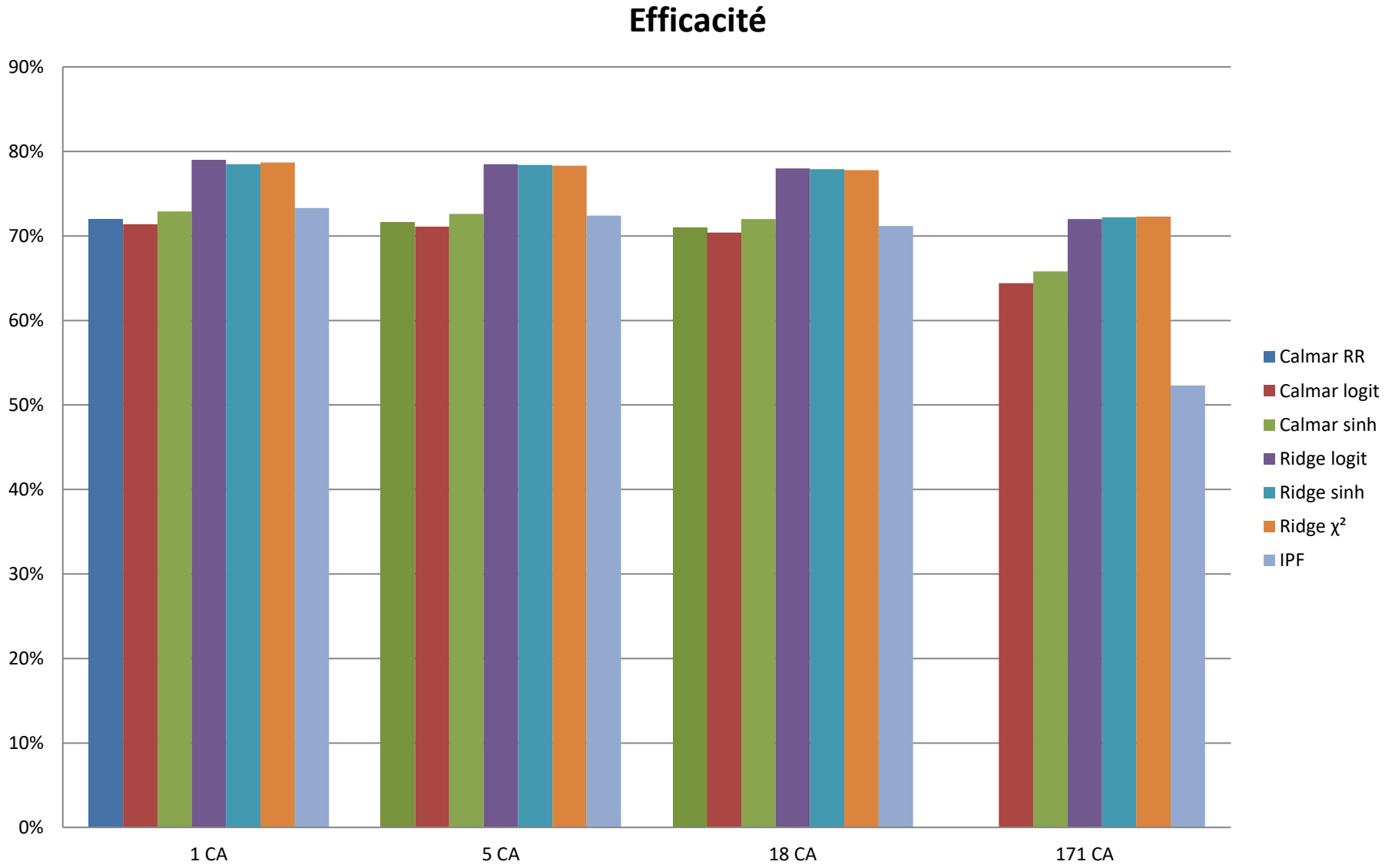
## Indicateurs observés :

- Précision du calage effectivement observé
- Temps moyen de traitement pour un jour redressé
- Efficacité moyenne (sur les 7 jours)
- Rapport de poids moyen (sur les 7 jours)
- Rapport de poids individuel moyen et maximum (sur les 7 jours)
- Coefficient de variation individuel moyen et maximum (sur les 7 jours)

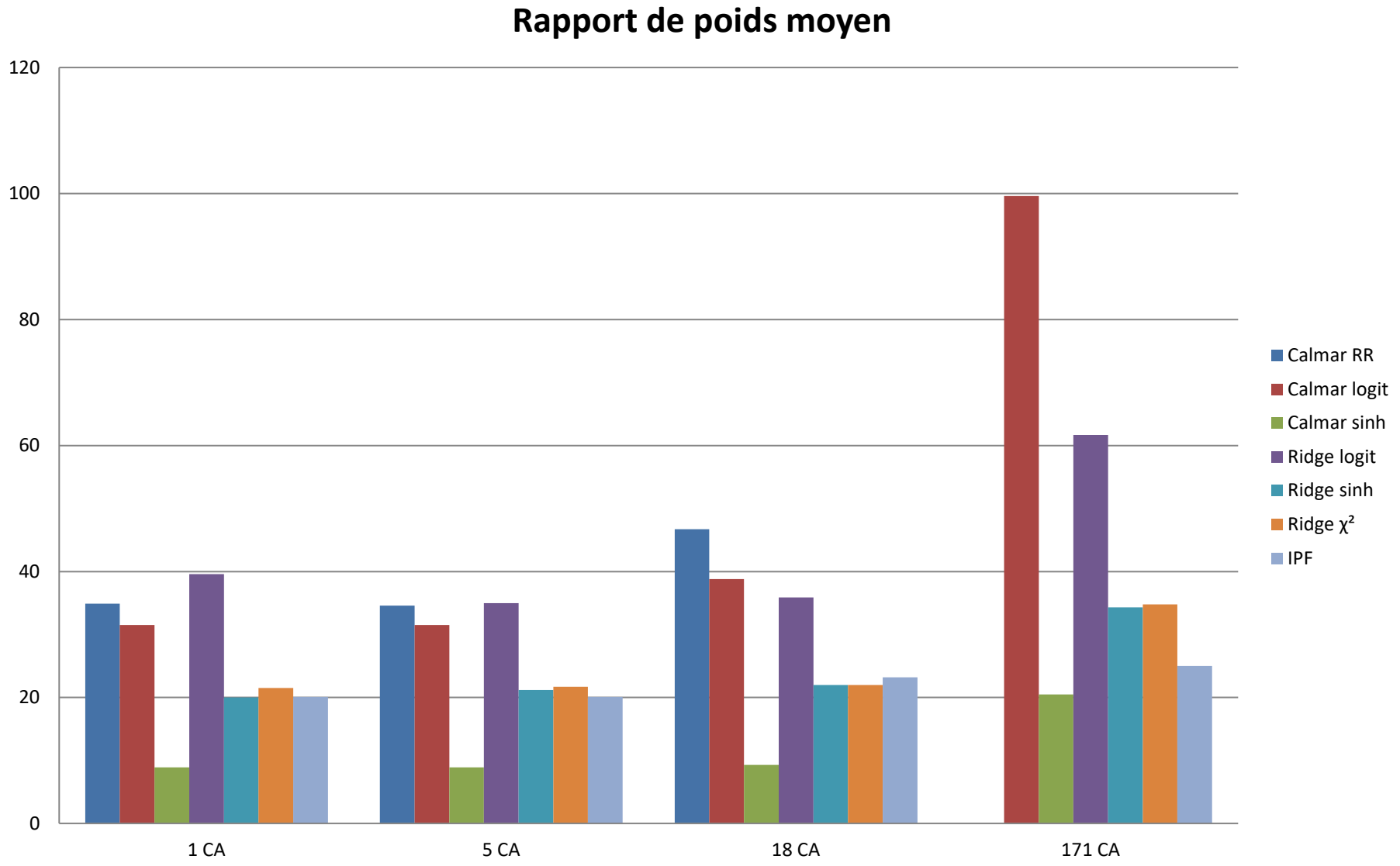
# Résultats


Temps de traitement (en sec)
160
140
120
100
80
60
40
20
0
1 CA
5 CA
18 CA
171 CA
Calmar RR
Calmar logit
Calmar sinh
Ridge logit
Ridge sinh
Ridge χ²
IPF

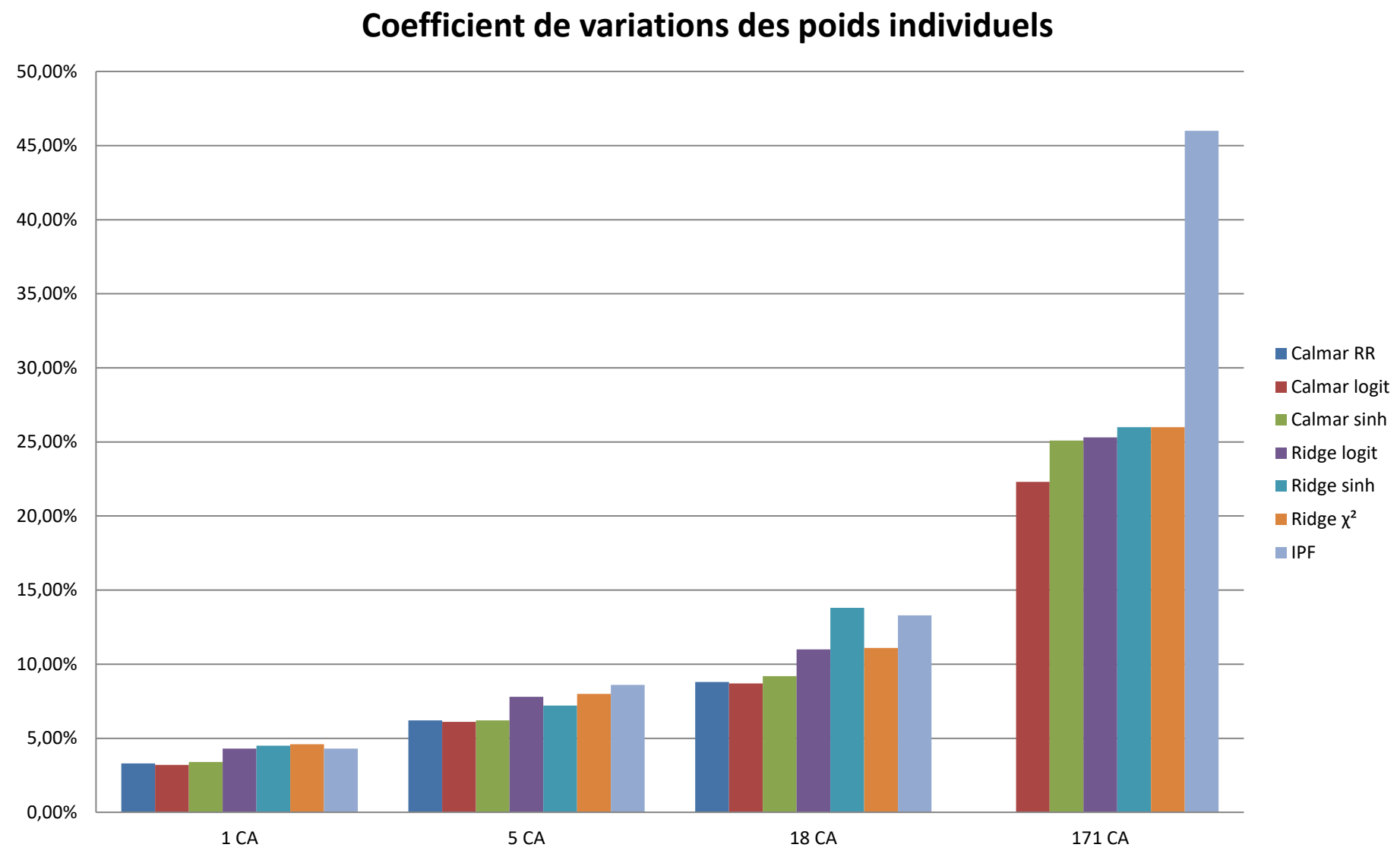
# Résultats


Efficacité
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 CA
5 CA
18 CA
171 CA
Calmar RR
Calmar logit
Calmar sinh
Ridge logit
Ridge sinh
Ridge χ²
IPF

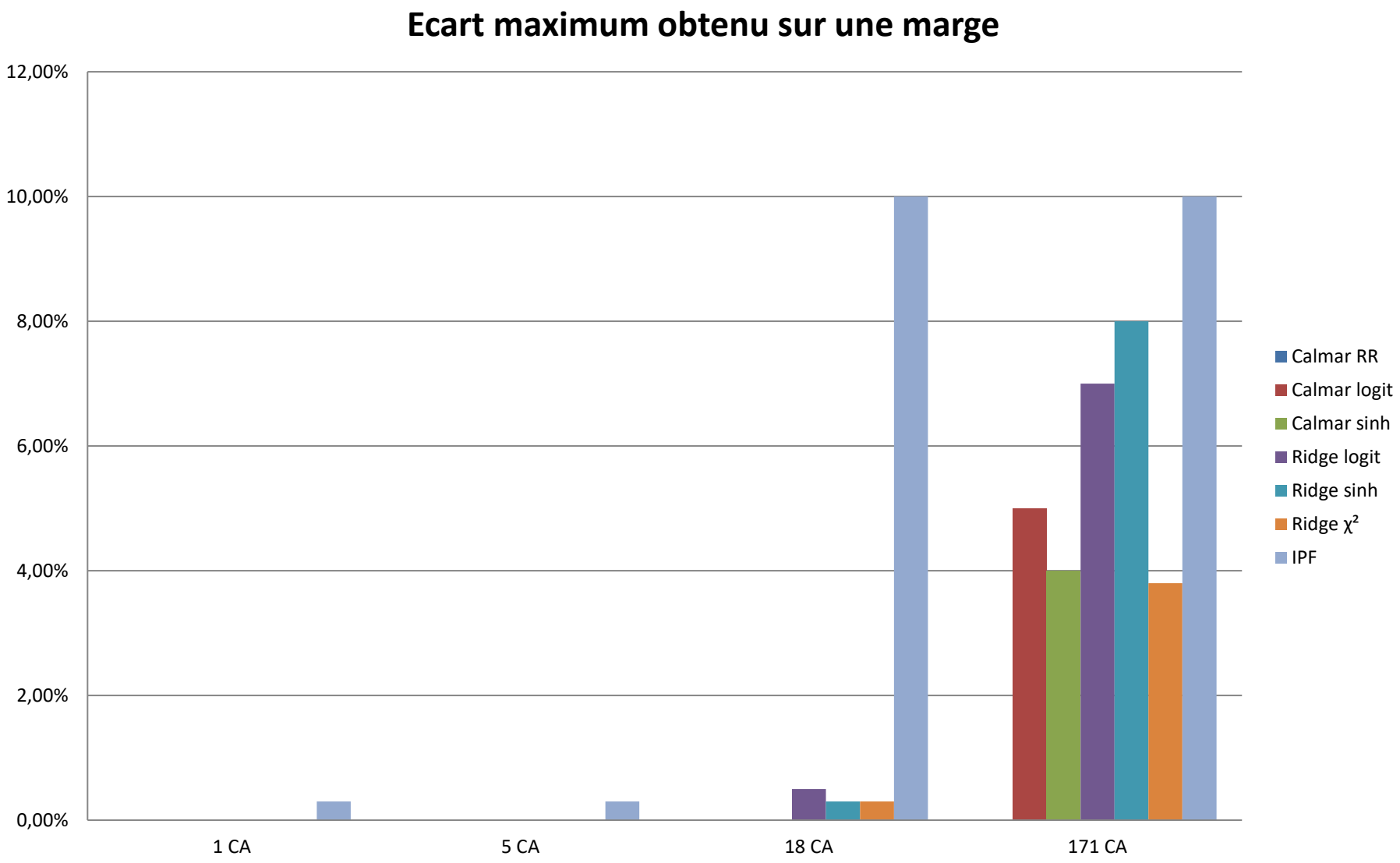
# Résultats


Rapport de poids moyen
120
100
80
60
40
20
0
1 CA
5 CA
18 CA
171 CA
Calmar RR
Calmar logit
Calmar sinh
Ridge logit
Ridge sinh
Ridge χ²
IPF

# Résultats

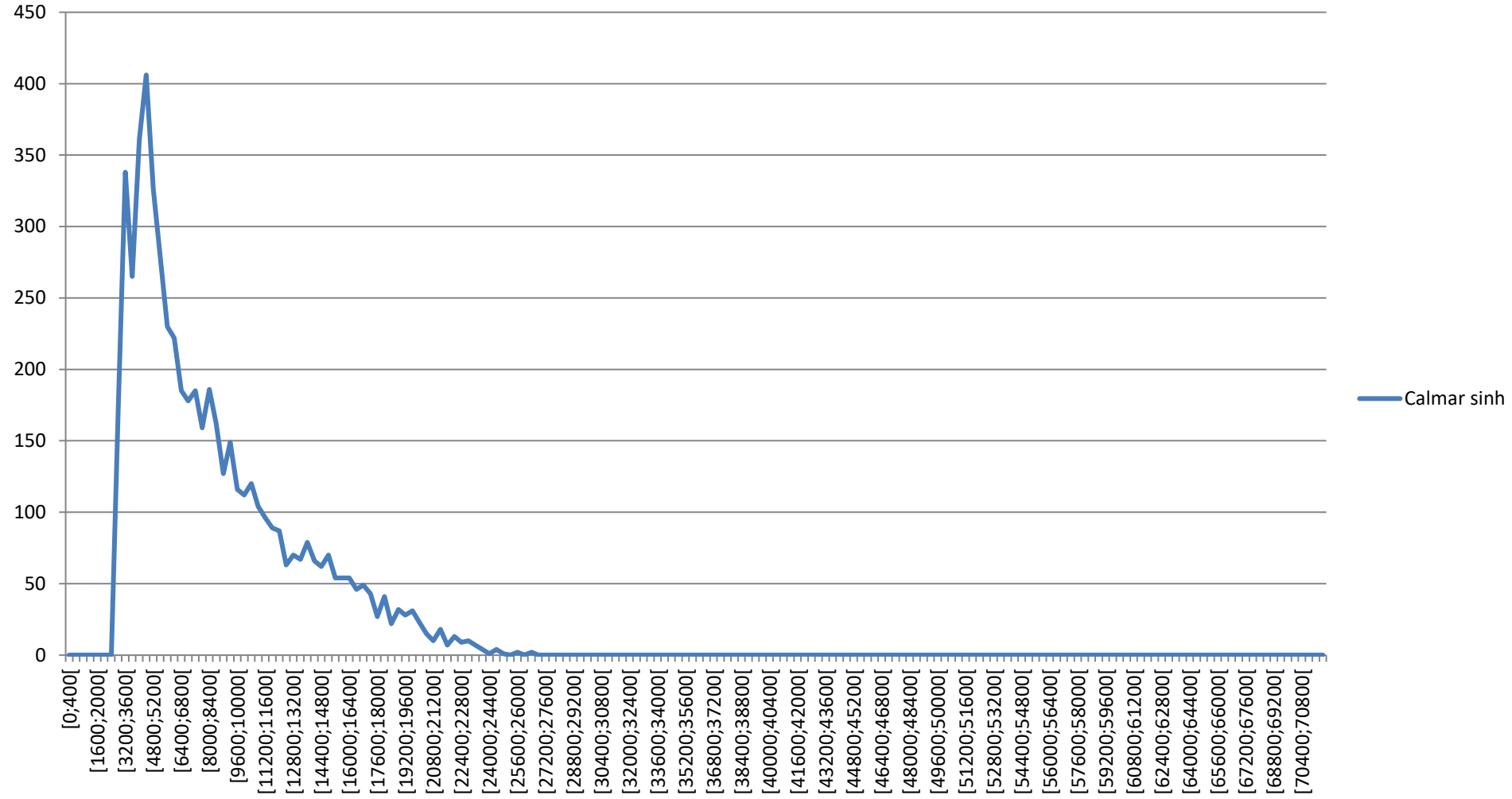

Coefficient de variations des poids individuels
50,00%
45,00%
40,00%
35,00%
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
1 CA
5 CA
18 CA
171 CA
Calmar RR
Calmar logit
Calmar sinh
Ridge logit
Ridge sinh
Ridge χ²
IPF

# Résultats


Ecart maximum obtenu sur une marge
12,00%
10,00%
8,00%
6,00%
4,00%
2,00%
0,00%
1 CA
5 CA
18 CA
171 CA
Calmar RR
Calmar logit
Calmar sinh
Ridge logit
Ridge sinh
Ridge χ²
IPF

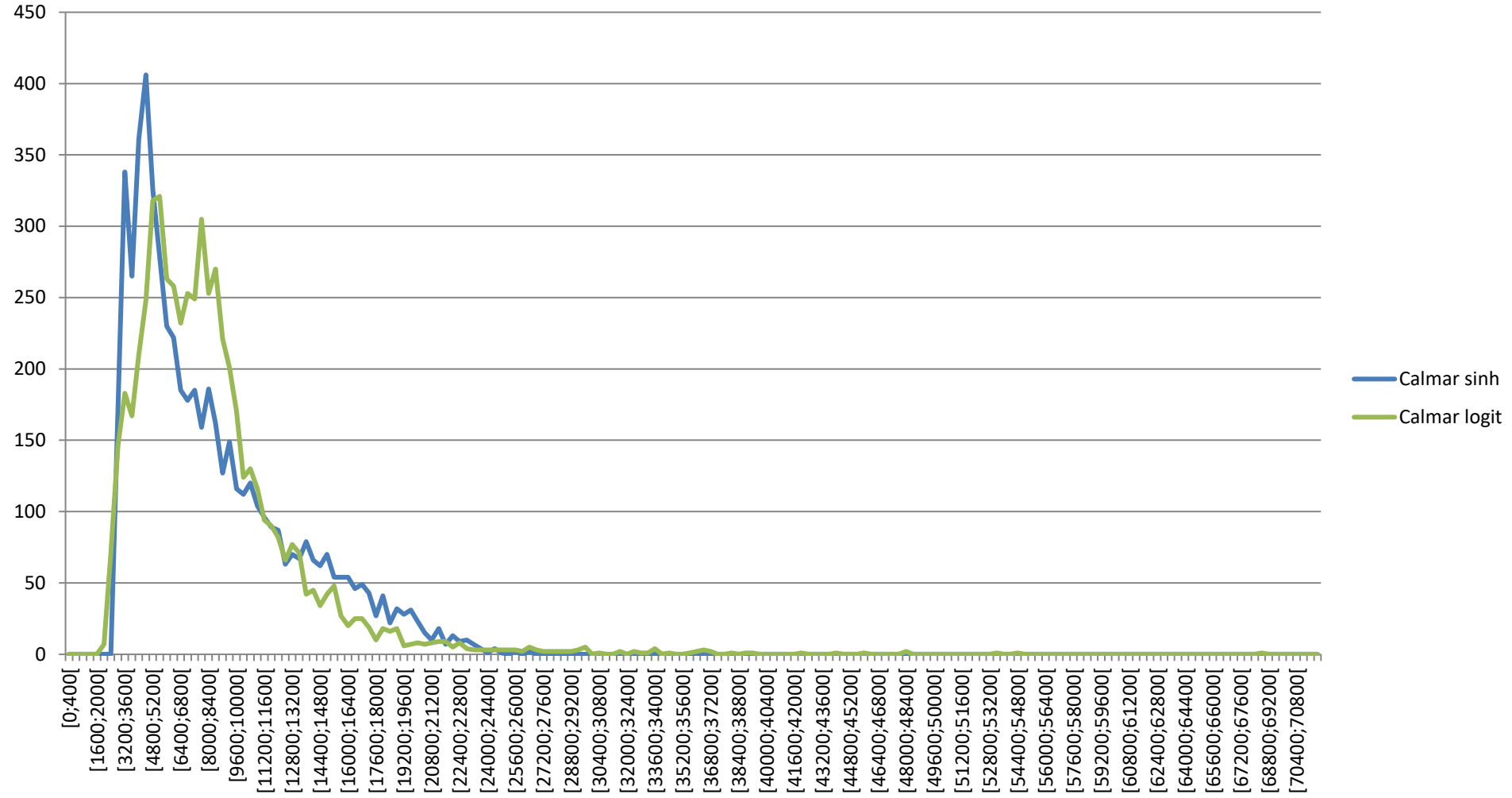
**Résultats**

# Redressement socio-démo + TOTAL TV + TOP17



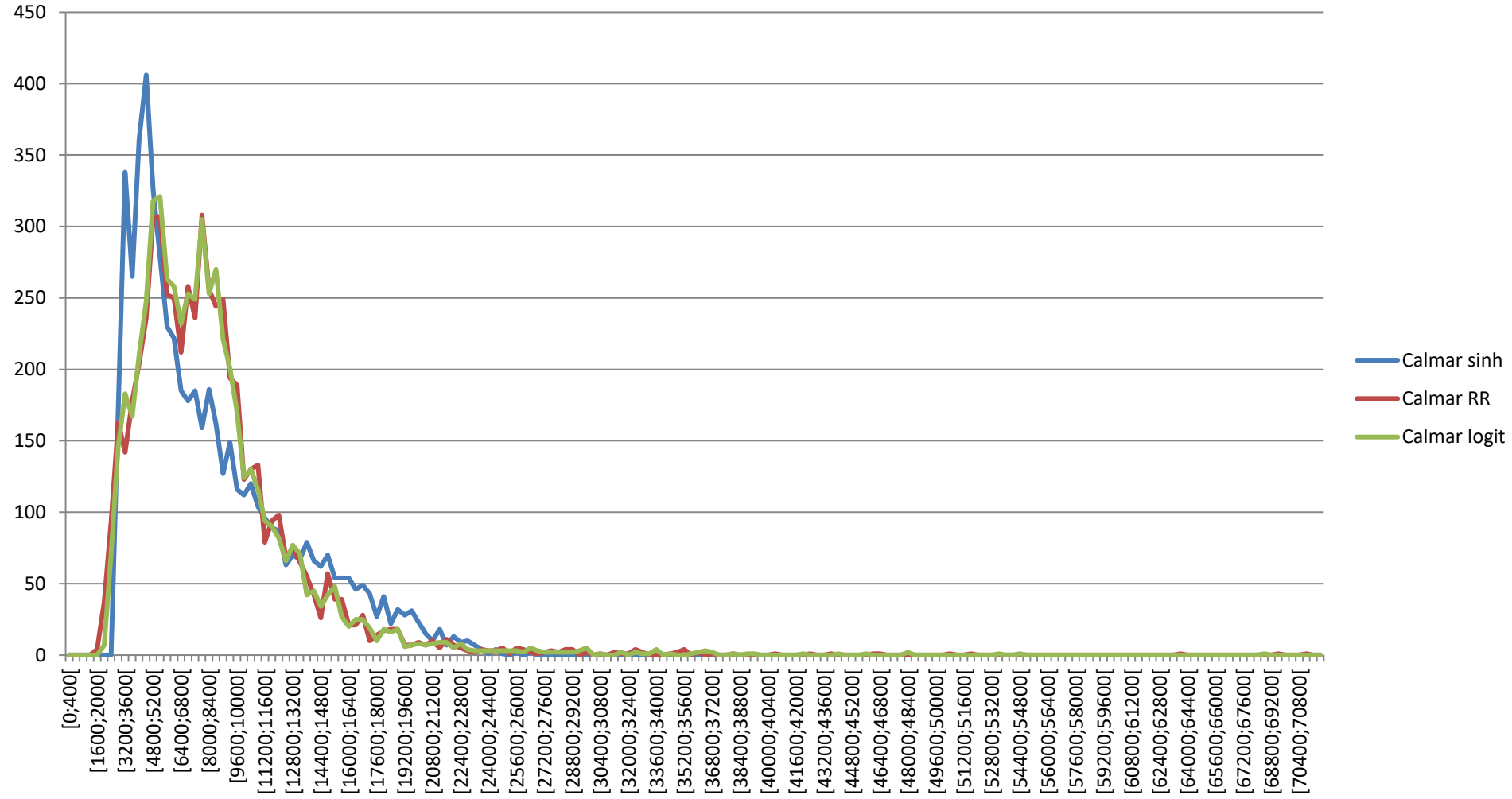
# Résultats

# Redressement socio-démo + TOTAL TV + TOP17



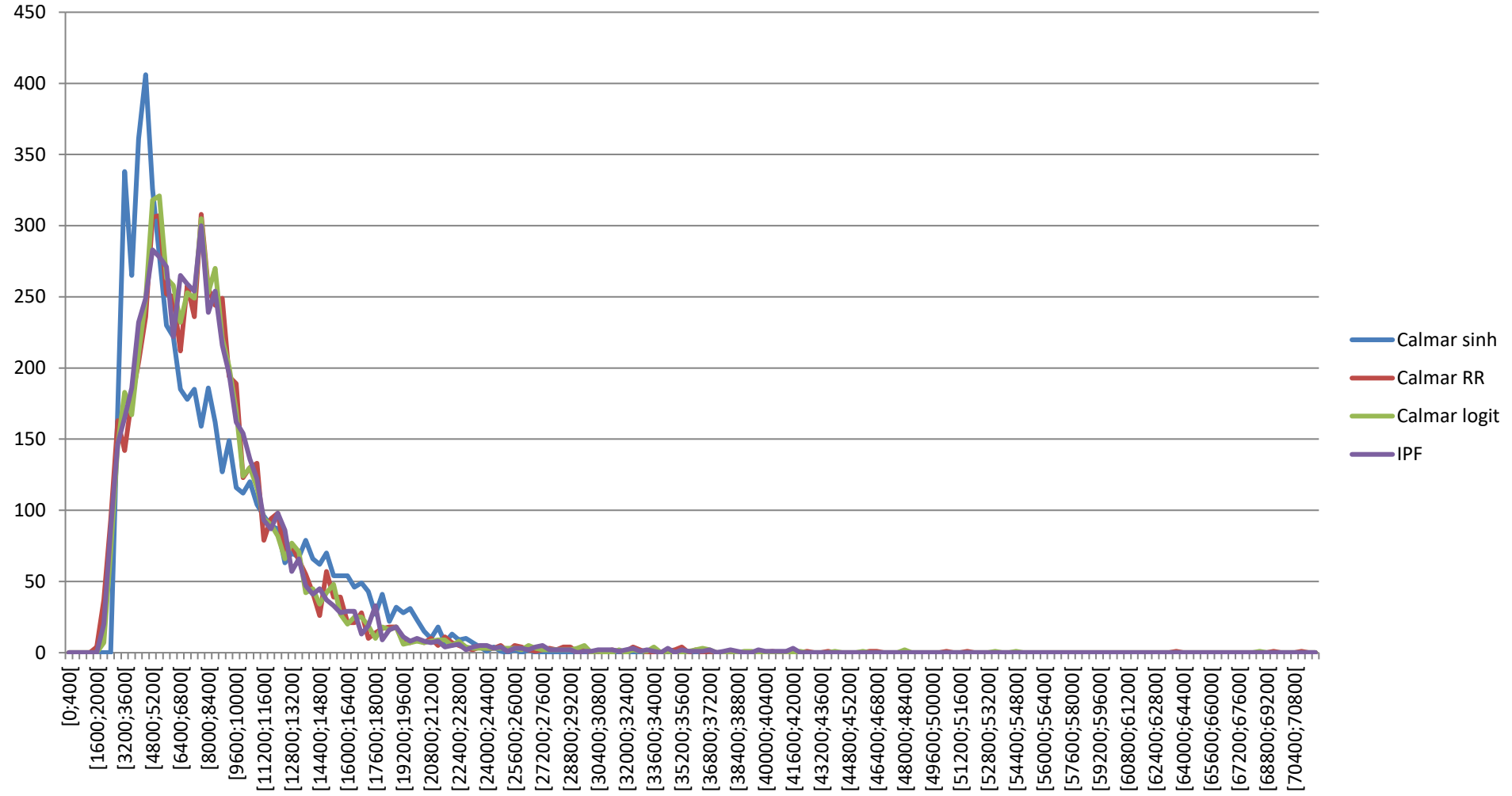
# Résultats

# Redressement socio-démo + TOTAL TV + TOP17



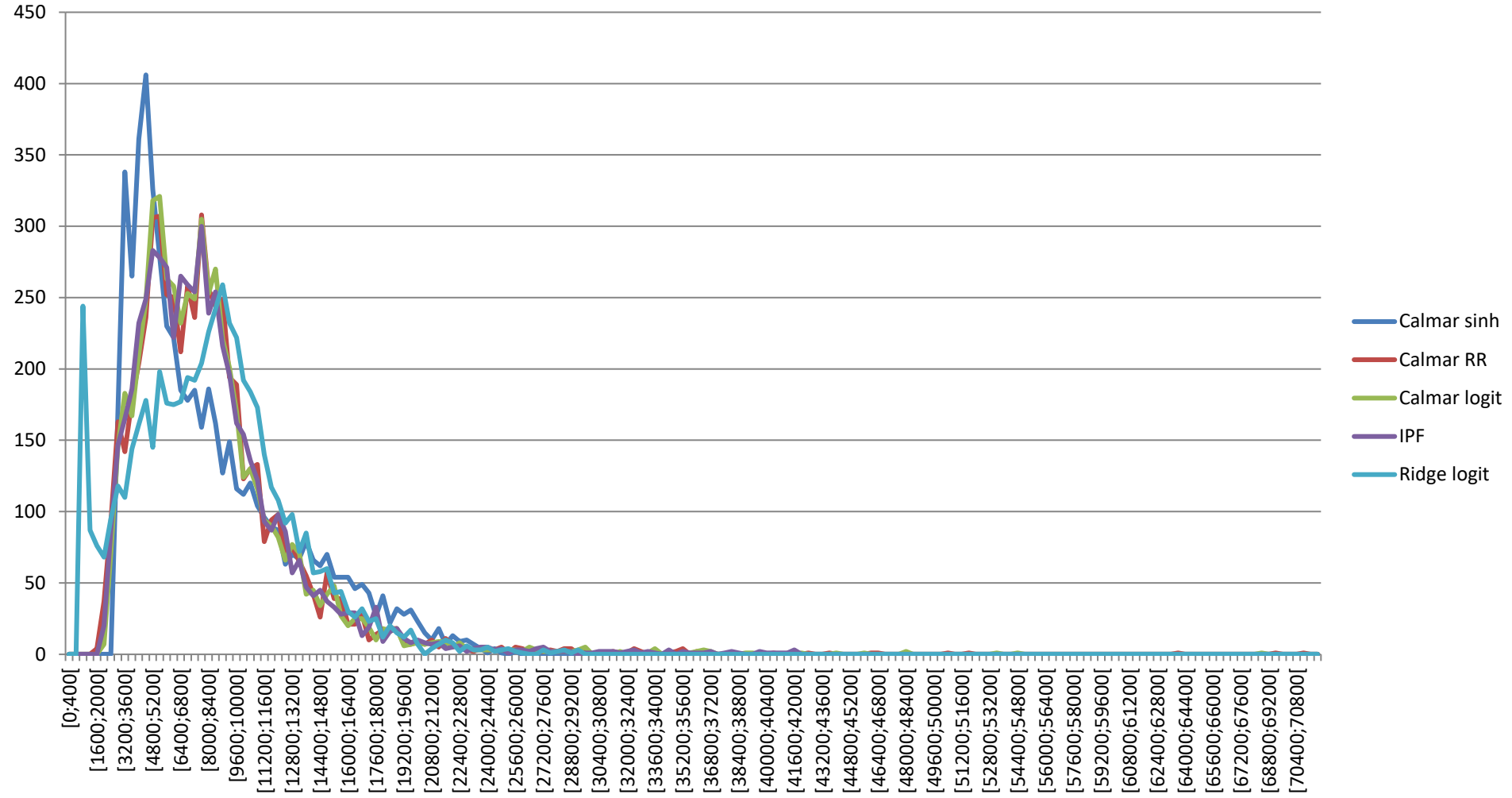
## Résultats

# Redressement socio-démo + TOTAL TV + TOP17



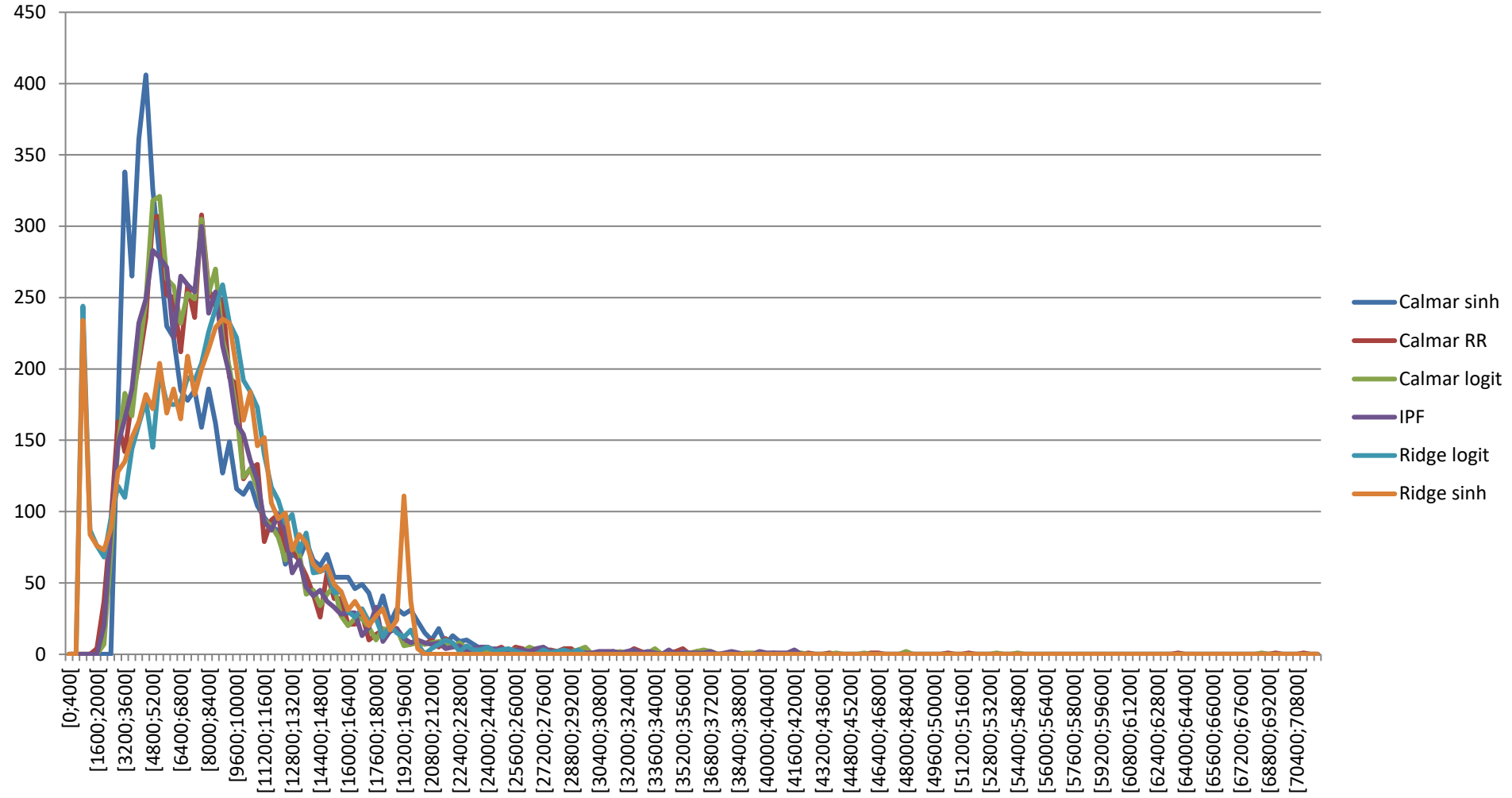
# Résultats

# Redressement socio-démo + TOTAL TV + TOP17



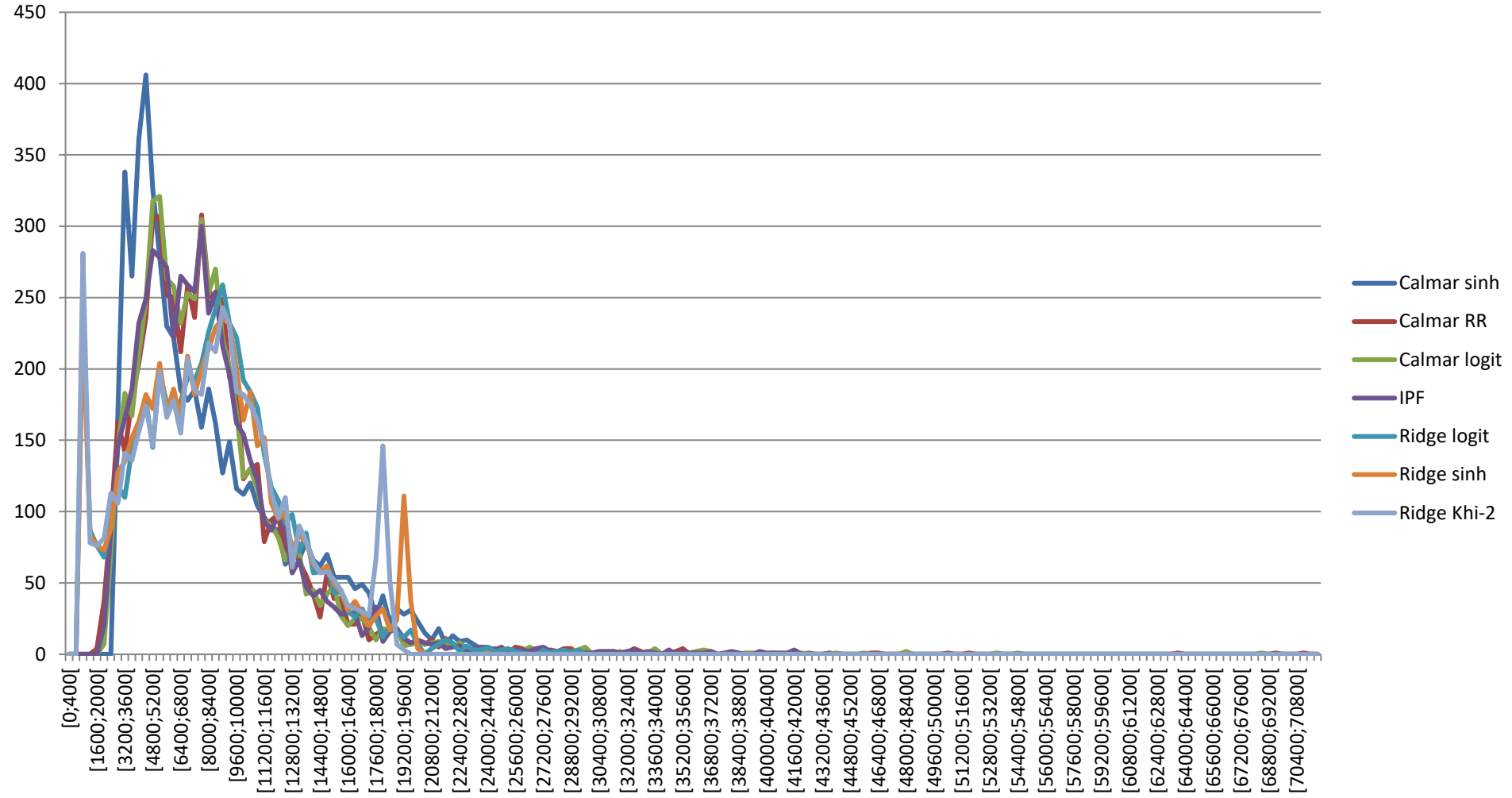
# Résultats

# Redressement socio-démo + TOTAL TV + TOP17



**Résultats**

# Redressement socio-démo + TOTAL TV + TOP17



# 4

# Conclusion

# Conclusion

**Une méthode innovante :**

- Possibilité d'intégrer de nombreuses marges sans les caler parfaitement.

- Une efficacité augmentée par rapport à d'autres méthodes.

**Mais des points restent à améliorer :**

- Le biais n'est contrôlé que globalement.

- Le paramétrage est long et complexe.

- Temps de traitement très dépendant de la taille du vecteur de marges à caler : mise en production potentiellement délicate lorsque le nombre d'individus et le nombre de marges est élevé.