Imputations de données de revenu à l'aide de calage généralisé et de lois GB2 illustration sur les données SILC 2009

Eric Graf

Institut de Statistique
Université de Neuchâtel
www.unine.ch/statistics

15 juin 2012

Plan

Motivation I

- La connaissance de la distribution des revenus de la population constitue un intérêt vital pour toutes les études de marchés économiques pour gouverner les prises de décisions économiques ou sociales
- L'étude de la distribution des revenus est au cœur des mesures d'inégalités et plus généralement des évaluations du bien-être social
- Dans les enquêtes par échantillonnage auprès des ménages et des personnes, les questions sur le revenu sont souvent sensibles et donc sujettes à un taux de non réponse (NR) plus élevé
- La NR partielle s'ajoute à la NR totale

Motivation II

- Le phénomène de la NR altère notre vision de la distribution qu'on tente de cerner par l'enquête
- Sans traitement, les mesures d'inégalités calculées uniquement sur les répondants risques d'être biaisées
- Volonté de fournir des jeux de données complets, i.e. sans valeurs manquantes, à EUROSTAT et aux utilisateurs
- Nécessité d'avoir une stratégie d'imputation pour « boucher les trous »

Motivation III

- Le projet européen AMELI (2011), reposant sur les données EU-SILC, a montré que la loi bêta généralisée de seconde espèce *GB2* s'ajustait bien aux revenus récoltés dans cette enquête.
- Forts de ce résultat, nous avons choisi un système d'imputation par modèle par opposition à des imputations non paramétriques (p. ex. plus proche voisin).
- Le jeu de données suisses de SILC a pu être couplé aux données de la caisse de compensation (CDC), on dispose de la valeur relevée par CATI et de celle du registre
  $\rightarrow$ on applique le vrai mécanisme de NR affectant le CATI à la variable CDC et on s'entraîne à ré-imputer.

But I

Le système d'imputation doit

- être transparent : la qualité de chaque étape doit pouvoir être évaluée (pas de programme-boîte-noire)
- être reproductible : pas d'intervention « à la main »ou non argumentable méthodologiquement
- respecter le plus possible la distribution originale, naturelle et inconnue (!) des revenus à imputer
- pouvoir prendre en compte une pondération (poids d'échantillonnage ou poids déjà corrigés pour certaines étapes de la NR, ou poids obtenus par calage)
- permettre un calcul de la variance due à l'imputation
- fournir un modèle robuste face aux valeurs aberrantes ou extrêmes, mais s'accommoder tout de même de la nature d'une distribution de revenus

Quelle est la distribution des revenus ?

Elle n'est pas normale ni log-normale !

Dans le cas d'une imputation paramétrique, les hypothèses des modèles de régressions classiques pour expliquer les revenus (ou leur logarithme) ne sont donc pas vérifiées.

De telles imputations modifient la distribution originale (et « naturelle » !) des données : peut conduire l'utilisateur à interpréter faussement les résultats calculés à partir d'un jeu de données ainsi imputé $\rightarrow$ risque d'introduire un biais.

Loi GB2 I

La loi bêta généralisée de seconde espèce est une distribution à quatre paramètres : $GB2(a, b, p, q)$. Elle a été développée par McDonald (1984).
Des études empiriques sur le revenu - voir p. ex. Jenkins (2007) ; Dastrup et al. (2007) ; Kleiber et Kotz (2003) ; Sepanski et Kong (2007) - montrent que la GB2 s'ajuste bien à de telles données et qu'elle est souvent plus adaptée que d'autres distribution à quatre paramètres.
Résultats du projet européen AMELI (2011) confirment pour EU-SILC.

Loi GB2 II

fonction de densité d'une variable aléatoire suivant une loi GB2 :

$$f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p, q)} \frac{(y/b)^{ap-1}}{(1 + (y/b)^a)^{p+q}} \qquad (1)$$

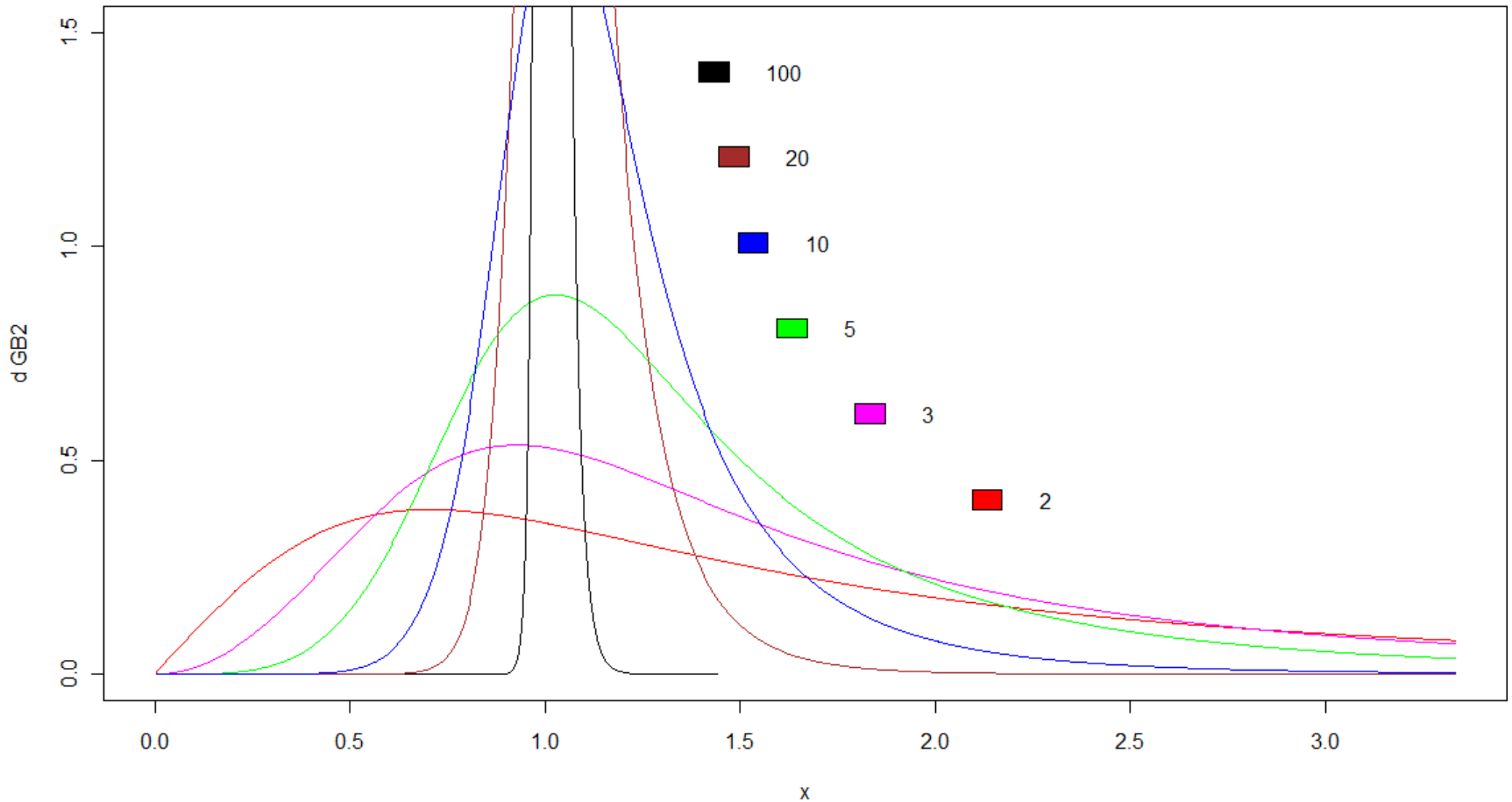
où $B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt$ est la fonction bêta.

- a représente la courbure générale, il détermine la vitesse avec laquelle les queues de la distribution approchent l'axe. Une grande valeur de a implique une distribution assez pointue.
- b est paramètre d'échelle, pour de grandes valeurs de a, b tend vers la moyenne (l'espérance)
- p gouverne la queue gauche
- q gouverne la queue droite
- Les paramètres p et q déterminent ensemble l'asymétrie de la distribution

Loi GB2 III

densités GB2, a variable, $b = 1$, $p = 1$, $q = 0.5$

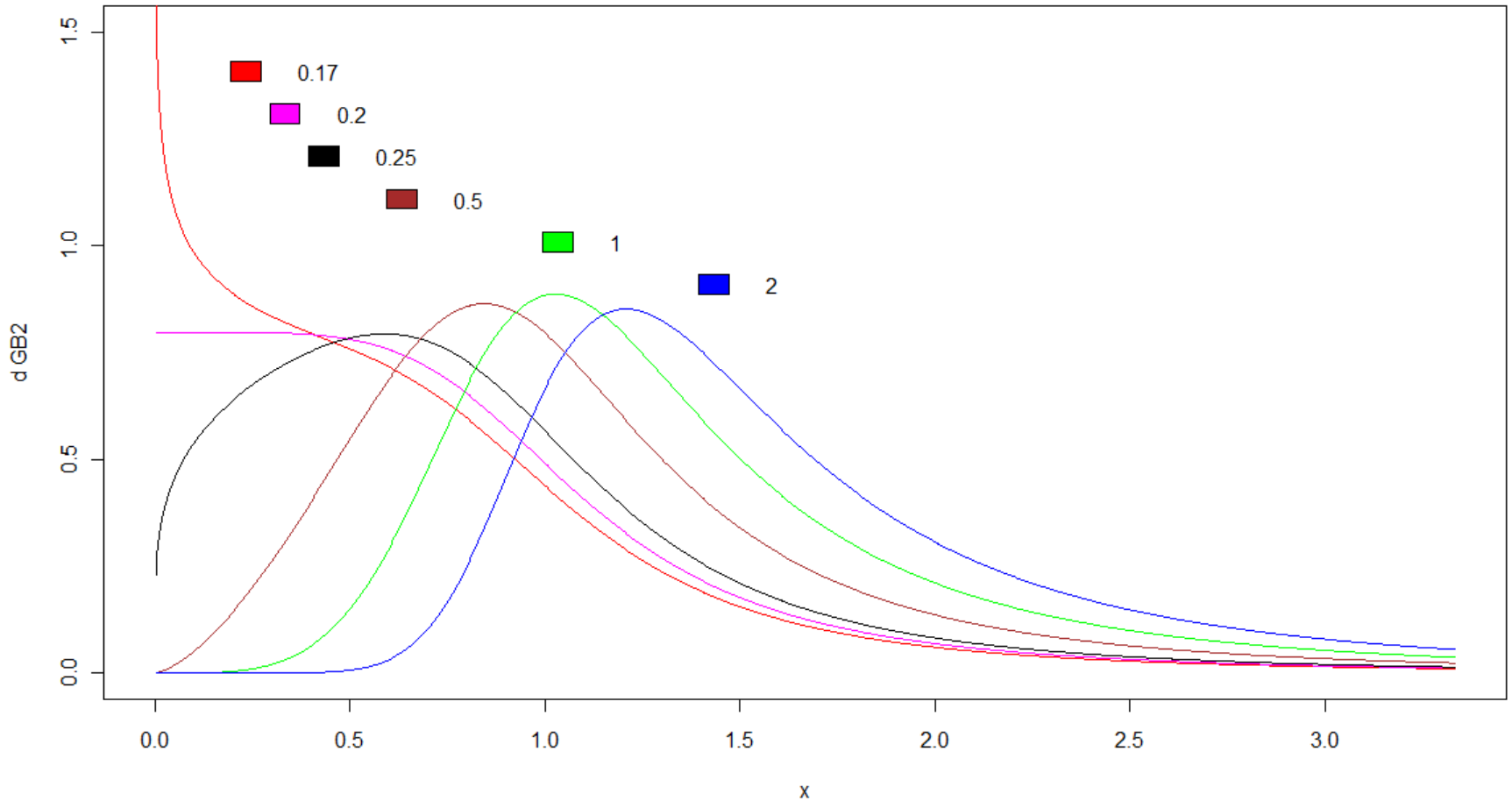
$\rightarrow$ distribution $\pm$ pointue



Loi GB2 IV

densités GB2, $a = 5$, $b = 1$, p variable, $q = 0.5$

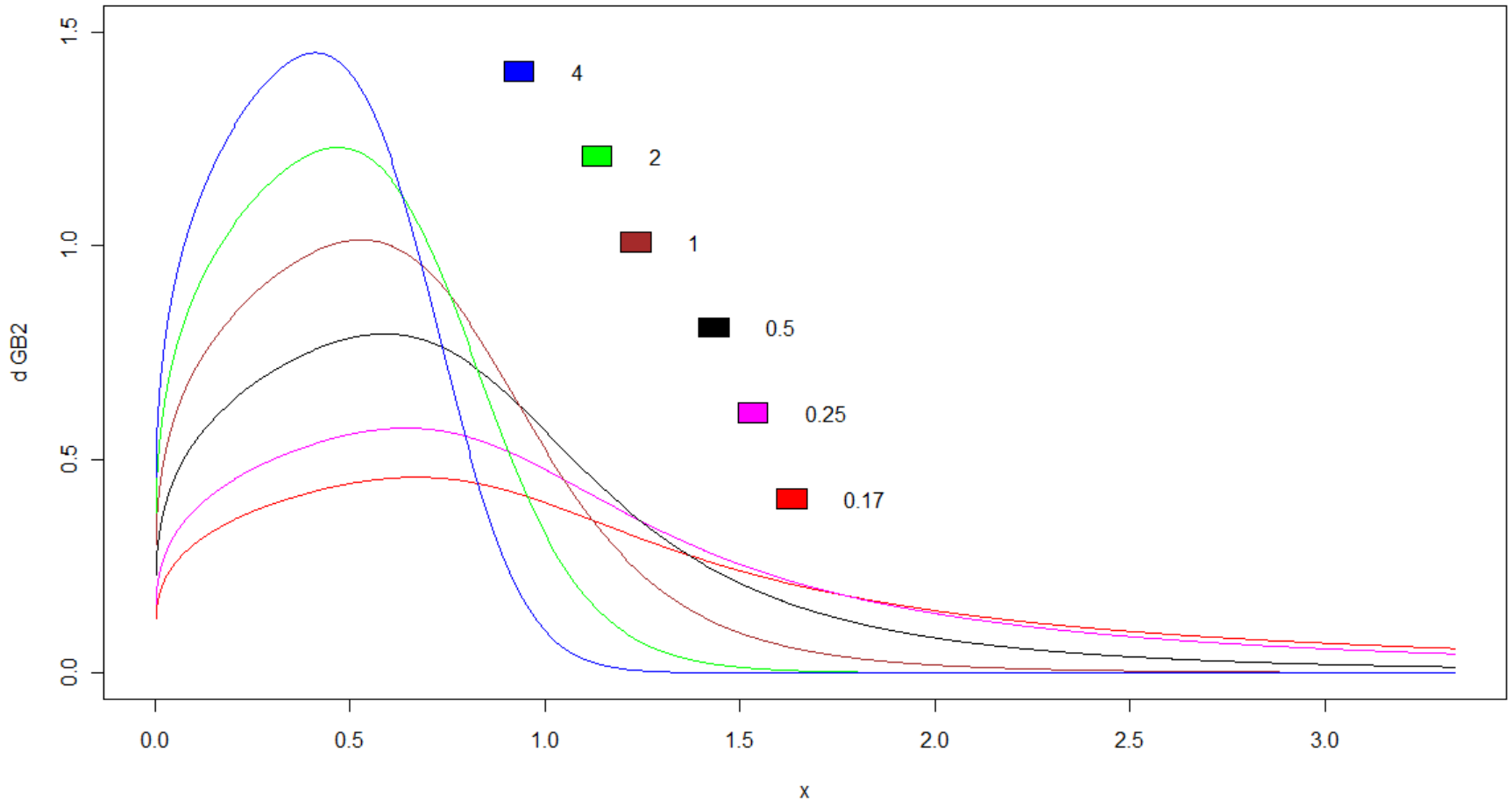
$\rightarrow$ on joue sur la forme de la queue de gauche



Loi GB2 V

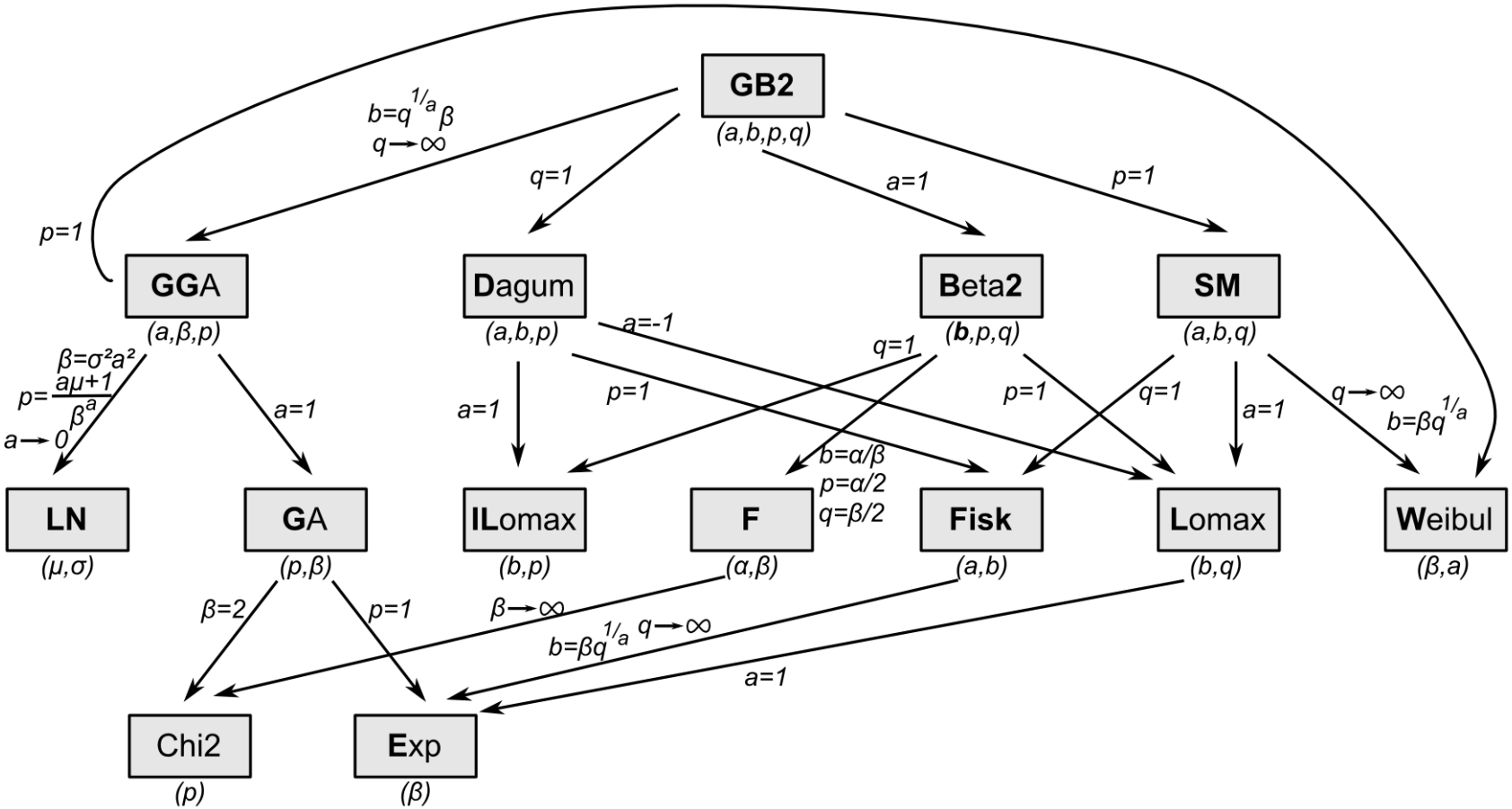
densités GB2, $a = 5$, $b = 1$, $p = 0.25$, q variable

$\rightarrow$ on joue sur la forme de la queue de droite



Loi GB2 VI

Plusieurs lois peuvent être vues comme des cas particuliers de la GB2



Loi GB2 et indices d'inégalité

Avantage d'une estimation paramétrique d'une distribution de revenu :
il existe des formules explicites pour les mesures d'inégalité comme des fonctions des quatre paramètres de la loi GB2 ajustée aux données - McDonald (1984), Graf, M. (2009), Ameli (2011).

Seuil de risque de pauvreté ARPT(a,b,p,q)
Taux de risque de pauvreté ARPR(a(,b),p,q)
Relative median at-risk-of poverty gap RMPG(a(,b),p,q)
Quintile share ratio (S_{80}/S_{20}) QSR(a(,b),p,q)
Indice de Gini GINI(a(,b),p,q)

Mécanisme de non réponse I

$\rightarrow$ description de la relation entre la NR avec les variables du jeu de données.

La distribution de la NR est caractérisée par la distribution conditionnelle de l'indicatrice de réponse $\mathcal{R} \in \{0, 1\}$ étant donné $y = (y_{obs}, y_{mis})$:

$$P(\mathcal{R}|y) = P(\mathcal{R}|y_{obs}, y_{mis}) \tag{2}$$

Classiquement, on distingue trois types de mécanismes de NR pouvant affecter la variable d'intérêt : MCAR, MAR, NMAR.

Mécanisme de non réponse II

- MCAR (Missing Completely at Random) : la probabilité de réponse est constante, égale pour toutes les observations, elle n'est pas reliée aux valeurs manquantes de y ou d'autres variables $\mathbf{X}$: $P(\mathcal{R}|y) = P(\mathcal{R})$,
- MAR (Missing At Random) : la probabilité de réponse dépend d'une ou plusieurs variables auxiliaires x_j mais pas de y elle-même : $P(\mathcal{R}|y) = P(\mathcal{R}|y_{obs})$,
- NMAR (Not Missing At Random) : la probabilité de réponse dépend de la variable d'intérêt elle-même et de variables auxiliaires x_j : $P(\mathcal{R}|y) = P(\mathcal{R}|y_{obs}, y_{mis})$.

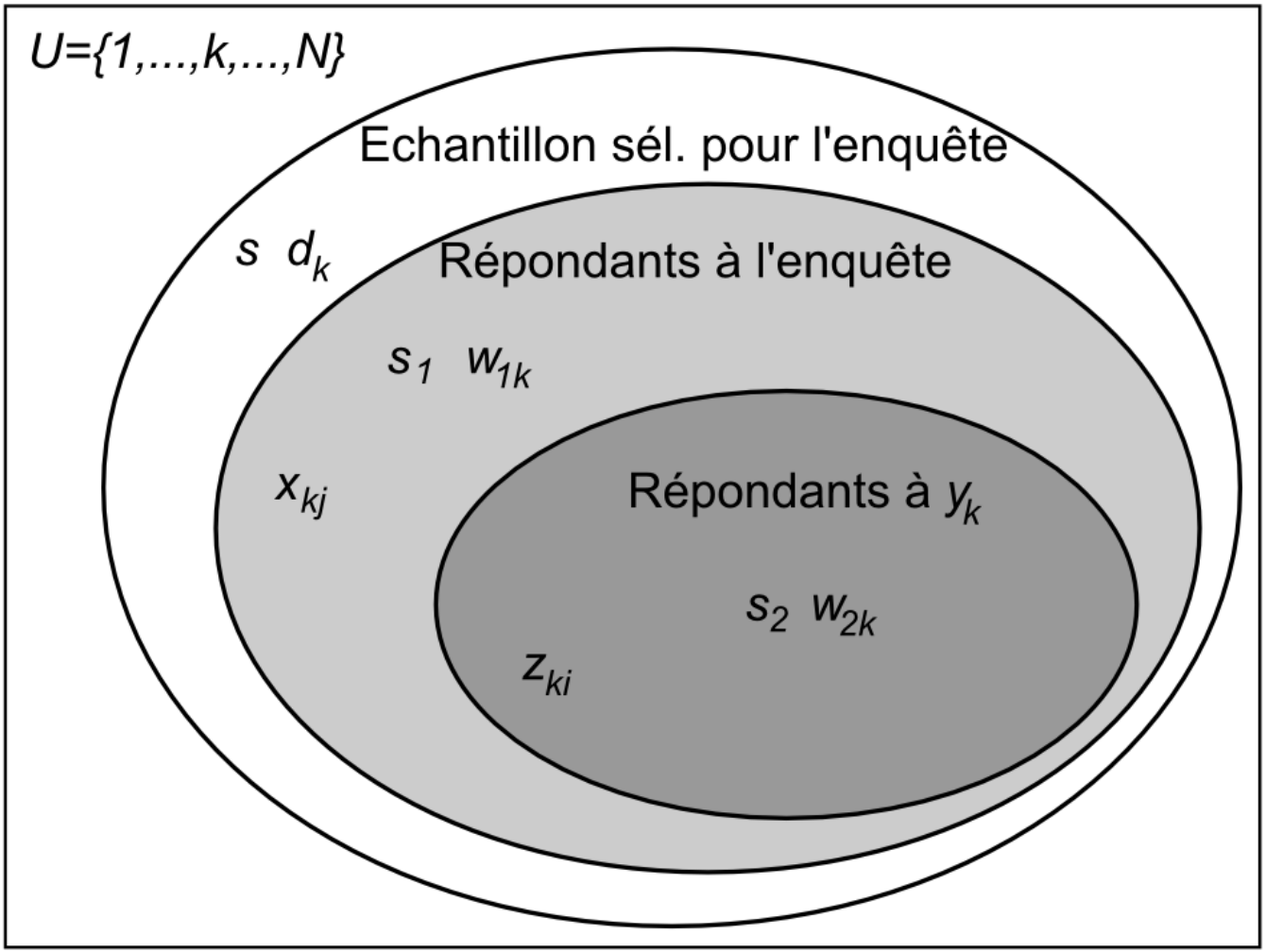
La plupart des variables de revenus que nous voulons imputer sont NMAR...

Calage et calage généralisé I

Pour plus de détails on se référera entre autre à Deville et Särndal (1992) ; Deville, Särndal et Sautory (1993) ; Legennec et Sautory (2002) ; Sautory (2003) ; Deville (2002) ; Kott (2006).

En résumé, on veut obtenir des nouveaux poids, spécifiques à y, corrigeant pour la NR et respectant certaines contraintes.

Notations


U={1,...,k,...,N}
Echantillon sél. pour l'enquête
s d_k
Répondants à l'enquête
s_1 w_1k
x_kj
Répondants à y_k
s_2 w_2k
z_ki

Calage et calage généralisé II

On dispose encore de

- $\mathbf{x}_k = (x_{k1}, ..., x_{kj}, ..., x_{kJ})$ les variables auxiliaires de calage connues sur s_1 et dont on connaît les totaux sur la population $\mathbf{t}_x = \sum_U \mathbf{x}_k$.
- $\mathbf{z}_k = (z_{k1}, ..., z_{ki}, ..., z_{kI})$ les variables instrumentales connues sur s_2.
- On observe $(y_k, \mathbf{x}_k, \mathbf{z}_k)$ et dispose de $\mathbf{t}_x$, de plus on suppose que $J = I$.

L'objectif est d'estimer le total sur la population $t_y = \sum_U y_k$, en général par $\hat{t}_y = \sum_s d_k y_k$ (estimateur HT du total, sans biais).

Calage et calage généralisé III

On cherche des nouveaux poids w_2 proches des poids avant calages w_1 au sens d'une certaine (pseudo-)distance G sous la contrainte des *équations de calage* :

$$\mathbf{t}_x = \sum_{k \in s_2} w_{2k} \mathbf{x}_k \tag{3}$$

Ceci revient à chercher des poids w_2 solutions du programme suivant pour tout échantillon s_2 :

$$\min_{w_{2k}} \sum_{k \in s_2} \frac{G_k(w_{2k}, w_{1k})}{q_k} \text{ sous la contrainte } (3) \tag{4}$$

On peut donner $\pm$ d'importance à certaines unités en pondérant chaque G_k par $1/q_k$.

Calage et calage généralisé IV

On trouve alors que les nouveaux poids calés sont de la forme

calage

$$w_{2k} = w_{1k} F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$$

calage généralisé

$$w_{2k} = w_{1k} F(\mathbf{z}_k' \boldsymbol{\lambda})$$

Ces poids doivent satisfaire les équations de calage (3).
F dépend du choix de la forme de la pseudo-distance G. Par ex., dans le **cas linéaire**, $G_k(w_{2k}, w_{1k}) = \frac{(w_{2k}-w_{1k})^2}{2w_{1k}}$ et

$$F(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = (1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}) \quad \Big| \quad F(\mathbf{z}_k' \boldsymbol{\lambda}) = (1 + \mathbf{z}_k' \boldsymbol{\lambda})$$

On doit ensuite résoudre pour $\boldsymbol{\lambda}$.

Calage et calage généralisé V (cas linéaire)

On obtient

calage

$$\hat{t}_{ylin} = \mathbf{t}_x' \hat{\mathbf{B}}_{s_2} + \sum_{k \in s_2} w_{1k} e_k$$

où $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_{s_2}$ résidus de la régression de y sur les J variables auxiliaires x_k.

$\hat{\mathbf{B}}_{s_2} = \mathbf{T}_{s_2}^{-1} \sum_{k \in s_2} w_{1k} q_k \mathbf{x}_k y_k$ est le vecteur des J paramètres de la régression

$\mathbf{T}_{s_2}^{-1} = \left(\sum_{k \in s_2} w_{1k} \mathbf{x}_k q_k \mathbf{x}_k' \right)^{-1}$

calage généralisé

$$\hat{t}_{ylinG} = \mathbf{t}_x' \hat{\mathbf{B}}_{s_2 zx} + \sum_{k \in s_2} w_{1k} e_k$$

où $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_{s_2 zx}$ résidus de la régression instrumentale de y sur les J $\mathbf{x}_k$ dans l'échantillon s, avec les J variables instrumentales $\mathbf{z}_k$

$\hat{\mathbf{B}}_{s_2 zx} = \mathbf{T}_{s_2 zx}^{-1} \sum_{k \in s_2} w_{1k} \mathbf{z}_k y_k$ est le vecteur des J paramètres de la régr. instrumentale

$\mathbf{T}_{s_2 zx}^{-1} = \left(\sum_{k \in s_2} w_{1k} \mathbf{x}_k \mathbf{z}_k' \right)^{-1}$

Stratégie d'imputation I

1. Pour les répondants, calculer des poids *judicieusement* ajustés par calage généralisé (= qui tiennent compte de la NR) pour la variable à imputer
2. Utiliser cette pondération pour s'approcher de la « bonne »GB2 (i.e. celle qu'on obtiendrait si tout le monde avait répondu)
   $\rightarrow$ les indices de pauvreté peuvent être calculés sans qu'il y ait imputation
3. Ordonner les revenus des répondants selon leurs rang ou rang pondérés (robuste !).

Stratégie d'imputation II

4. Transformer les rangs en quantiles normaux
5. Imputer (prédire) les manquants par un modèle de régression classique reposant sur les variables auxiliaires $\mathbf{x}_k$ regroupées dans la matrice $\mathbf{X}$ et prenant les nouveaux poids w_{2k} en compte
6. Transformation inverse : quantiles normaux imputés $\rightarrow$ rangs imputés
7. Valeurs y imputées par la $GB2$ et les rangs imputés

Illustration avec les données SILC09

Variable d'intérêt y : revenu des personnes salariées

Variable relevée par téléphone	Variable du registre CDC
P09I57G_cati= y_{cati}	P09I57G_cdc= y_{cdc} (Centrale de Compensation)

Fichier d'entraînement : individus appariés au registre, $y_{cdc} > 0$ & $\{y_{cati} = (> 0$, *ne sait pas, pas de réponse, NR à la variable filtre, question filtrée* $)\}$.

7922 obs. $\rightarrow$ 6884 (86.9%) – NR partielle !

On restreint encore à *taux d'occupation*> 0 et *coûts du logement*> 0 pour pouvoir utiliser ces variables comme var. instrumentales.

$\rightarrow$ le nb. de cas complets passe à 6188 individus (21.9% de NR).

On applique ce mécanisme de NR réel à y_{cdc} dont on connaît toutes les valeurs.

① *Pour les répondants, calculer des poids judicieusement ajustés par calage généralisé pour la variable à imputer*

- $\rightarrow$ identifier les variables auxiliaires $\mathbf{X}$ et instrumentales $\mathbf{Z}$
- Les $\mathbf{X}$ doivent expliquer y_{cdc} et être disponibles pour les répondants et les non répondants, donc sur s_1.
  $\rightarrow$ choix raisonné de variables corrélant à $> 10\%$ avec y_{cdc}
- Les $\mathbf{Z}$ doivent expliquer la NR à y_{cdc}
  $\rightarrow$ choisi les variables intervenant dans un arbre de segmentation modélisant la NR + des variables continues expliquant y_{cdc} et liées à la NR

Extrait de l'arbre de segmentation.

Les taux de réponse des 73 feuilles de l'arbre obtenu corrèlent à 39.3% avec les valeurs des médianes par feuille, une preuve que la NR appliquée à y_{cdc} (et affectant y_{cati}) n'est pas ignorable. Les personnes mieux payées ont tendance à répondre mieux.

	$med_{y_{cdc}}^{GHR}$	$medNR_{y_{cdc}}^{GHR}$	tx_{rep}
$med_{y_{cdc}}^{GHR}$	1	0.77	0.39
$medNR_{y_{cdc}}^{GHR}$	0.77	1	0.32
tx_{rep}	0.39	0.32	1

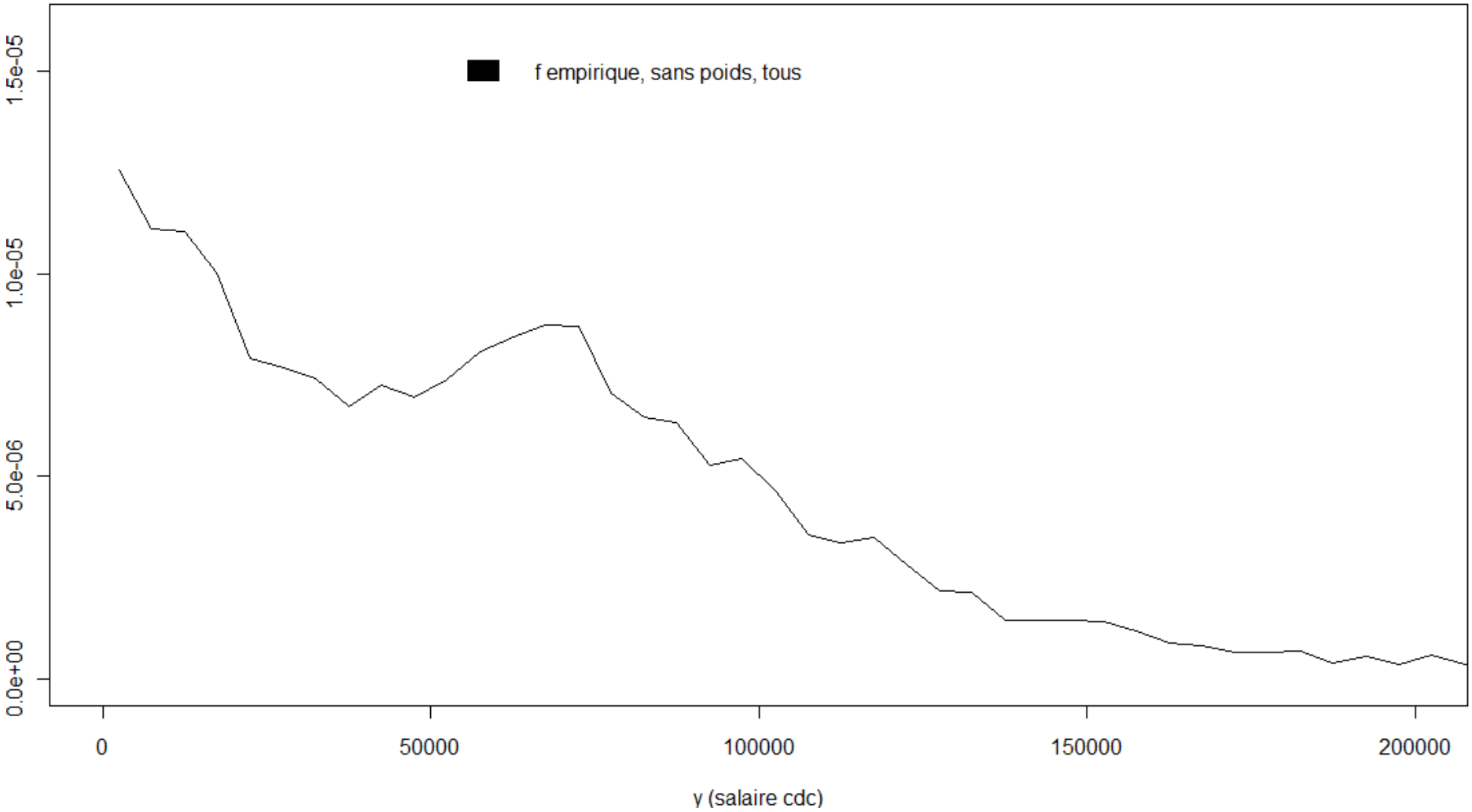
Par ailleurs, la haute corrélation entre $med_{y_{cdc}}^{GHR}$ et $medNR_{y_{cdc}}^{GHR}$ est un signe que l'arbre de segmentation crée de bons GHR en fonction des variables explicatives de la NR à dispo.

Calage généralisé

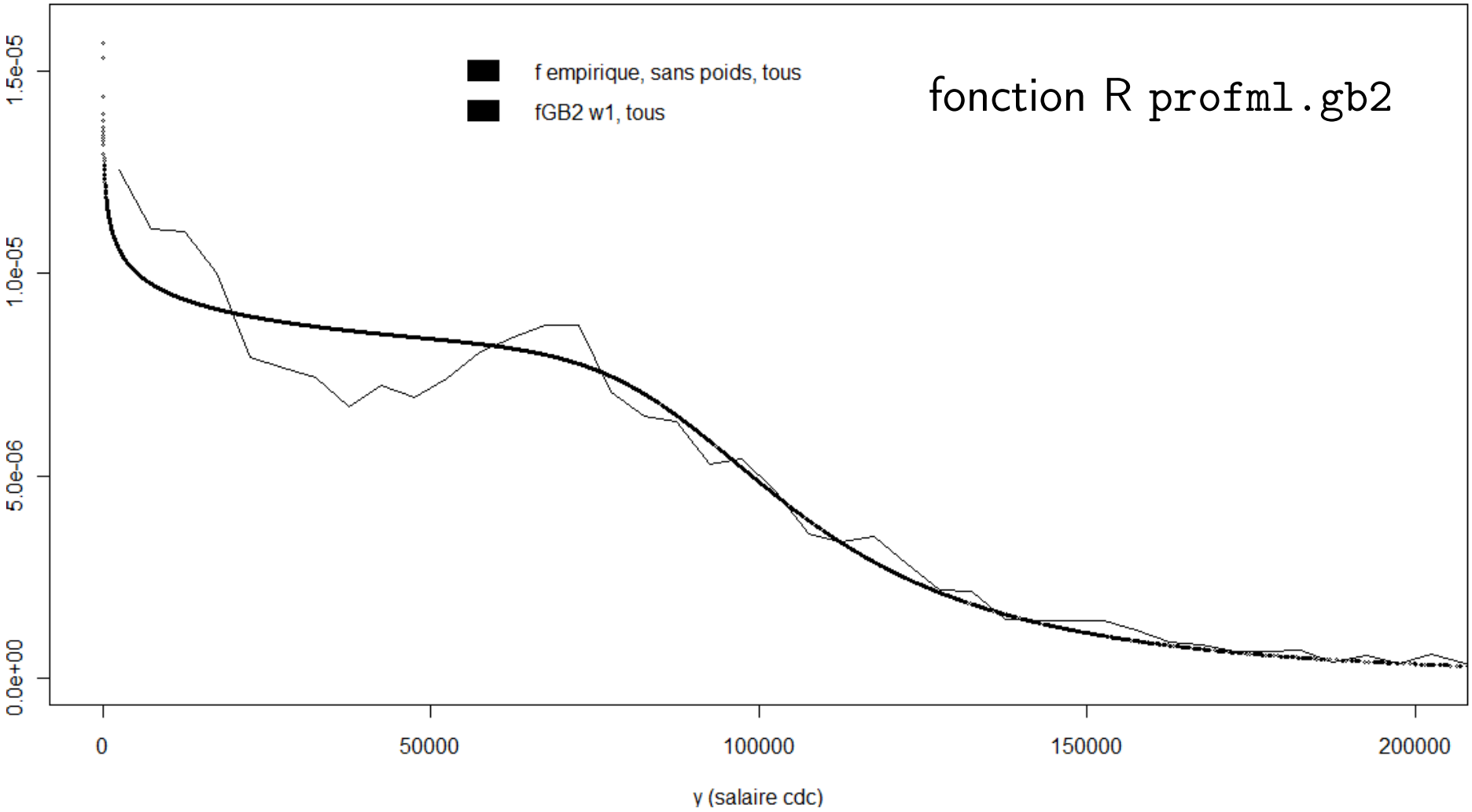
Comparaison des poids avant calage w_{1k}, à ceux obtenus par calage généralisé (ajustement logisitique).

	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
Poids avant calage w_{1k}	85.6	302.2	372.1	428.1	499.8	4583.0
Facteur d'ajustement $F^{logit}(\mathbf{z}_k'\boldsymbol{\lambda})$	1.00	1.02	1.07	1.31	1.19	11.25
Poids après calage gén. w_{2k}^{logit}	85.8	340.7	439.4	550.1	616.3	5901.0

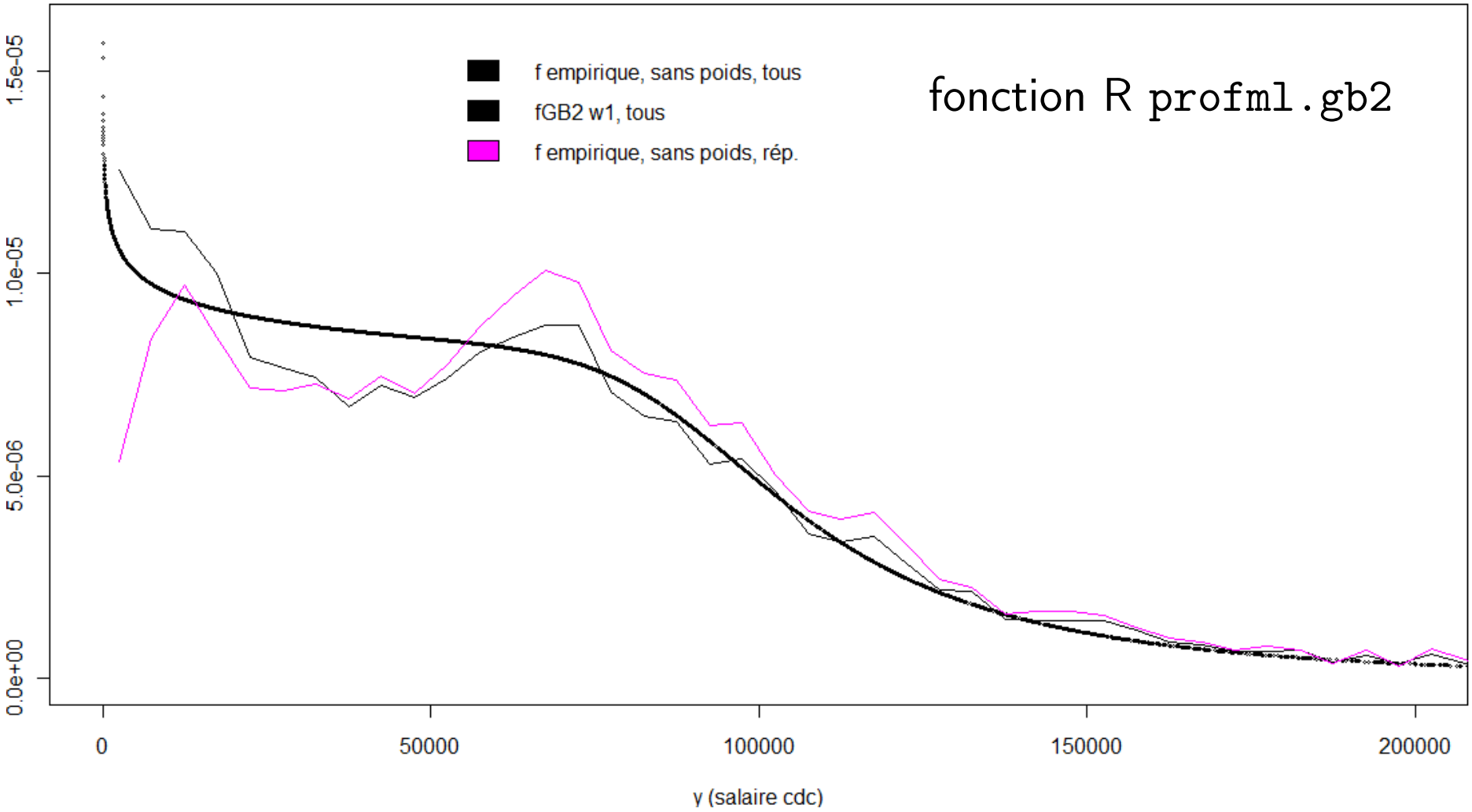
2 Ajustements GB2



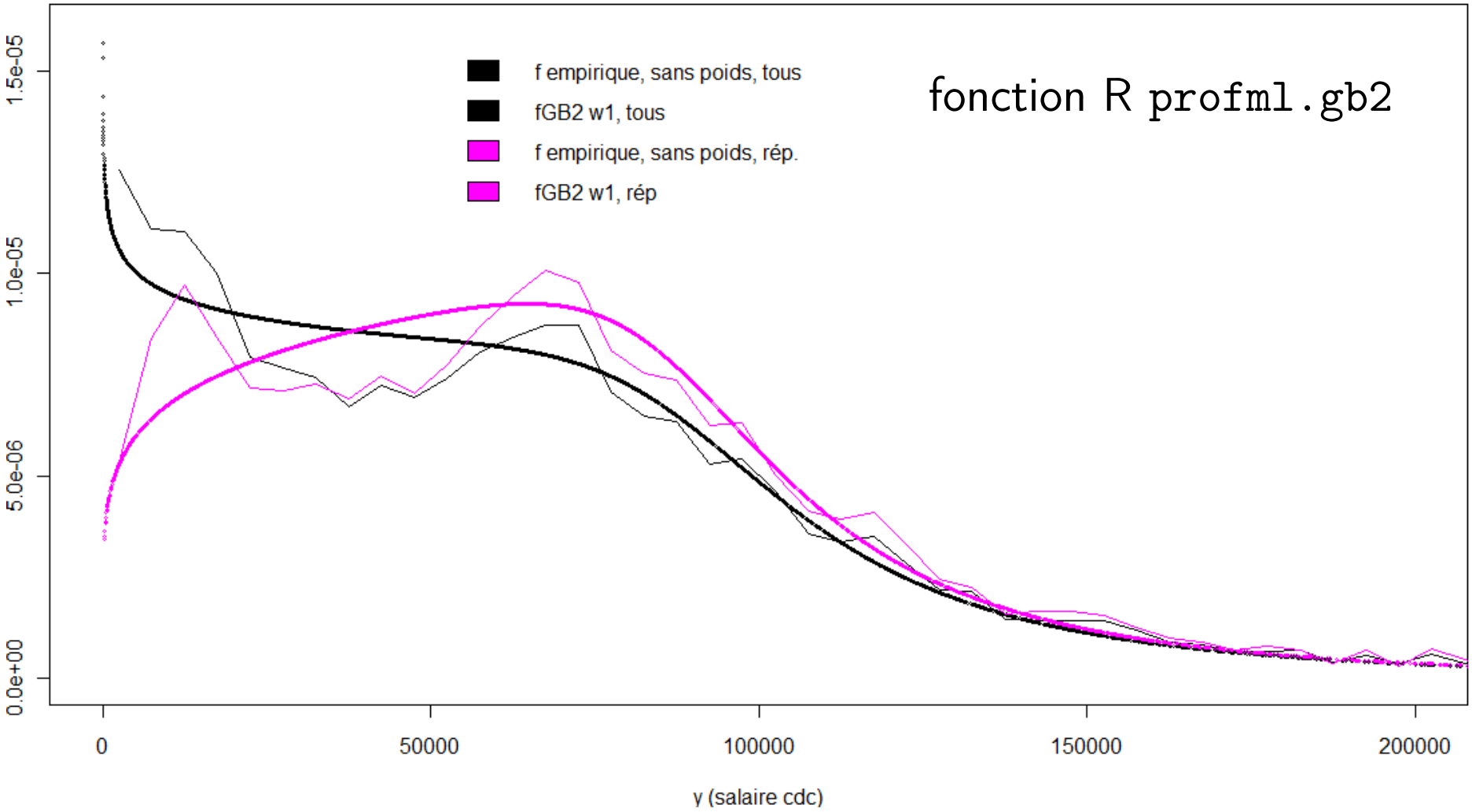
2 Ajustements GB2



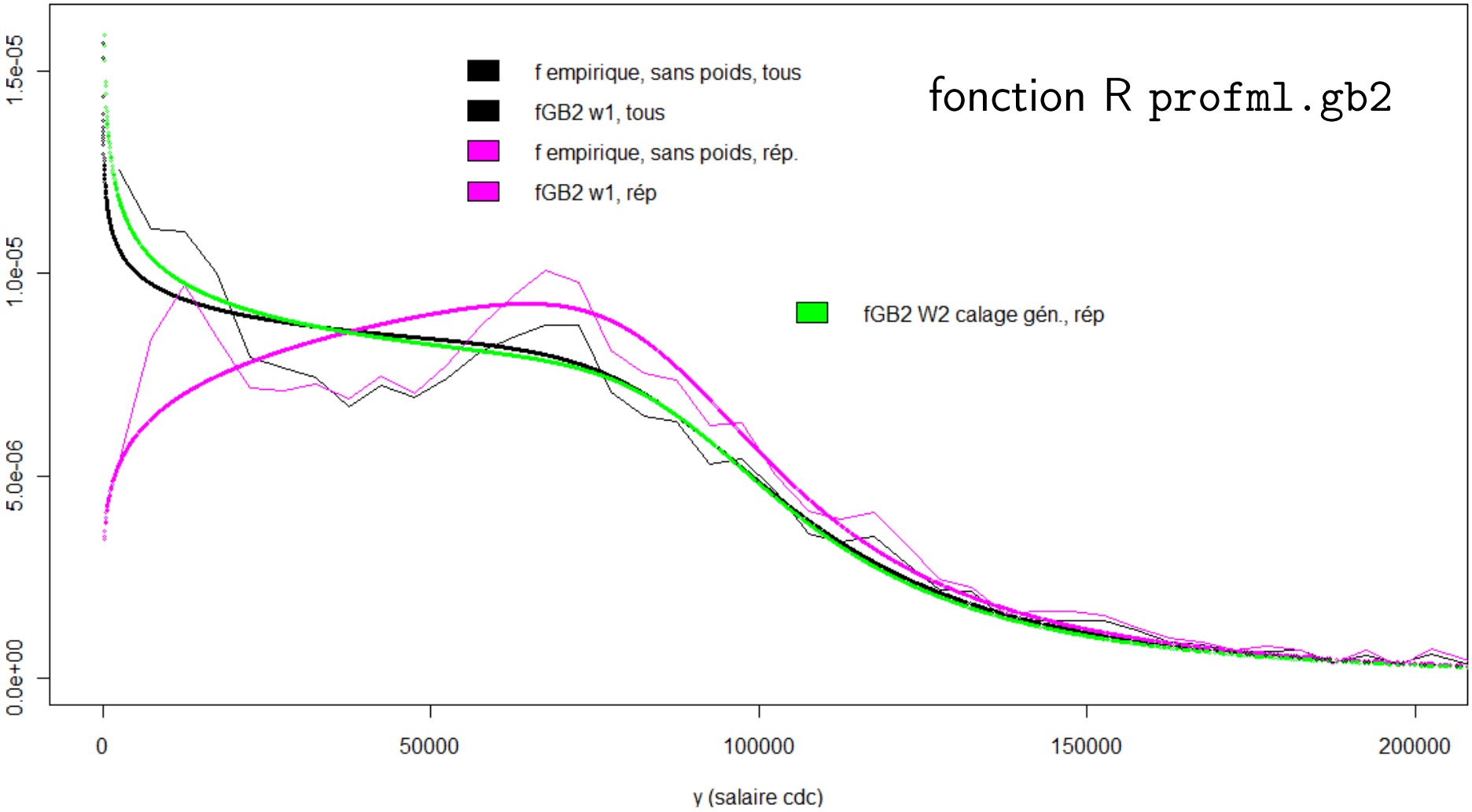
2 Ajustements GB2



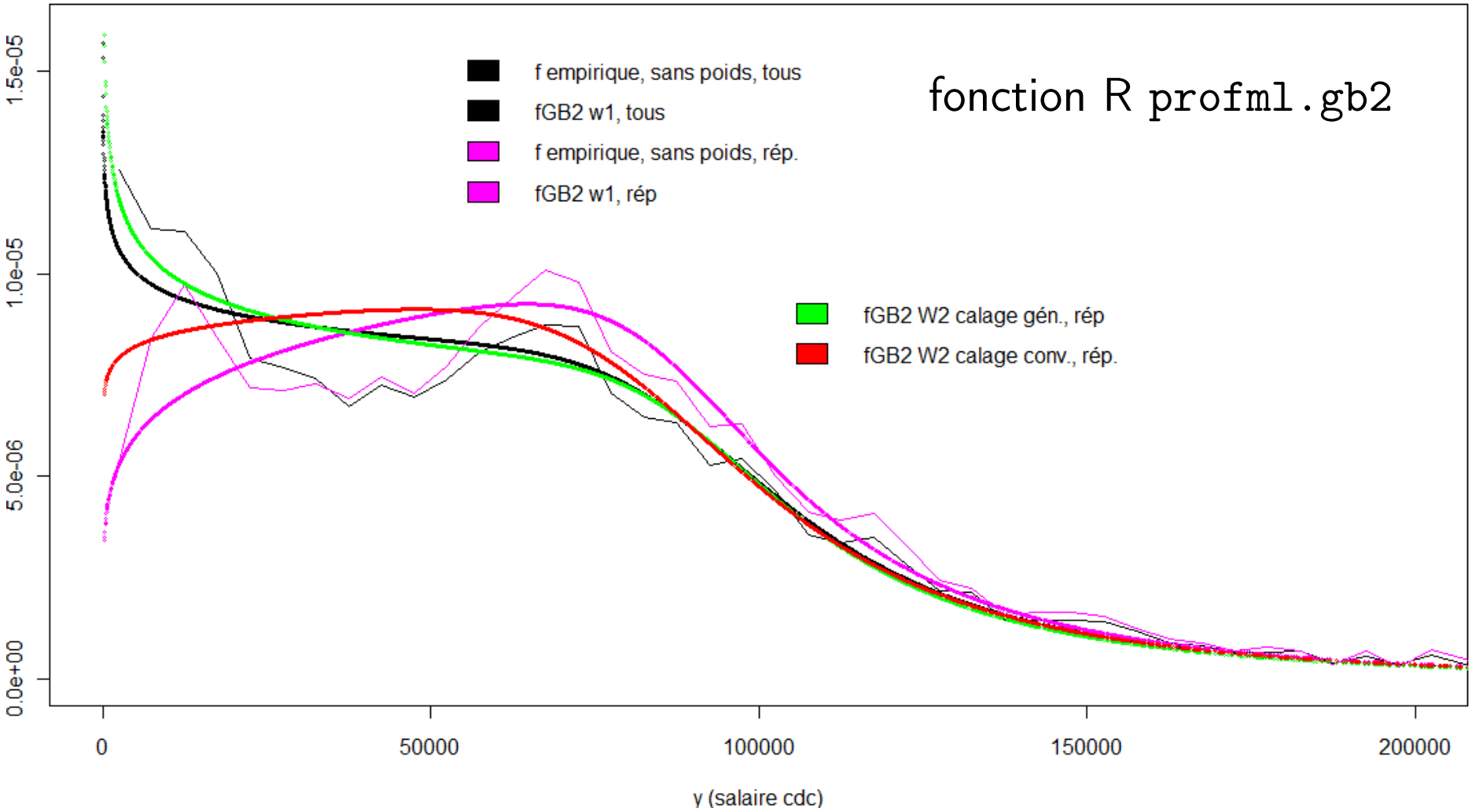
2 Ajustements GB2



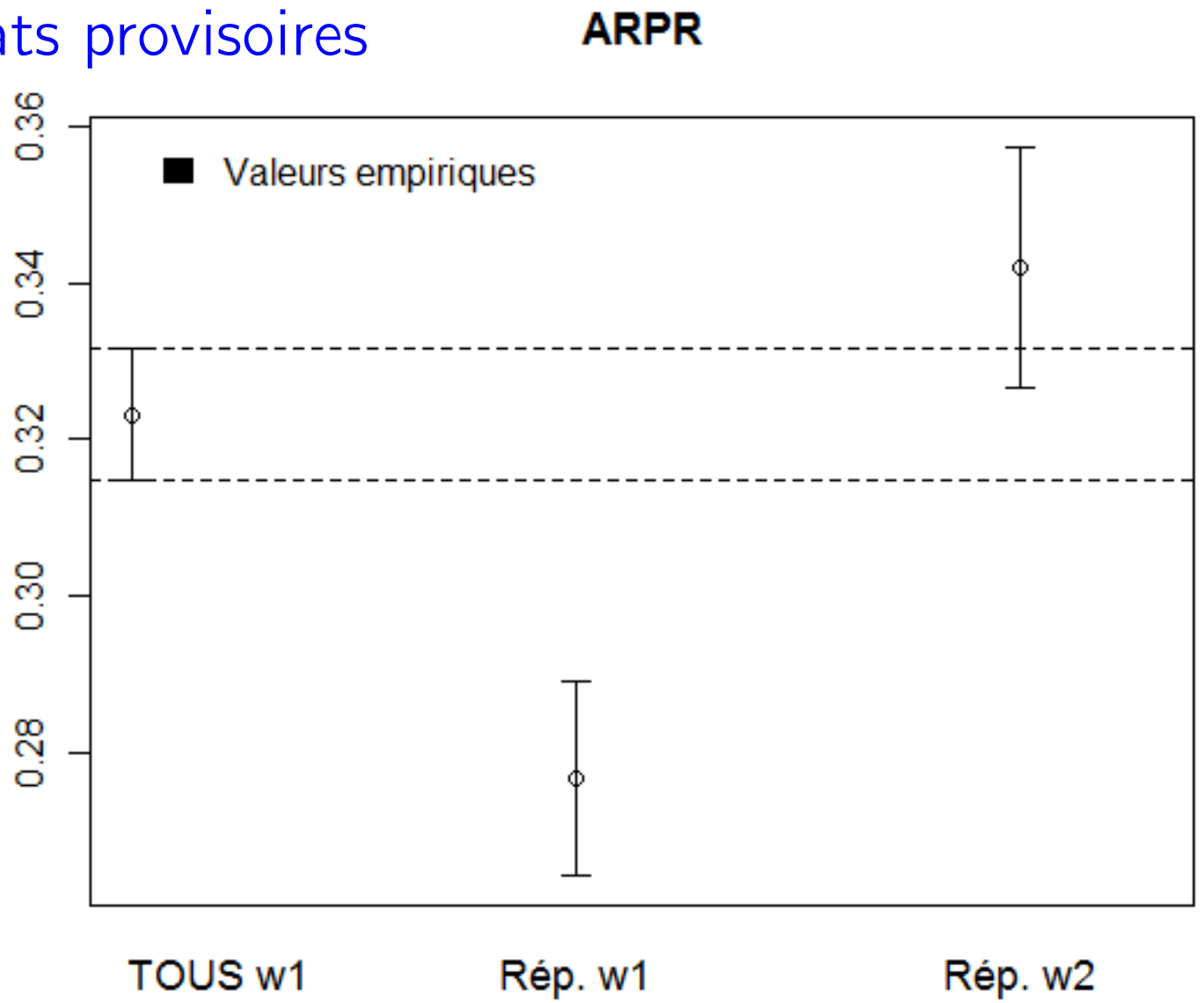
2 Ajustements GB2



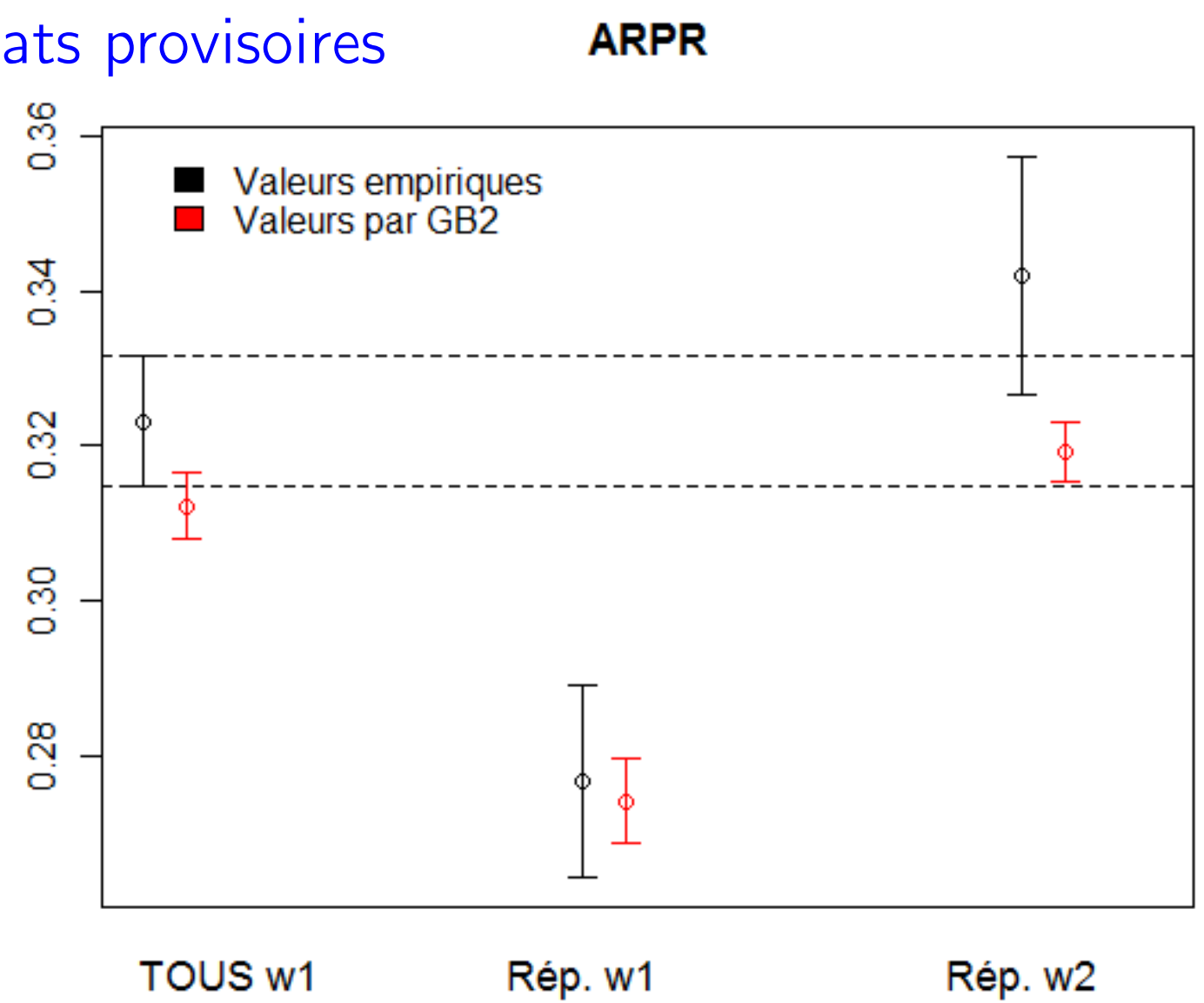
2 Ajustements GB2



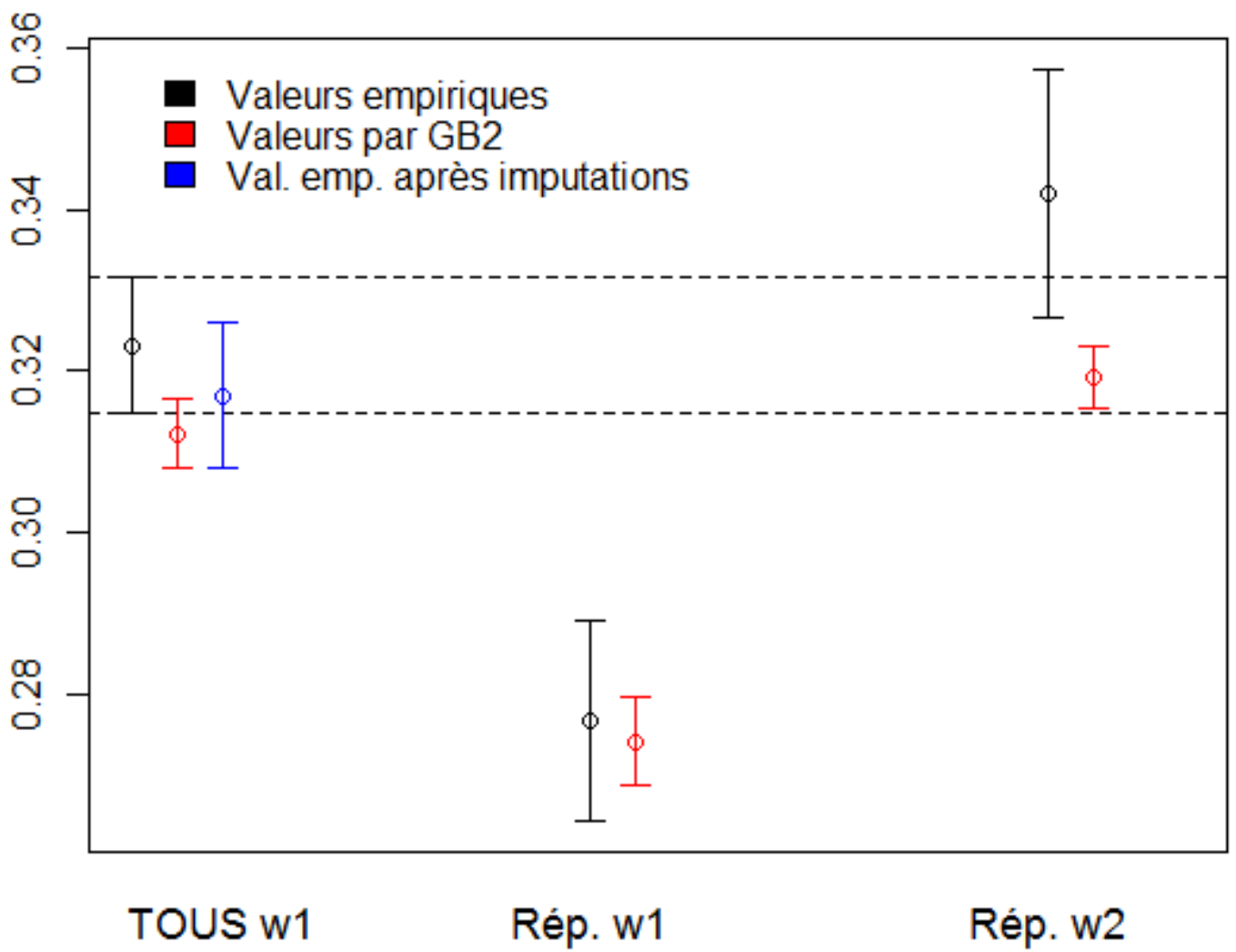
Résultats provisoires



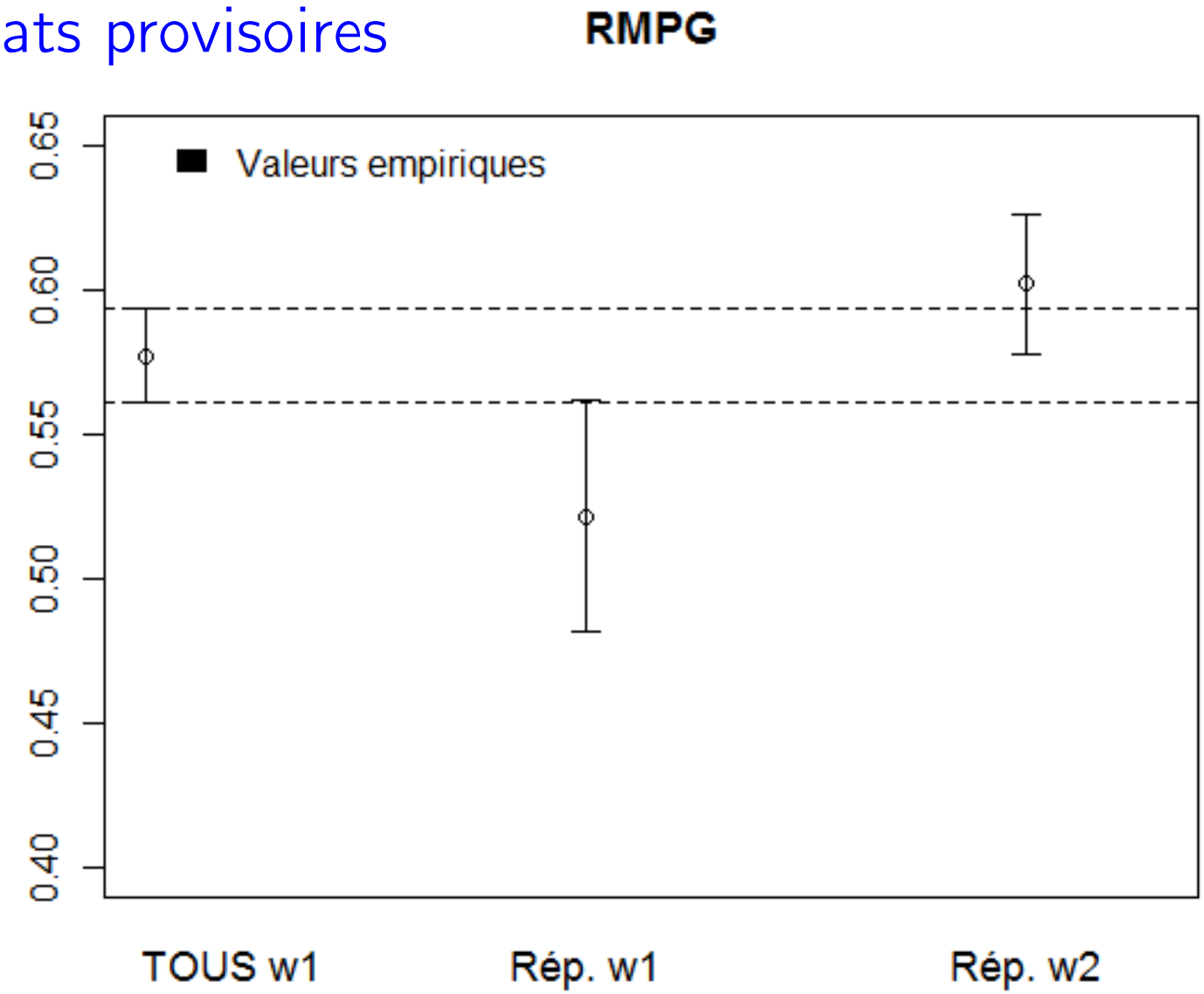
Résultats provisoires



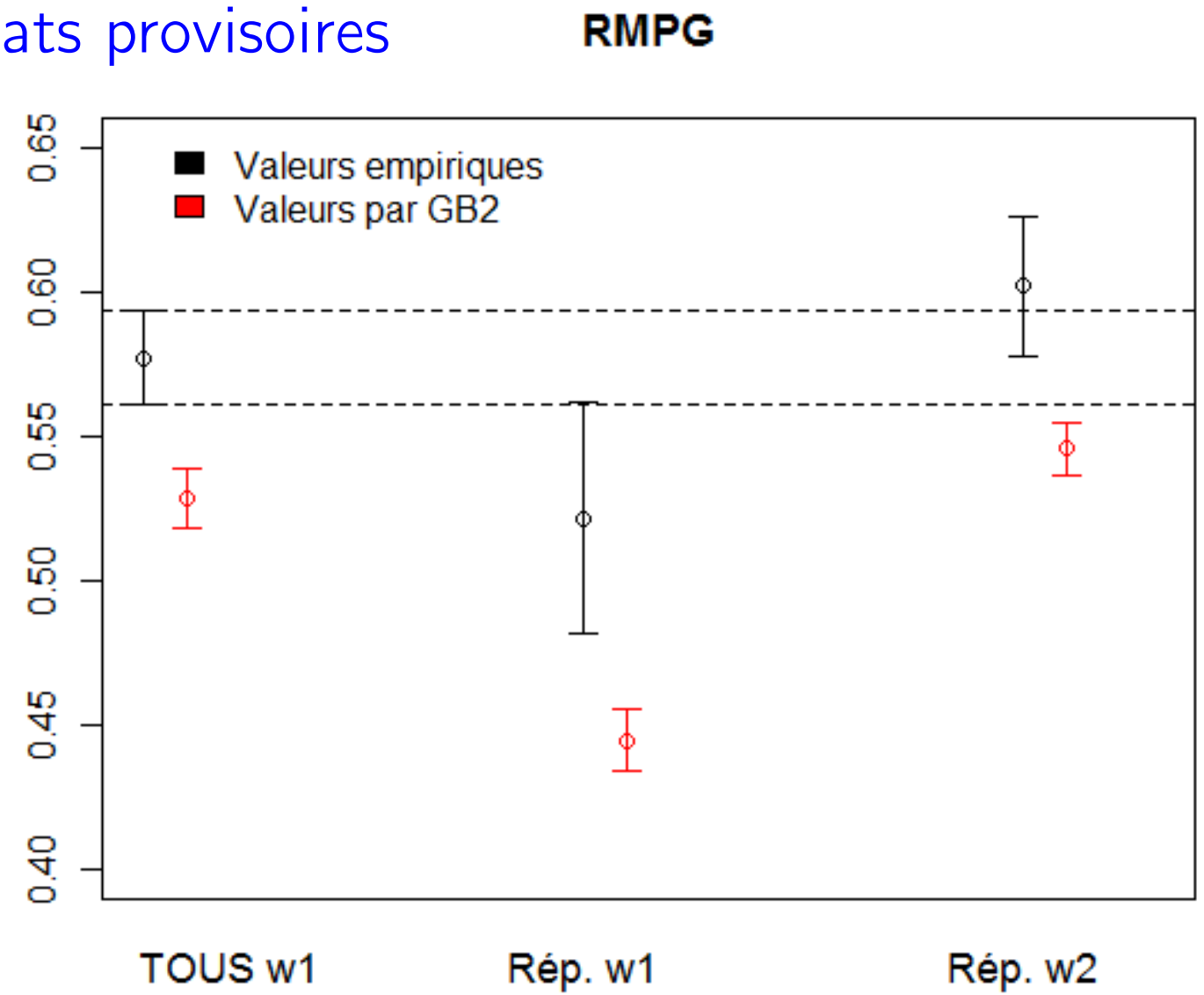
Résultats provisoires



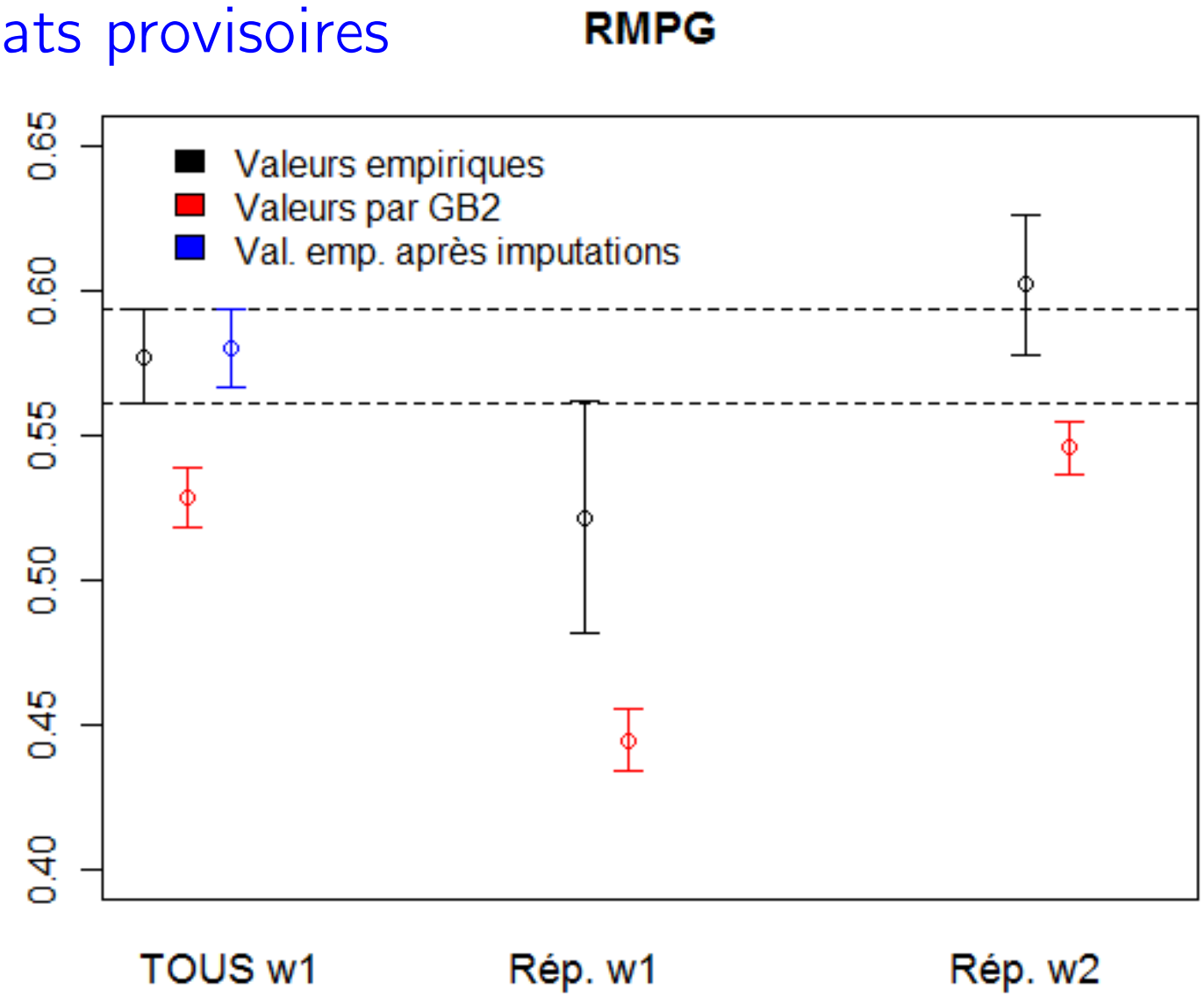
Résultats provisoires



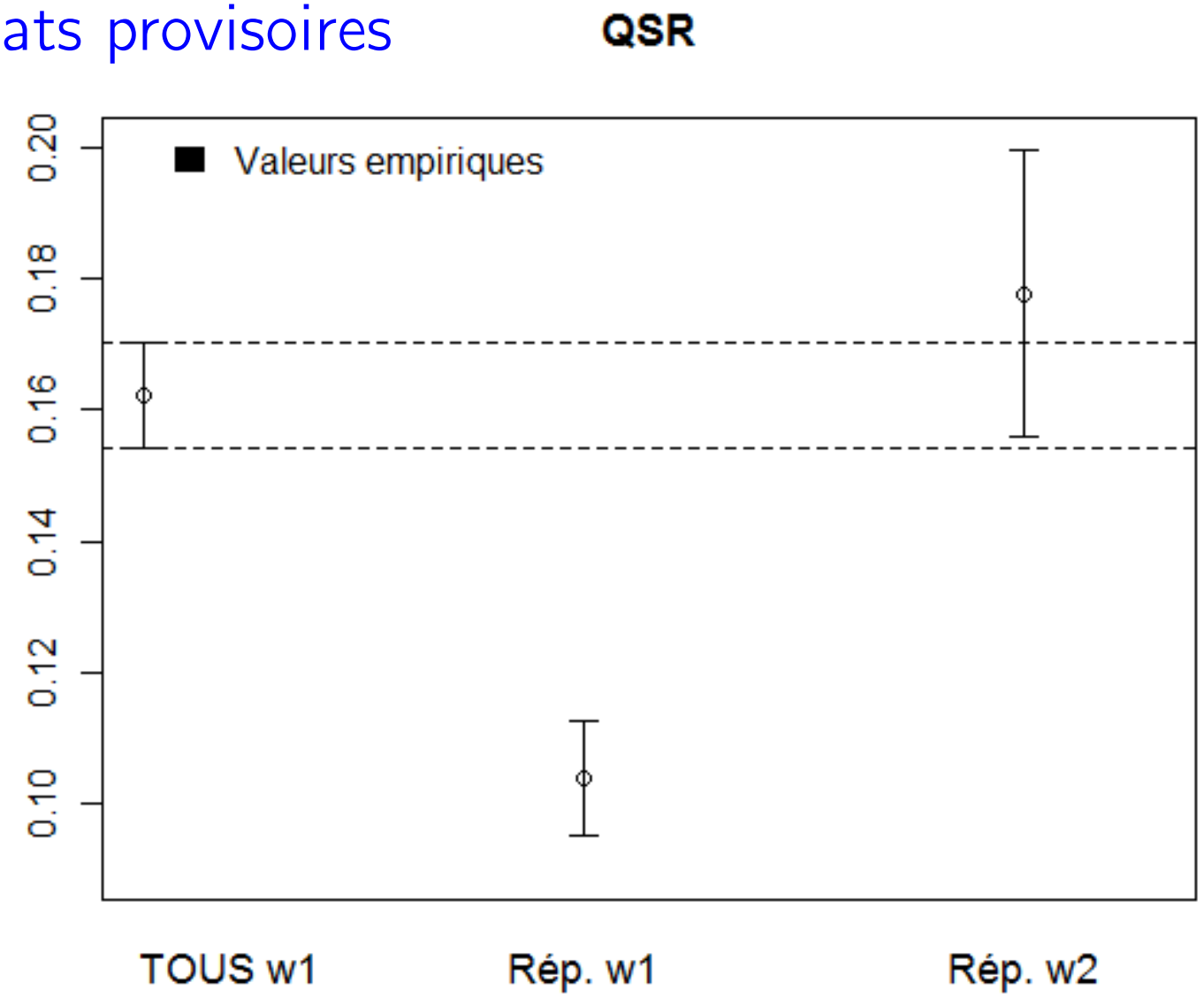
Résultats provisoires



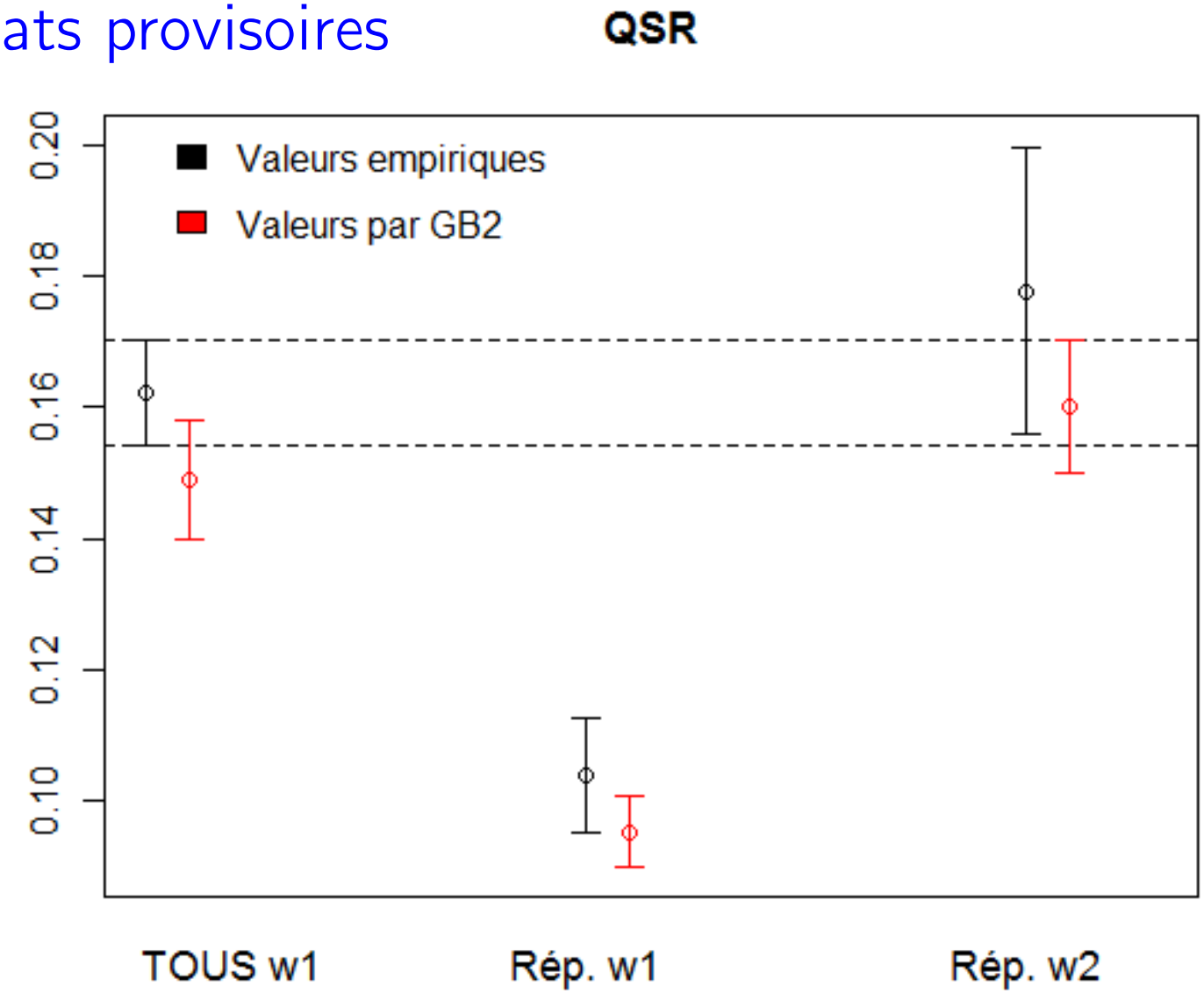
Résultats provisoires



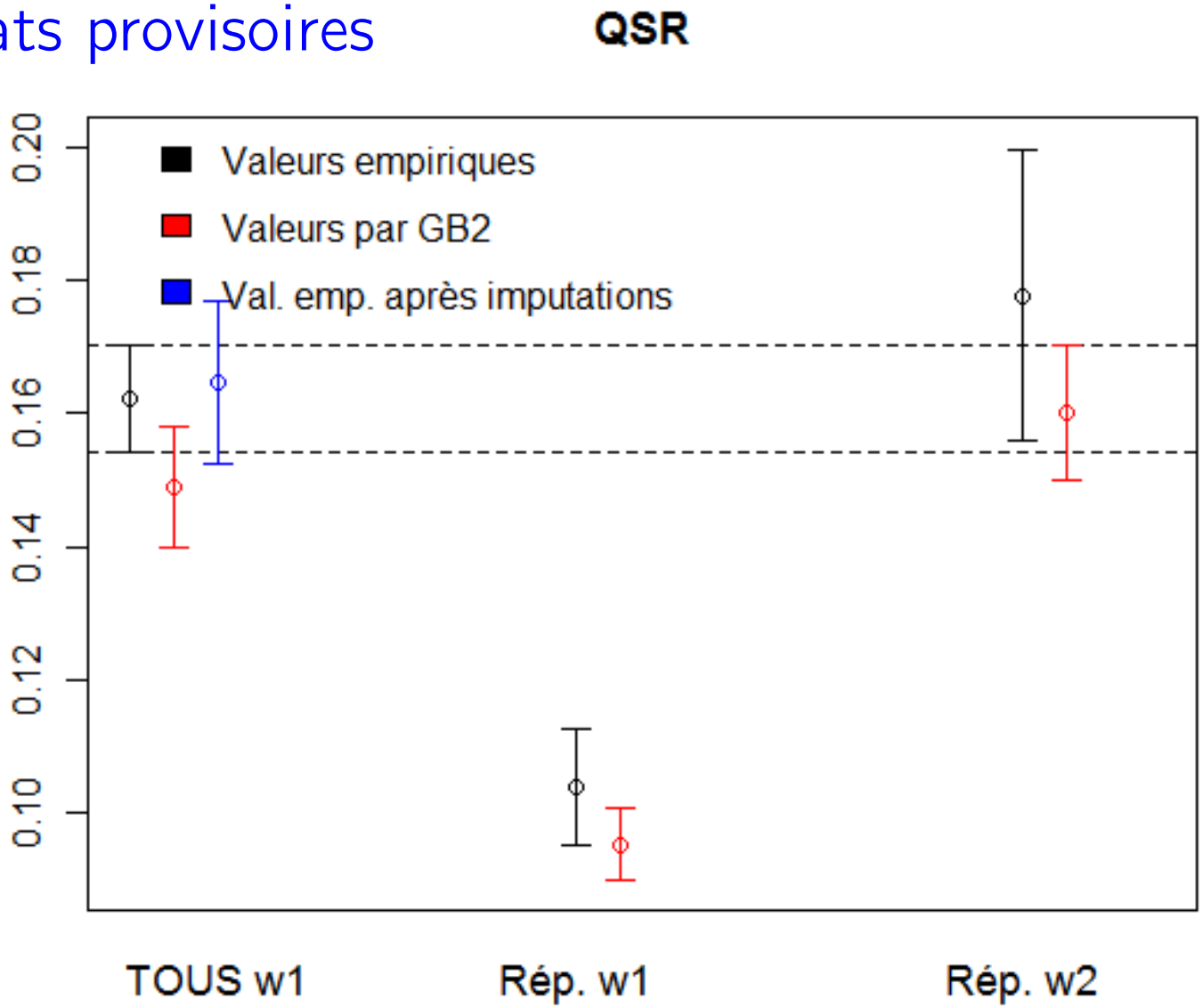
Résultats provisoires



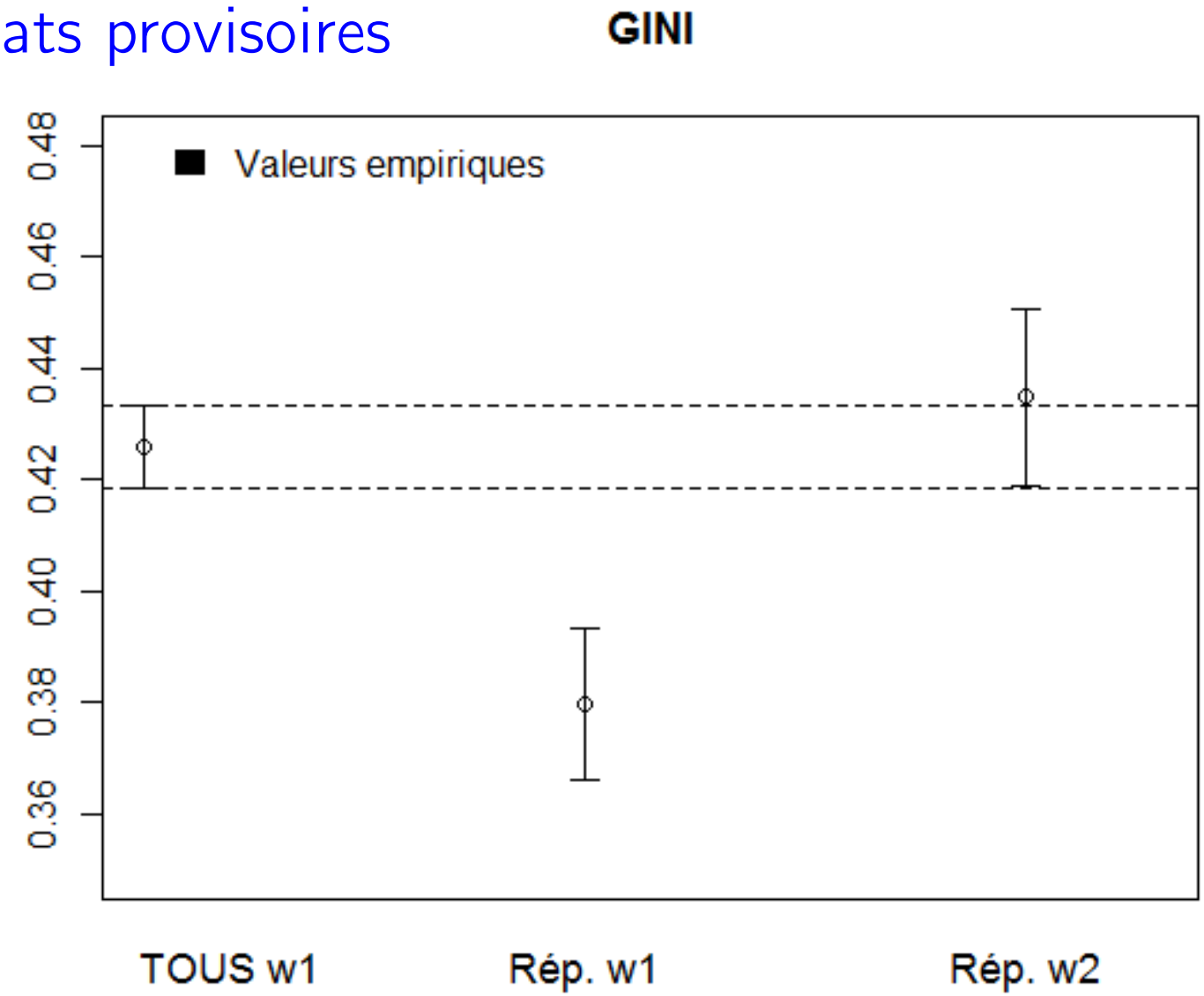
Résultats provisoires



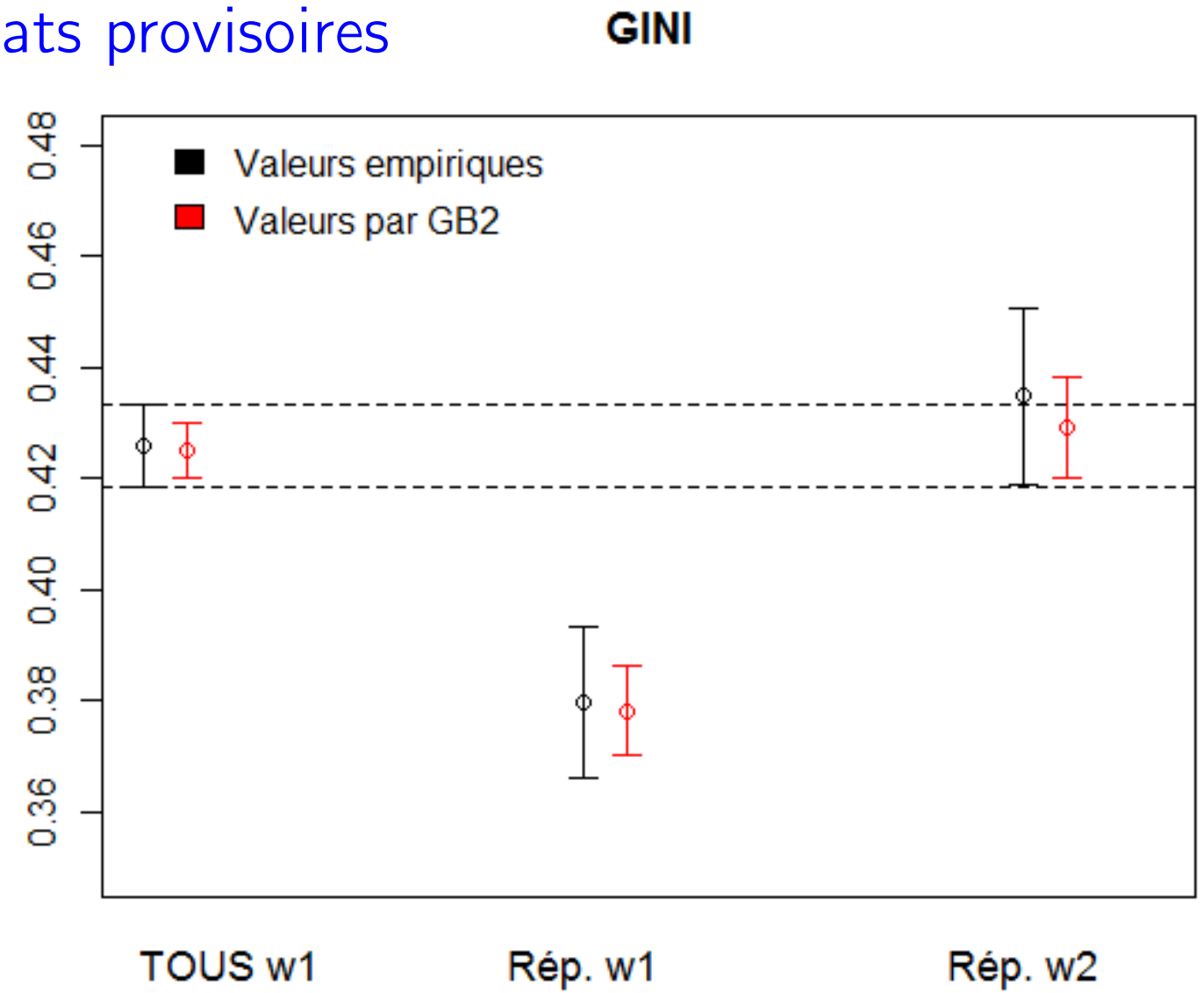
Résultats provisoires



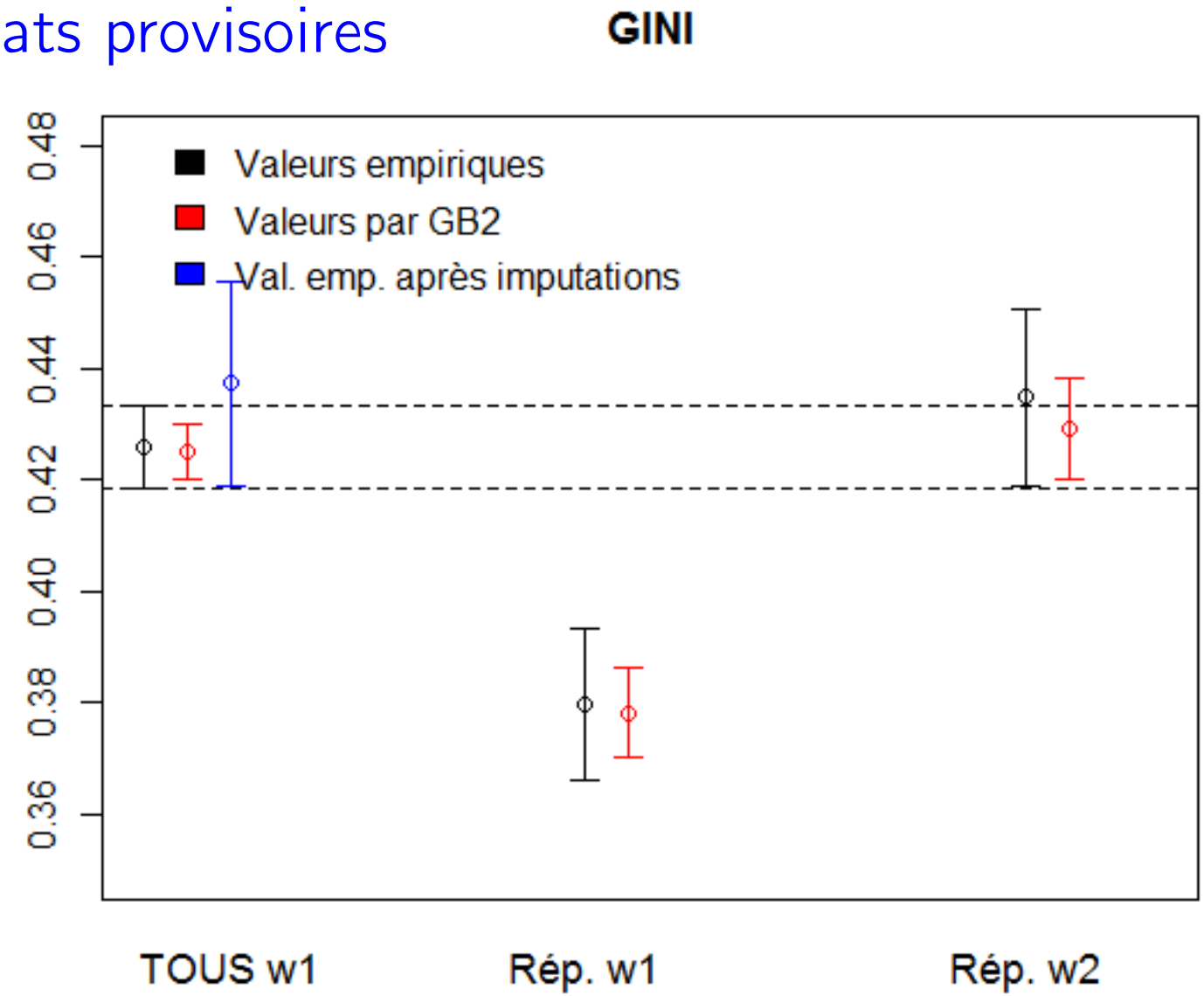
Résultats provisoires



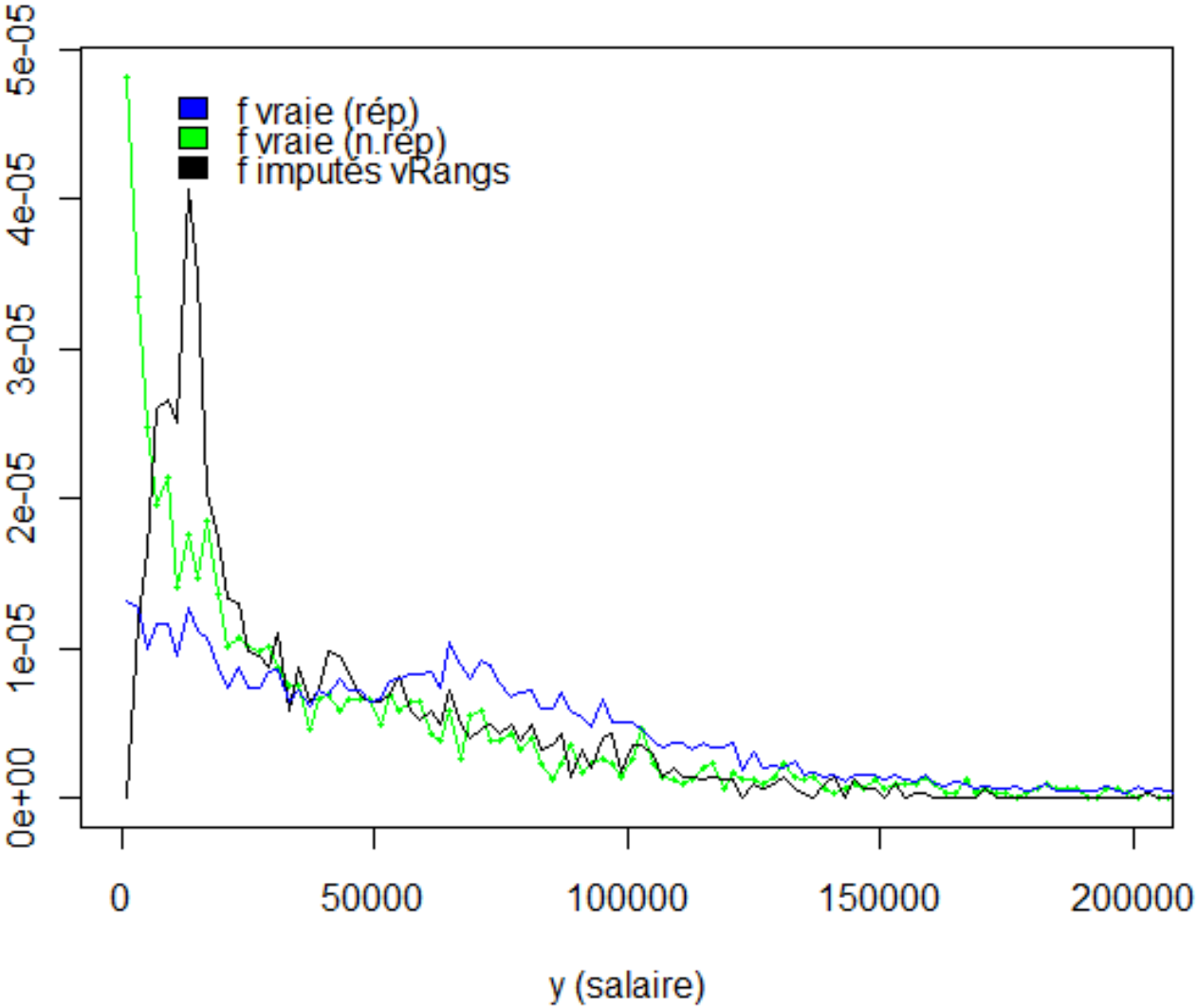
Résultats provisoires



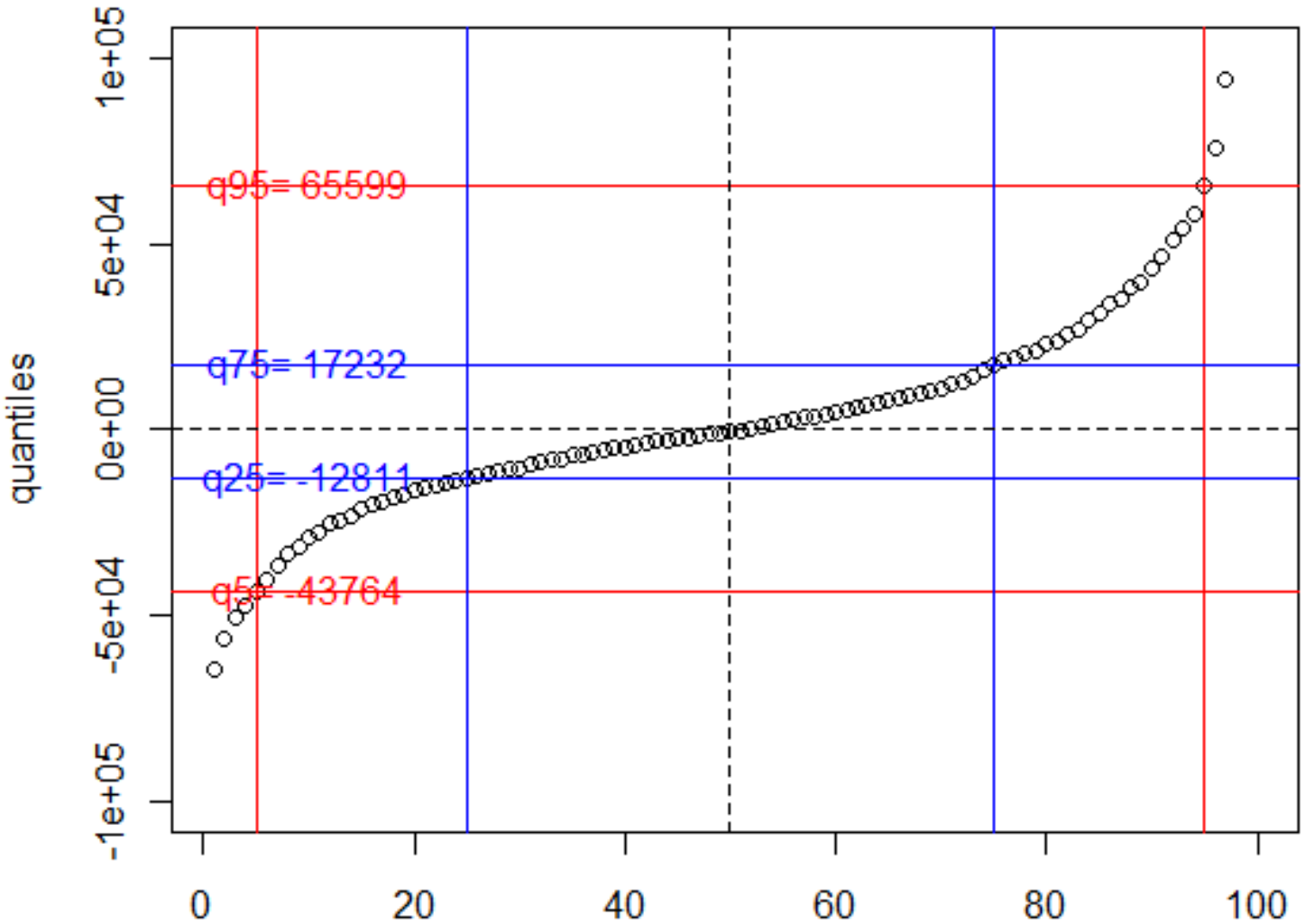
Résultats provisoires



Résultats provisoires densités empiriques



Résultats provisoires



Conclusions I

Méthode prometteuse :

- met en évidence l'importance et l'utilité d'une pondération capable de corriger pour la NR non ignorable
- s'adapte et respecte beaucoup mieux la distribution naturelle de variables de revenus
- permet, par la loi GB2 ajustée, de calculer les indices de pauvreté sans imputer
- la précision des imputations calculées au niveau unitaire dépend du pouvoir explicatif des variables auxiliaires à disposition et des poids obtenus par calage généralisé

Conclusions II

Travail en cours et futur :

- Le choix des instruments pour le calage généralisé est crucial. Comment trouver les bons instruments parmi les variables à disposition ? (Tests de Durbin-Hausman-Wu...)
- Calage généralisé et régression instrumentale dans le contexte de la méthodologie d'enquête
- Calculs des variances des indices et des variances dues à l'imputation à peaufiner
- Simulations

Références citées

- AMELI (2011), Deliverable 2.1, Graf, M., Nedyalkova, D., Münnich, R., Seger, J., Zins, S.. *Parametric Estimation of Income Distributions, and Indicators of Poverty and Social Exclusion*.
- Dastrup, S. R., Hartshorn, McDonald, J. B.. Journal of Economic Inequality. *The impact of taxes and transfer payments on the distribution of income : A parametric comparison*.
- Deville, J.-C. (2002). Actes des Journées de Méthodologie Statistique INSEE. *La correction de la non-réponse par calage généralisé*.
- Deville, J.-C., Särndal, K. E. (1992). Journal of ASA. *Calibration Estimators in Survey Sampling*.
- Deville, J.-C. (1993). Document interne, INSEE. Calage, calage généralisé et hypercalage
- Deville, J.-C., Särndal, K. E., Sautory, O. (1993). Journal of ASA.*Generalized Raking Procedures in Survey Sampling*.
- Graf, M. (2009). JSM Proceedings. *An Efficient Algorithm for the Computation of the Gini Coefficient of the Generalized Beta Distribution of the Second Kind*.
- Jenkins, S. P. (2007). *Inequality and the GB2 Income Distribution*.
- Kleiber, C., Kotz, S. (2003). Wiley. *Statistical Size Distributions in Economics and Actuarial Sciences*.
- Kott, P.S. (2006). Survey Methodology. *Using calibration weighting to adjust for nonresponse and coverage errors*.
- McDonald James B. (1984). Econometrica. *Some Generalized Functions for the Size Distribution of Income*.
- Sautory, O. (2003). Symposium StatCan. *Calmar 2 : une nouvelle vers. du pgm. CALMAR de redr. d'échantillons par calage*.