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

Eric Graf & Yves Tillé

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

7 novembre 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
- Dans les enquêtes par échantillonnage auprès des ménages et des personnes, les questions sur le revenu sont sensibles et sujettes à un taux de non réponse (NR) plus élevé
- La NR partielle s'ajoute à la NR totale
- Sans traitement, les mesures d'inégalités calculées uniquement sur les répondants risques d'être biaisées

Motivation II

- Volonté de fournir des jeux de données complets, i.e. sans valeurs manquantes, à EUROSTAT et aux utilisateurs
- La distribution des revenus n'est pas normale ni log-normale !
- 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.
- 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

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
- 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

Loi bêta généralisée de seconde espèce (GB2)

- distribution à quatre paramètres : $GB2(a, b, p, q)$.
- a été développée par McDonald (1984).
- densité : $$f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p,q)} \frac{(y/b)^{ap-1}}{(1+(y/b)^a)^{p+q}}$$
  où $B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$ est la fonction bêta.

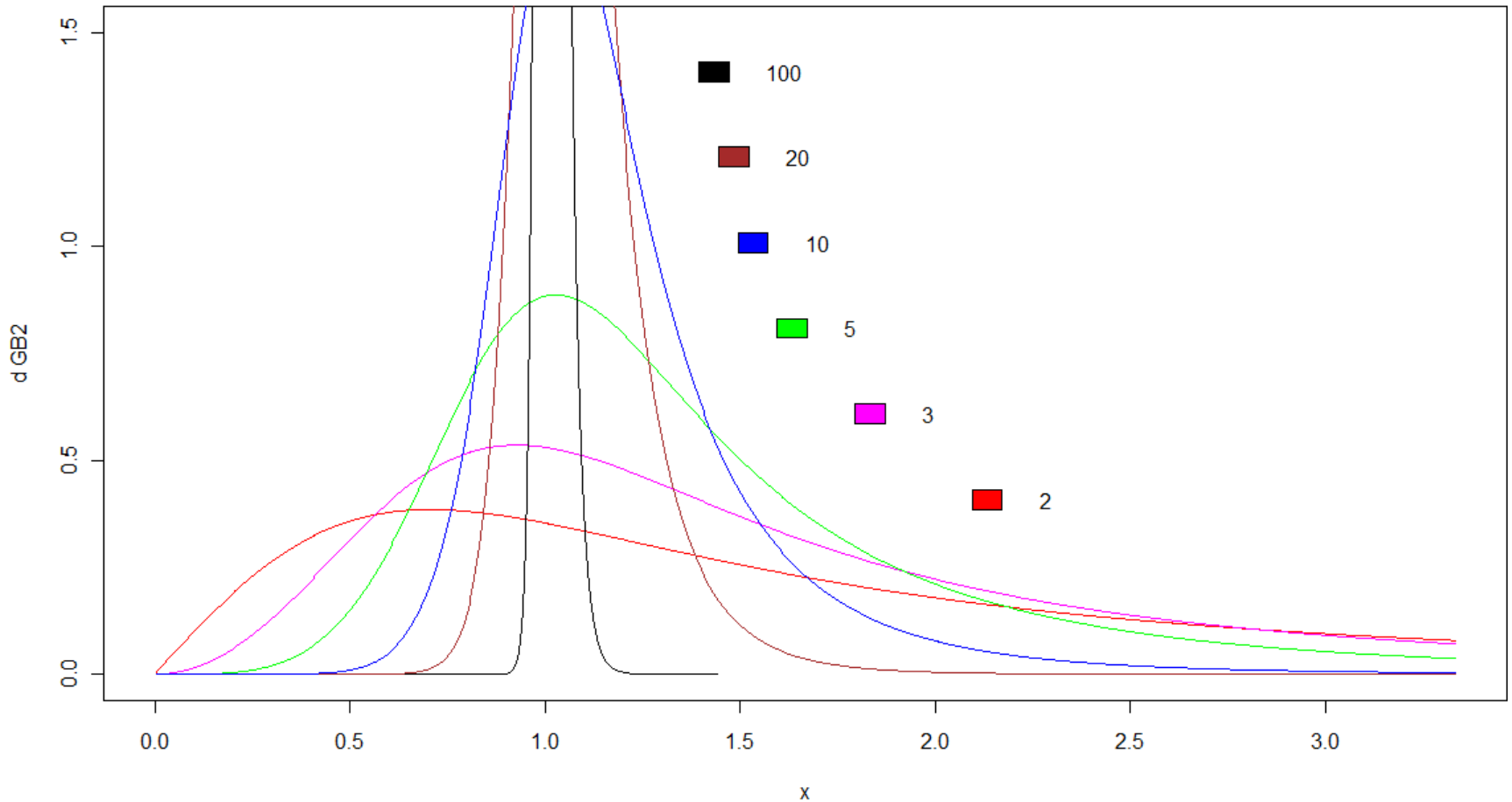
Des études empiriques sur le revenu - voir p. ex. Jenkins (2007) ; Dastrup et al. (2007) ; Kleiber et Kotz (2003) ; Sepanski et Kong (2007) - montrent que la GB2 s'ajuste bien à de telles données et qu'elle est souvent plus adaptée que d'autres distribution à quatre paramètres.

Résultats du projet européen AMELI (2011) confirment pour EU-SILC.

Loi GB2 II

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

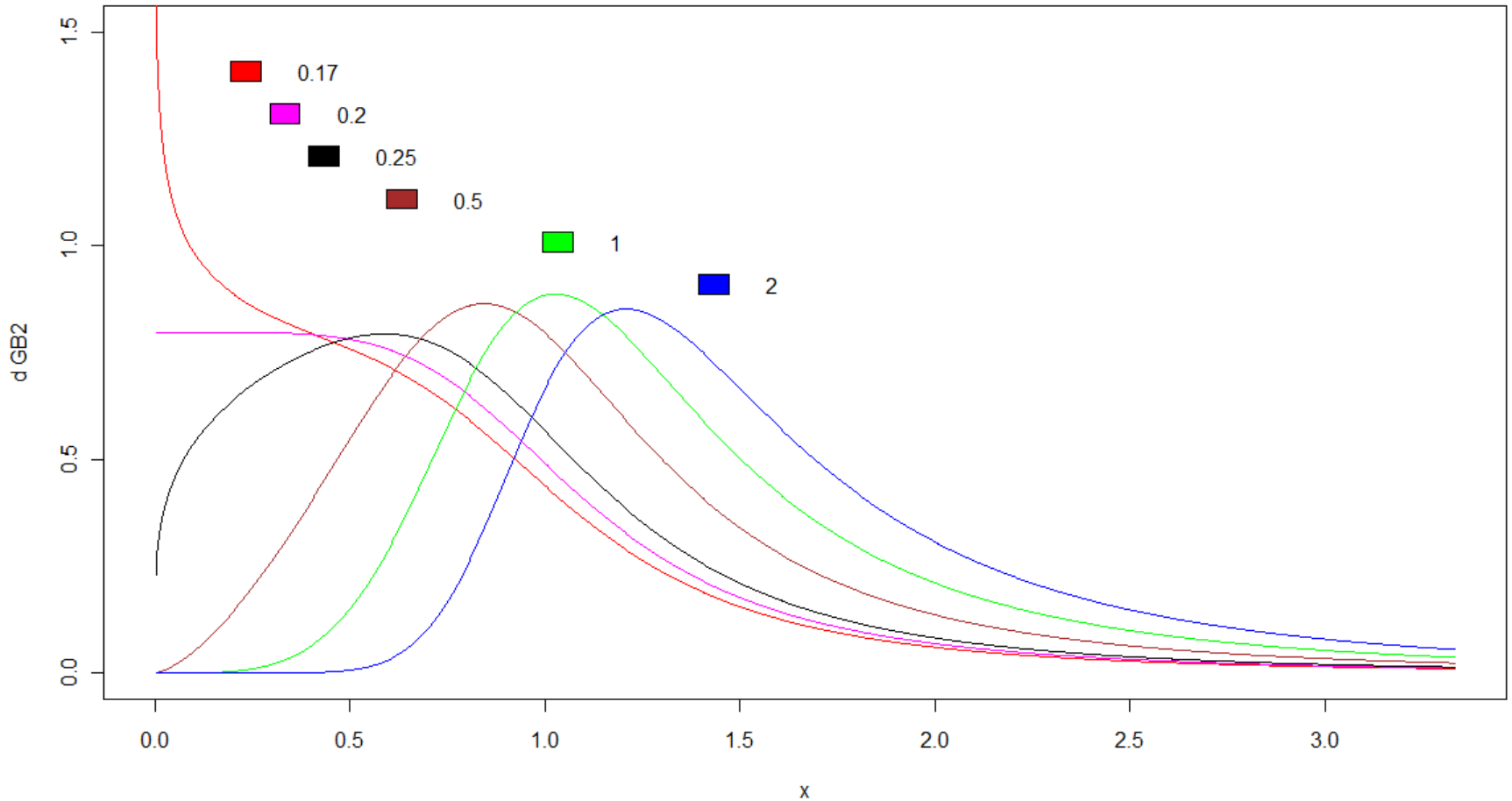
$\rightarrow$ distribution $\pm$ pointue



Loi GB2 III

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

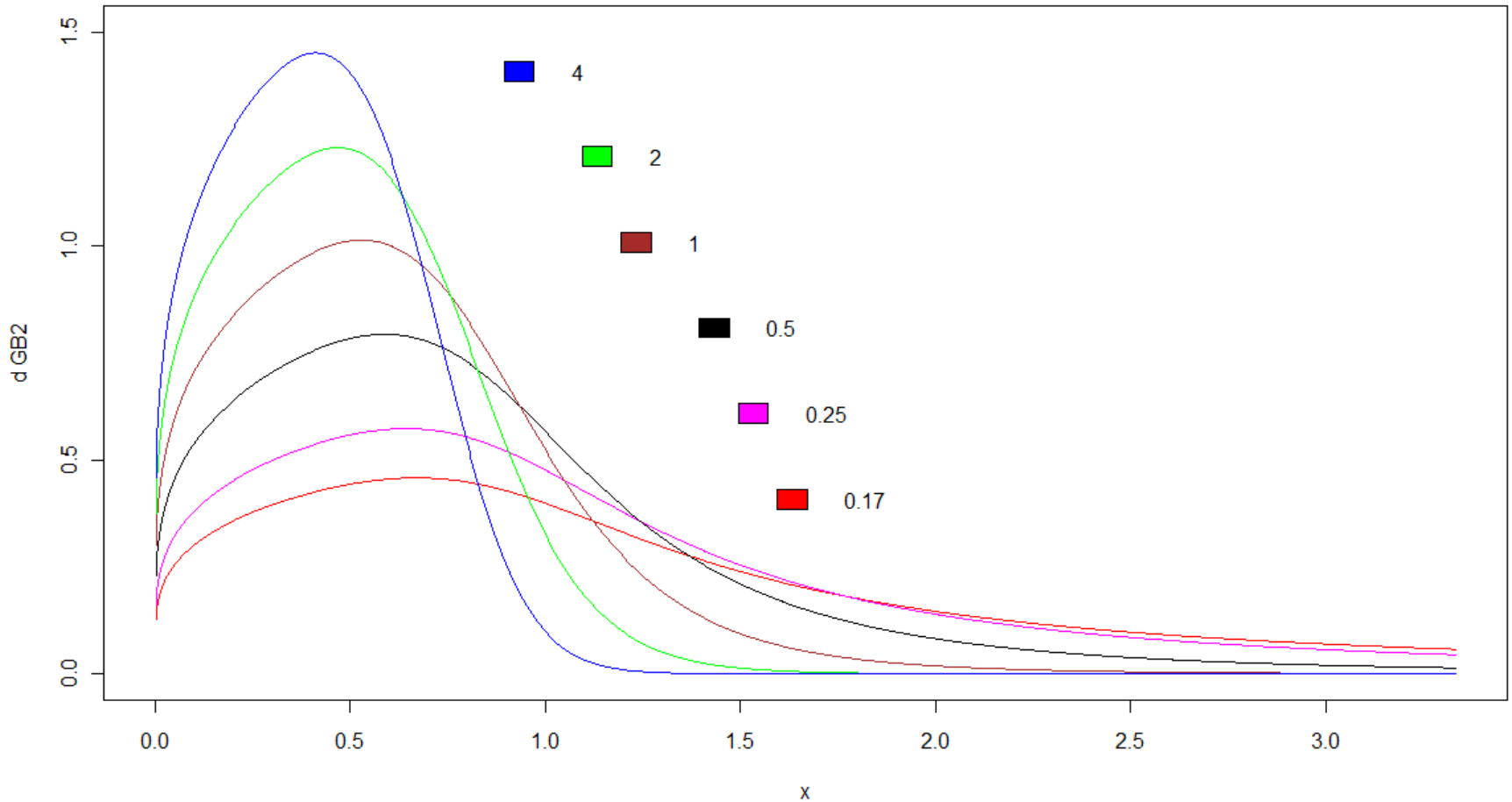
$\rightarrow$ on joue sur la forme de la queue de gauche



Loi GB2 IV

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

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



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

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

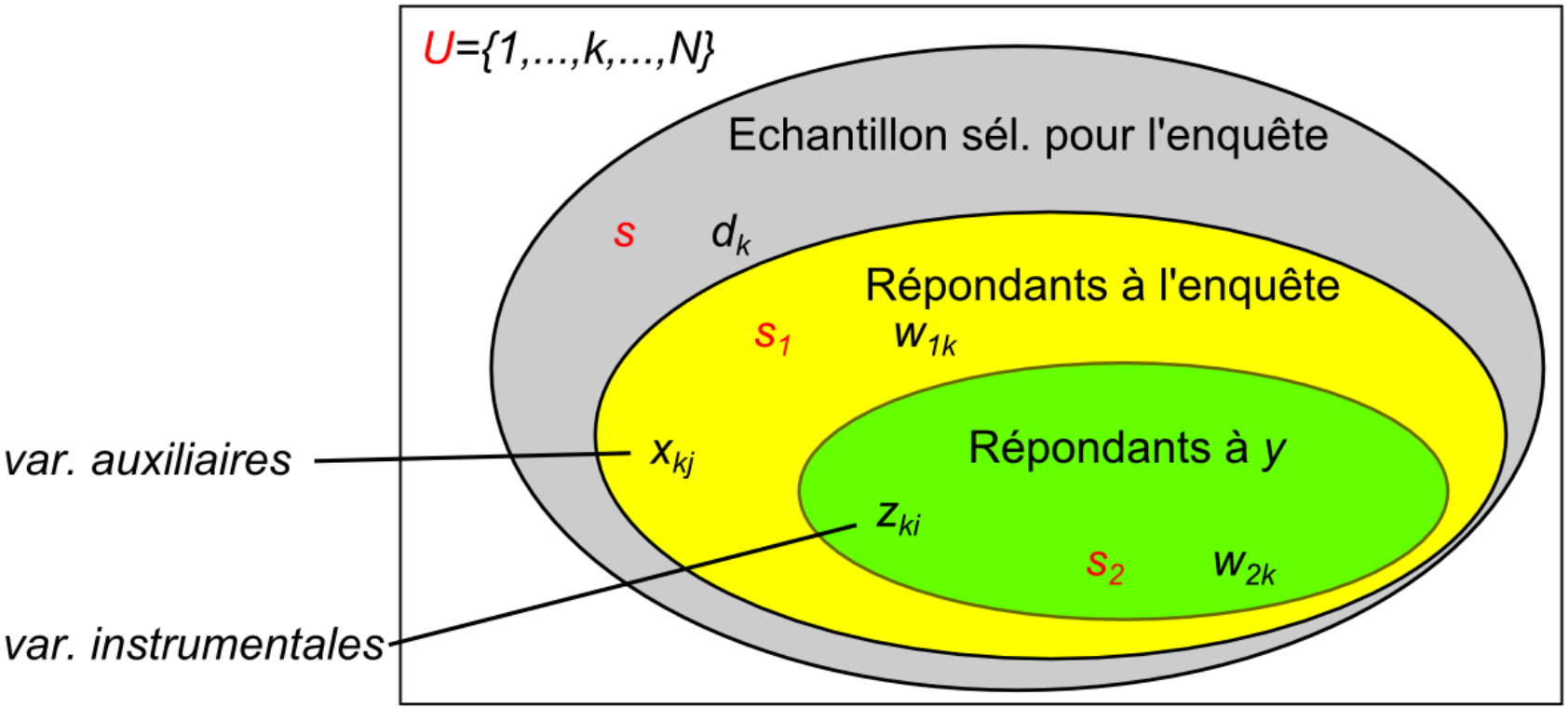
Classiquement, on distingue trois types de mécanismes de NR pouvant affecter la variable d'intérêt : MCAR, MAR, NMAR.
La plupart des variables de revenus que nous voulons imputer sont NMAR ...

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 « proches »des poids avant calage, spécifiques à y , corrigeant pour la NR et respectant certaines contraintes.

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).



On observe $(y_k, \mathbf{x}_k, \mathbf{z}_k)$, dispose de $\mathbf{t}_x$, et suppose que $J = I$.

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 pour tout échantillon s_2

$$\min_{w_{2k}} \sum_{k \in s_2} \frac{G_k(w_{2k}, w_{1k})}{q_k}$$

sous les contraintes

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

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

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

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 résout pour $\boldsymbol{\lambda}$ de manière à ce que les w_{2k} satisfassent les contraintes.

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 (Centrale de Compensation)	nombre d'observations
P09I57G_cati= y_{cati}	P09I57G_cdc= y_{cdc}	7922

Fichier d'entraînement : individus appariés au registre

> 0	> 0 & *ne sait pas, pas de réponse, NR à la variable filtre, question filtrée*	6884 (86.9%)
taux d'occupation> 0 et *coûts du logement*> 0		6188 (78.1%)

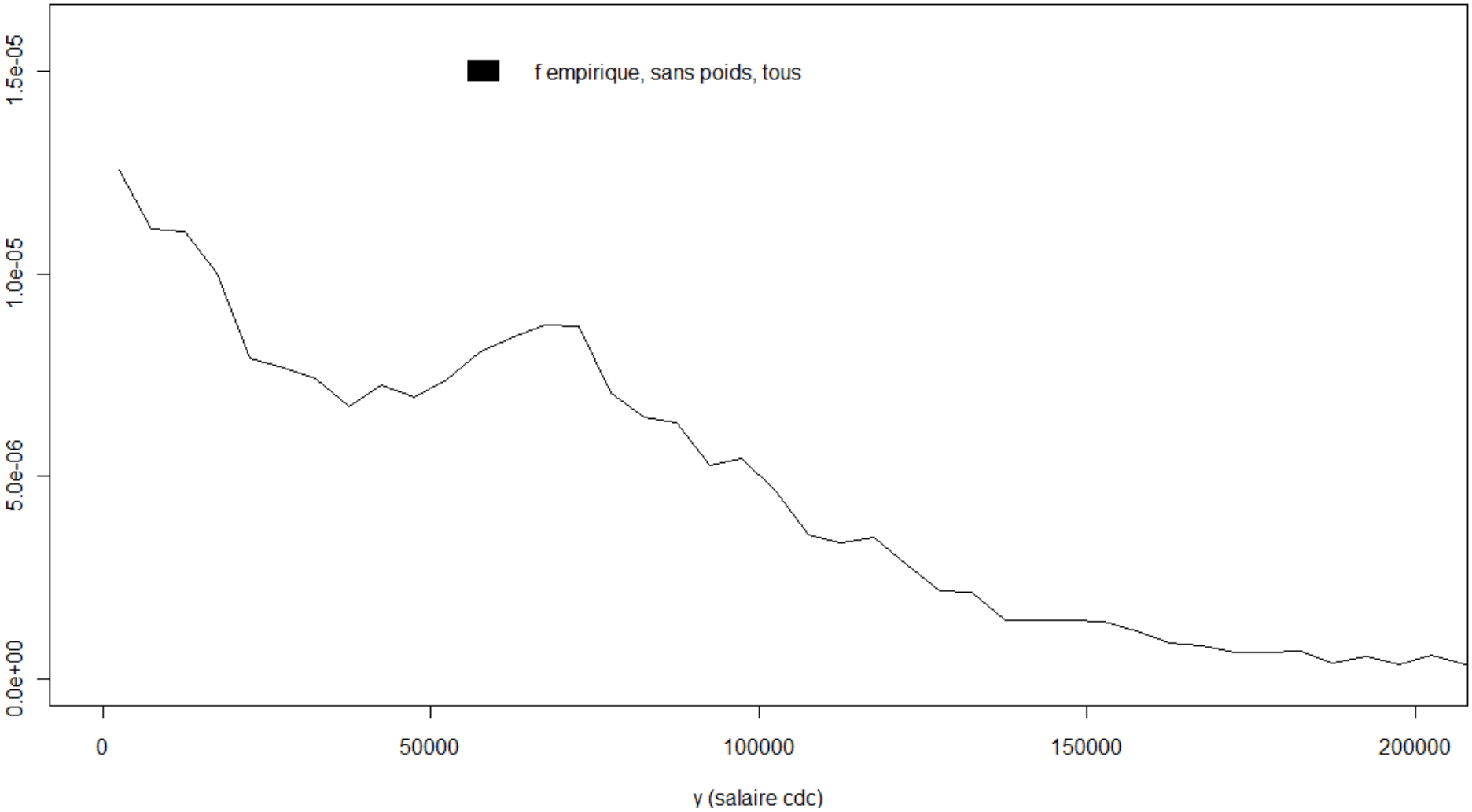
On applique ce mécanisme de NR partielle réel à y_{cdc} dont on connaît toutes les valeurs.

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

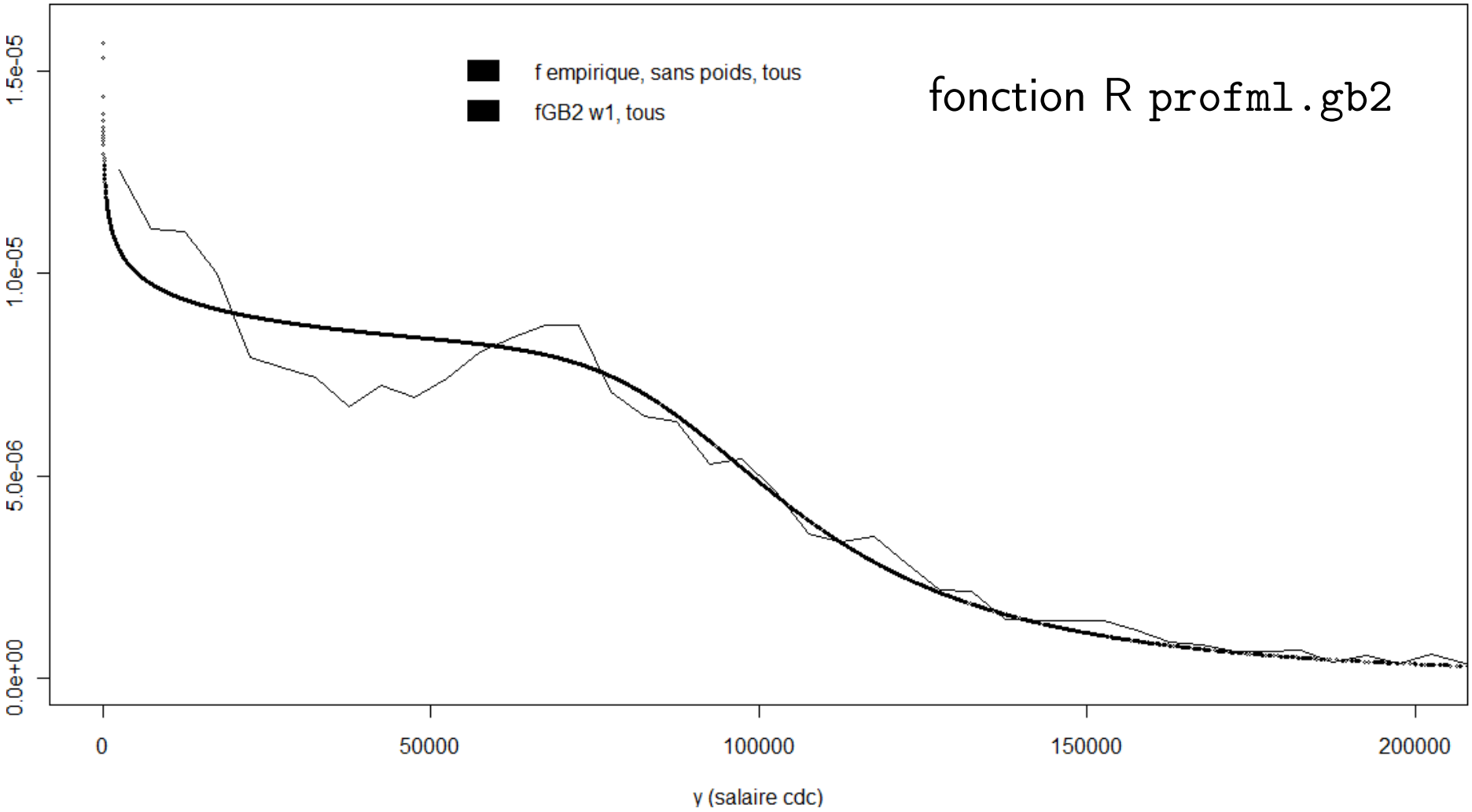
- $\rightarrow$ identifier les variables auxiliaires $\mathbf{X}$ et instrumentales $\mathbf{Z}$
- Les $\mathbf{X}$ doivent expliquer y_{cdc} et être disponibles pour les répondants et les non répondants, donc sur s_1.
  $\rightarrow$ choix raisonné de variables corrélant à $> 10\%$ avec y_{cdc}
- Les $\mathbf{Z}$ 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

▸ voir détails w_{2k}

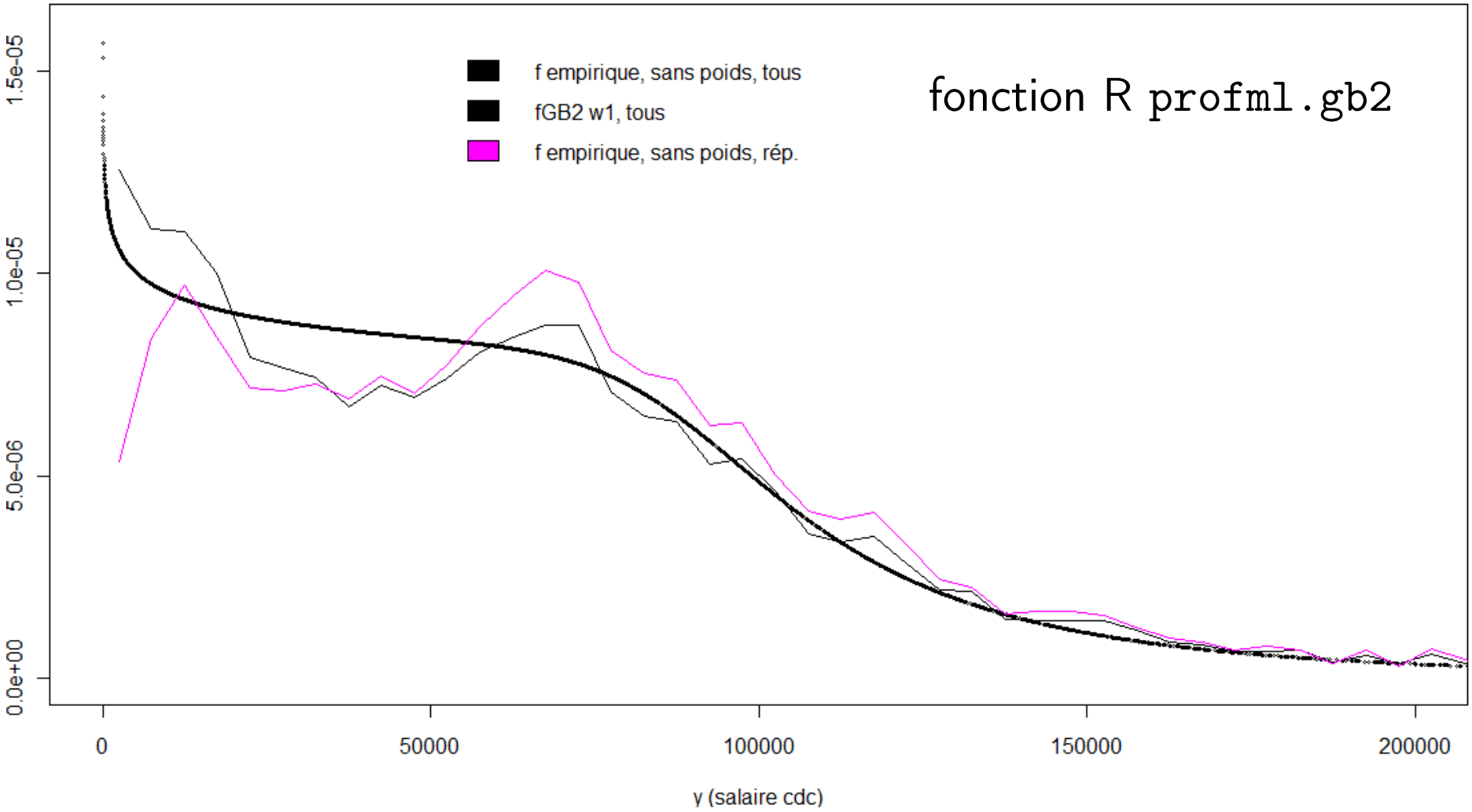
2 Ajustements GB2



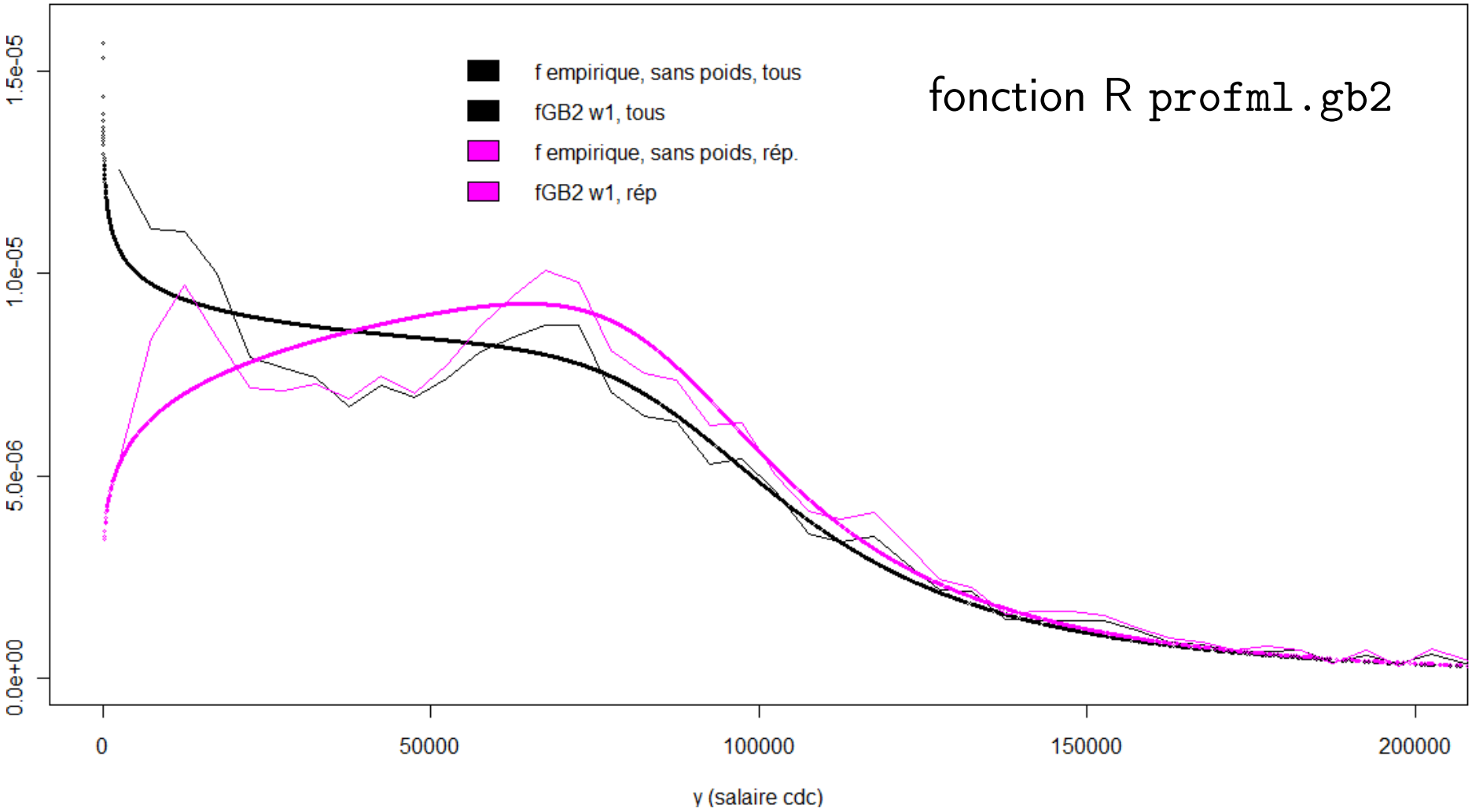
2 Ajustements GB2



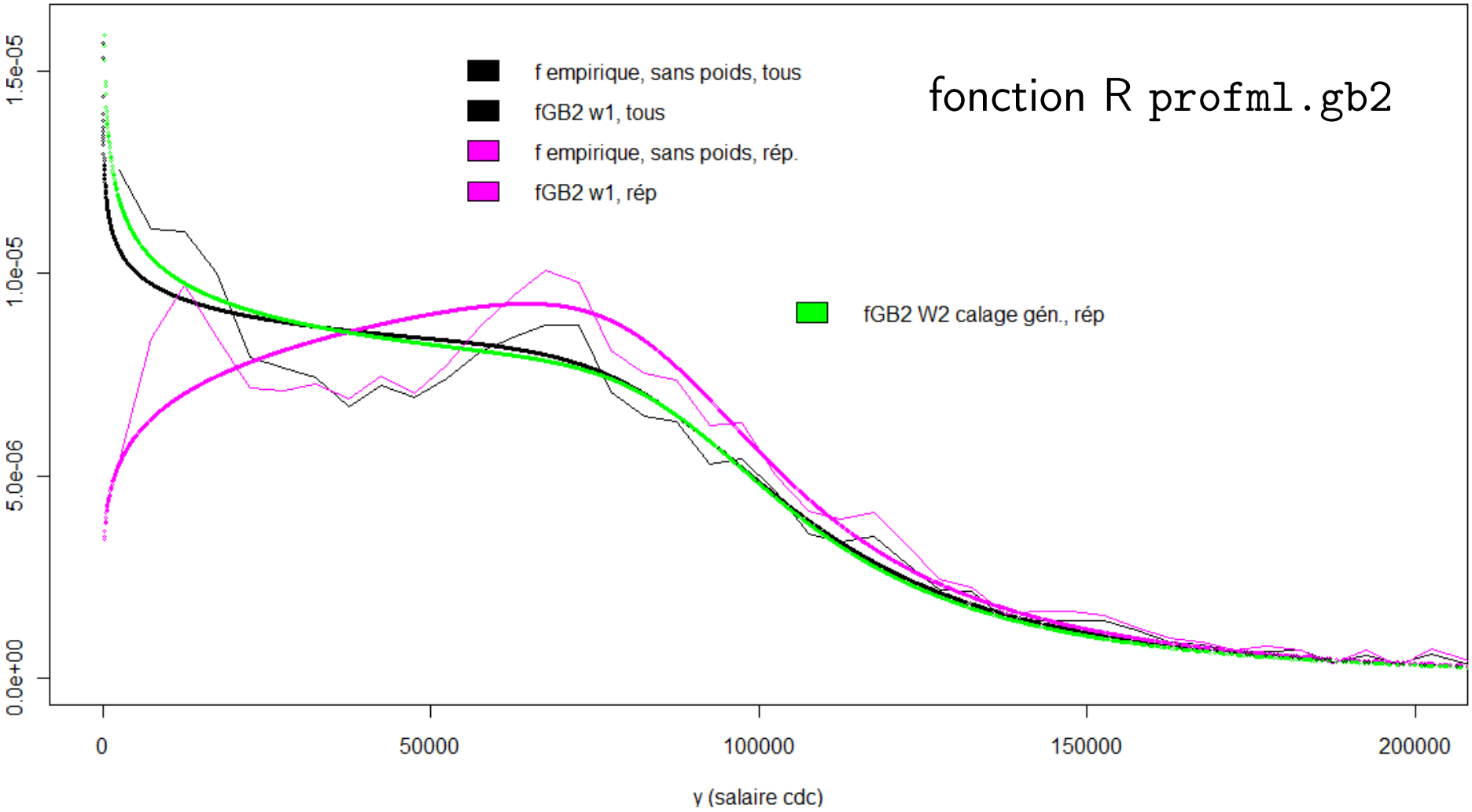
2 Ajustements GB2



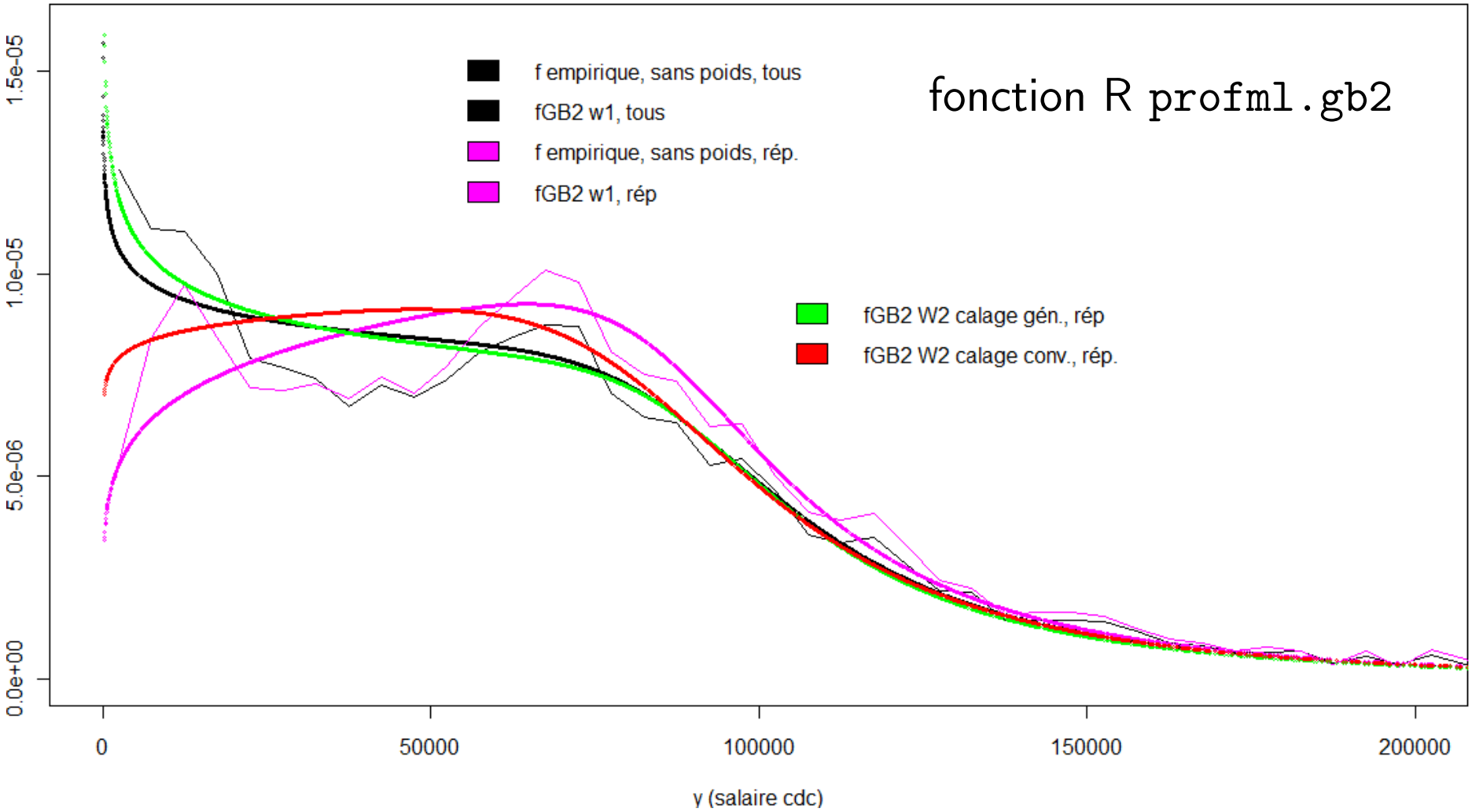
2 Ajustements GB2



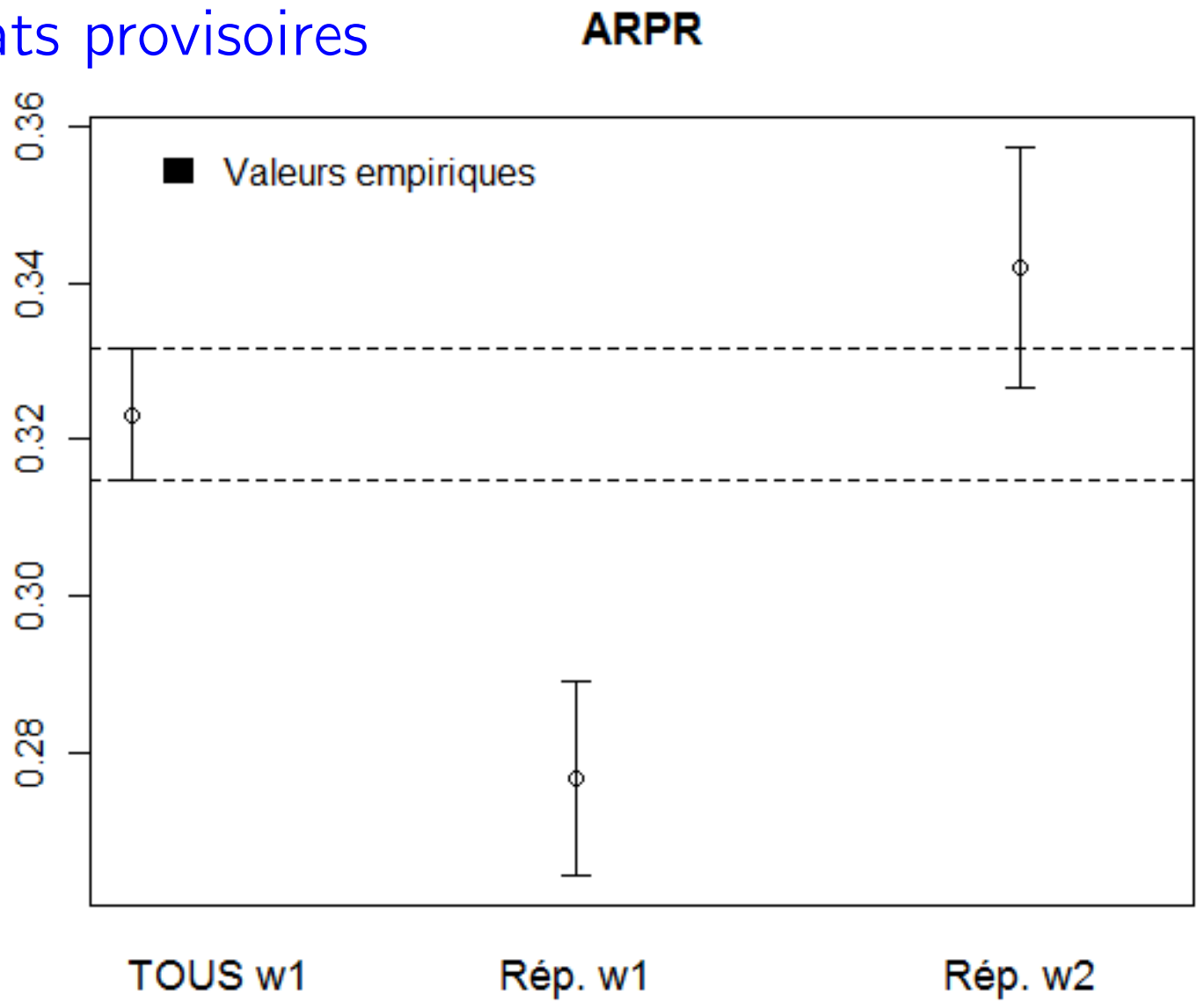
2 Ajustements GB2



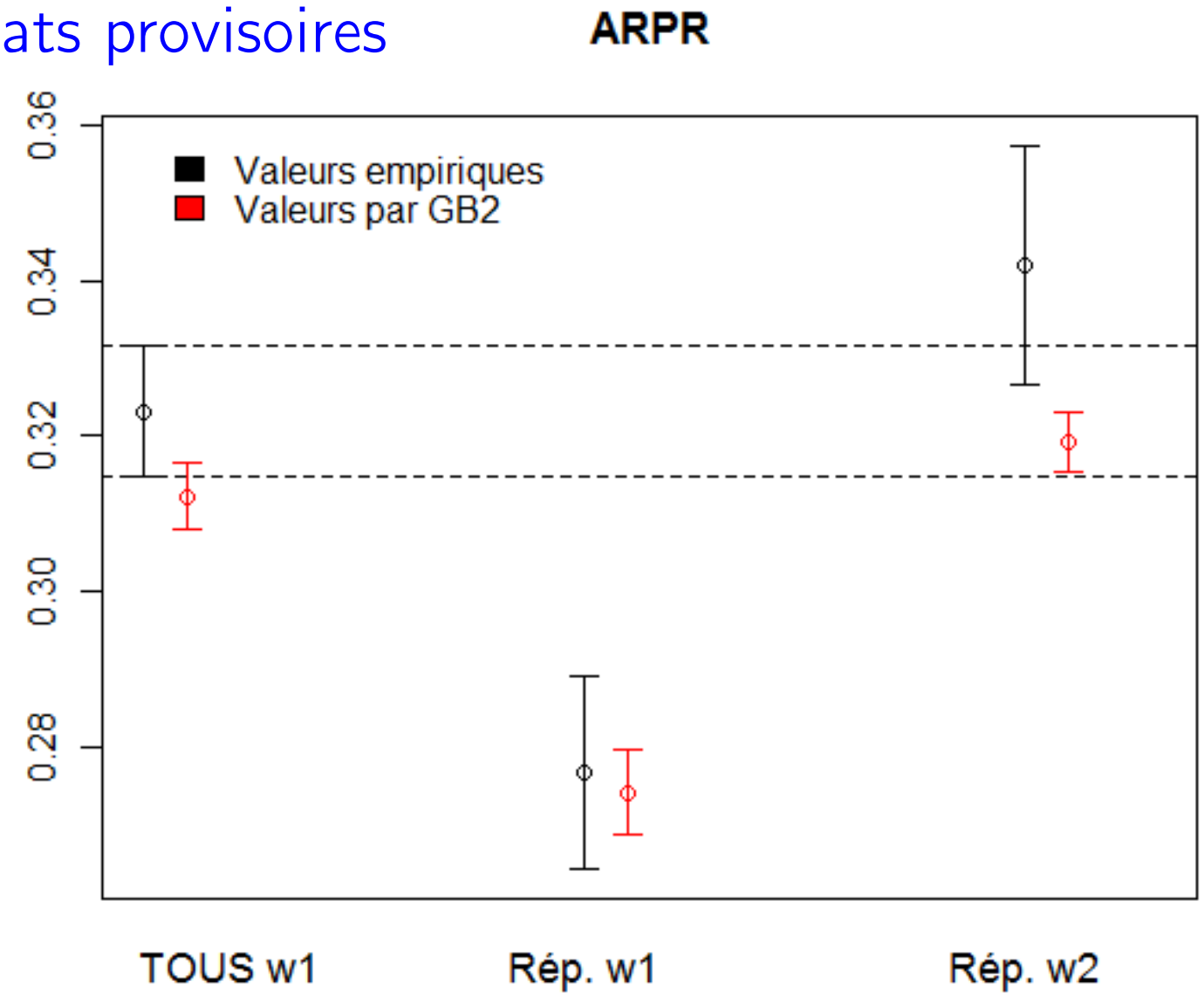
2 Ajustements GB2



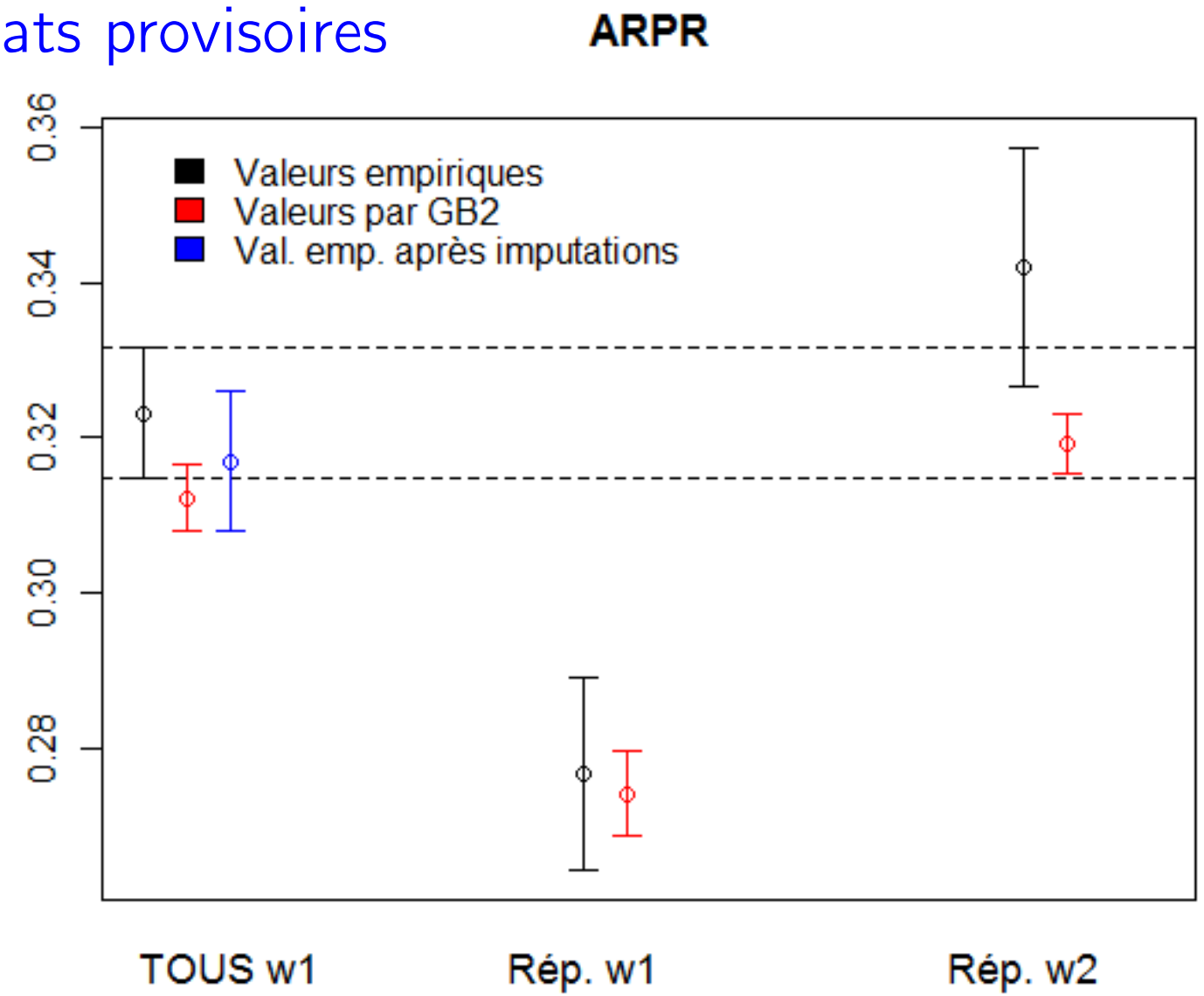
Résultats provisoires



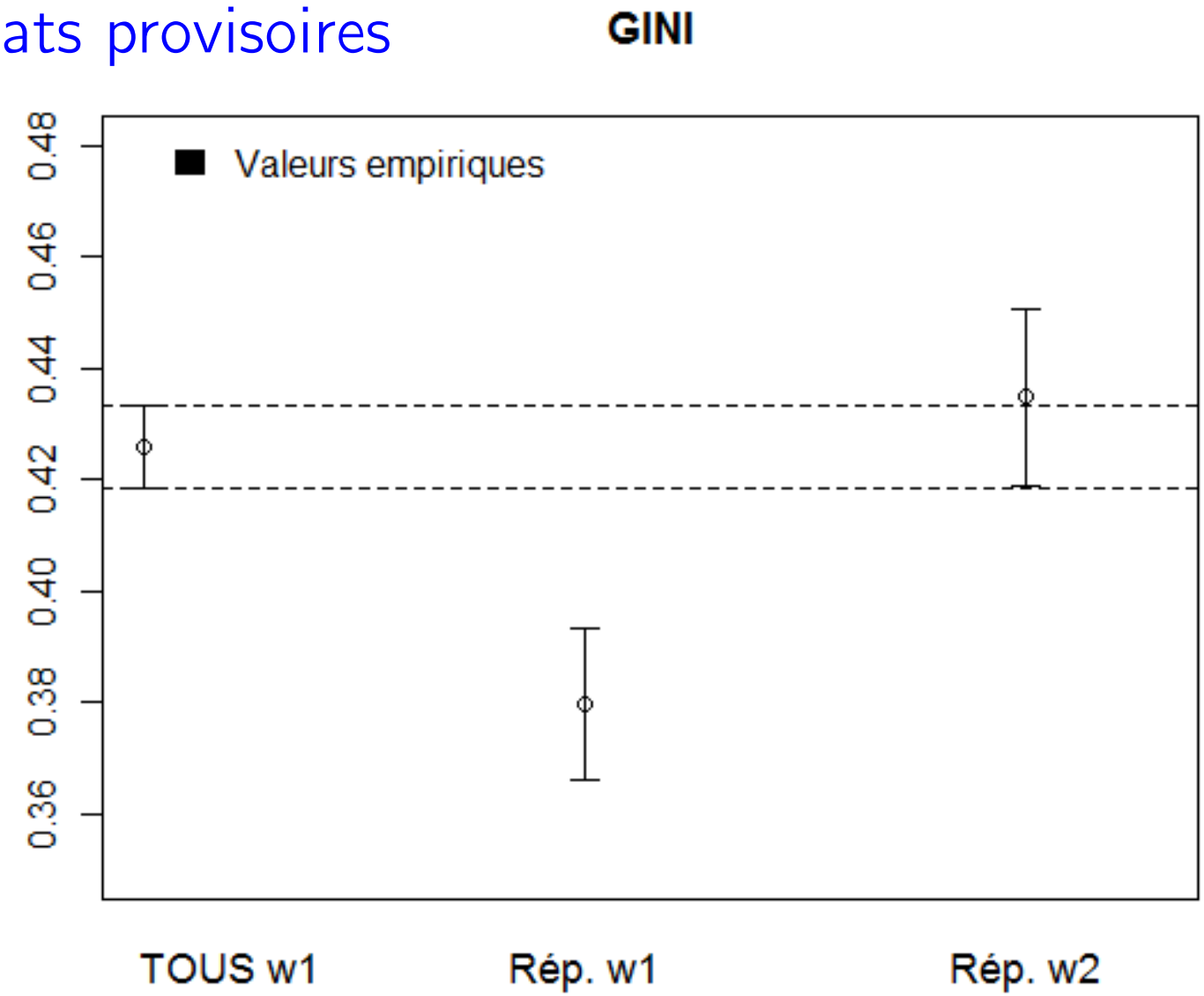
Résultats provisoires



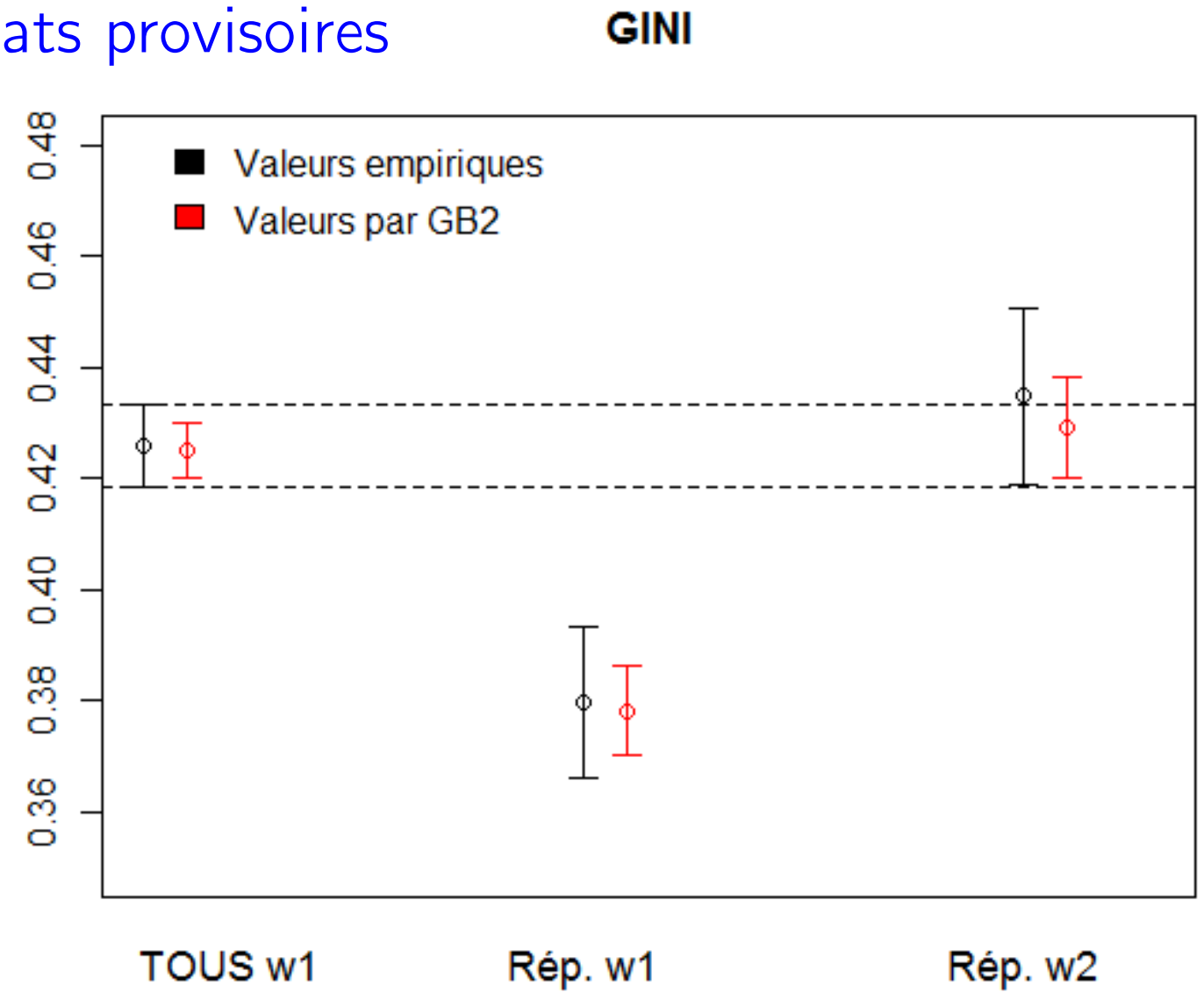
Résultats provisoires



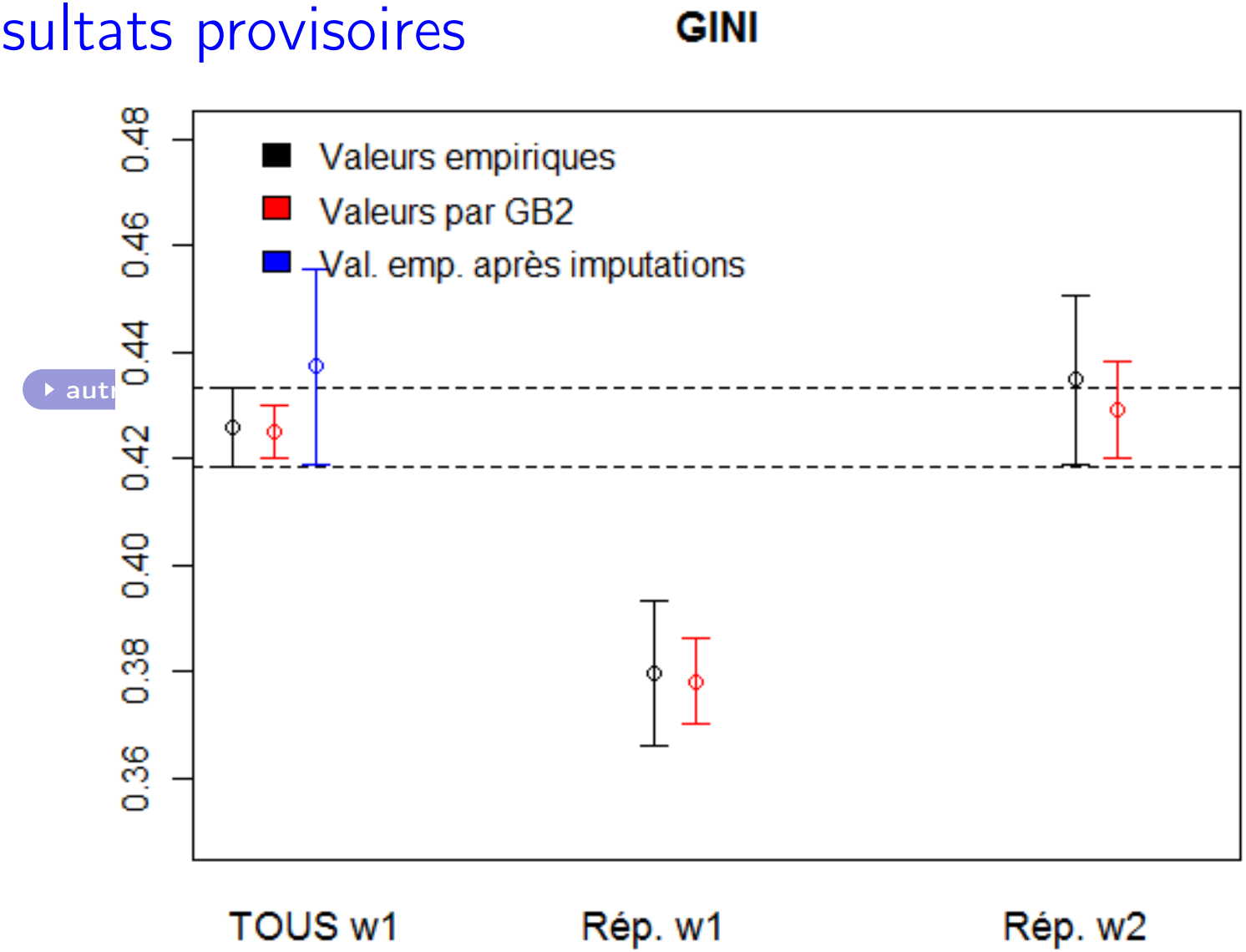
Résultats provisoires



Résultats provisoires



Résultats provisoires



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

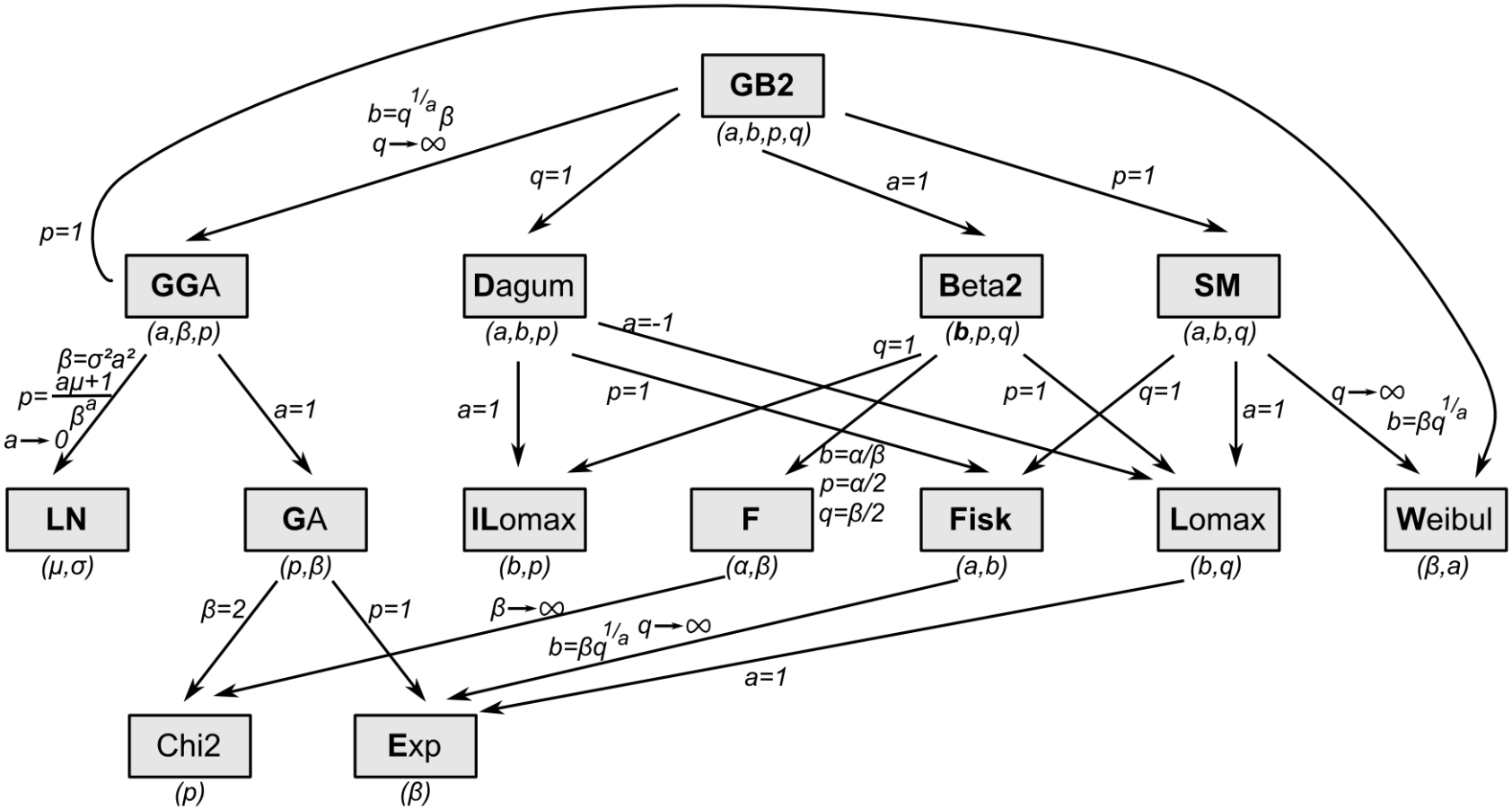
Références citées

- 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*.

ANNEXES

Loi GB2 V

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



Loi GB2 et indices d'inégalité I

Seuil de risque de pauvreté (ARPT)

Soit m la médiane de la $GB2(a, b, p, q)$, $F_{GB2}(m) = 0.5$. Alors l'ARPT est donné par :

$$ARPT(a, b, p, q) = 0.6\, m \tag{1}$$

Taux de risque de pauvreté (ARPR)

Le risque de pauvreté étant indépendant de l'échelle, le paramètre b peut-être choisi arbitrairement, par exemple fixé à 1.

$$\begin{aligned} ARPR(a, p, q) &= \mathbf{P}(Y < 0.6m) = \mathbf{P}(Y < ARPT) \\ &= F_{GB2}(ARPT; a, 1, p, q) \end{aligned} \tag{2}$$

Loi GB2 et indices d'inégalité II

Relative median at-risk-of-poverty gap (RMPG)

C'est la différence relative entre la seuil de pauvreté et la médiane des pauvres (=ceux en-dessous du seuil) :
Si $m_p = F_{GB2}^{-1}(ARPR/2) = q_{GB2}(ARPR/2)$ est la médiane des pauvres,

$$RMPG(A, a, p, q) = \frac{0.6m - m_p}{0.6m} = \frac{ARPT - m_p}{ARPT} \tag{3}$$

Le RMPG est défini comme un moins le ratio entre le revenu médian des personnes au-dessous de l'ARPT et l'ARPT.

$$RMPG(A, a, p, q) = 1 - \frac{q_{GB2}(ARPR/2, a, 1, p, q)}{ARPT} \tag{4}$$

où q_{GB2} est la fonction quantile de la GB2 considérée.

Loi GB2 et indices d'inégalité III

Quintile share ratio (QSR ou S_{80}/S_{20})

Soient q_{80} et q_{20} les 80^e et 20^e percentiles de la fonction de répartition de la GB2. Le quintile share ratio est le ratio de la somme des revenus des 20% les plus riches, sur la somme des revenus des 20% les plus pauvres :

$$QSR = \frac{\mathbf{E}(Y|Y > q_{80})}{\mathbf{E}(Y|Y < q_{20})}. \tag{5}$$

Il peut s'exprimer à l'aide des moments incomplets d'ordre 1 :

$$QSR(a,p,q) = 1 - \frac{F_{GB2(1)}(q_{80};a,b,p,q)}{F_{GB2(1)}(q_{20};a,b,p,q)} \tag{6}$$

Loi GB2 et indices d'inégalité IV

Si X et Y sont deux variables aléatoires de distribution F,

$$GINI(F) = \frac{\mathbf{E}(|X-Y|)}{2\mathbf{E}(X)} = \frac{\mathbf{E}(|X-Y|)}{2m}. \quad (7)$$

Indice de Gini est un indice d'inégalité mesurant l'espérance de la différence absolue de deux revenus sélectionnés indépendamment par rapport au revenu médian. Celui d'une distribution GB2 est donné par (McDonald, 1984) :

$$GINI(a,p,q) = \frac{B(2p+1/a, 2q-1/a)}{B(p,q)B(p+1/a, q-1/a)} \left\{ \frac{1}{p} G_1 - \frac{1}{p+1/a} G_2 \right\} \quad (8)$$

où

$$G_1 = {}_3F_2 \left[\begin{matrix} 1, & p+q, & 2p+1/a \\ & p+1, & 2(p+q) \end{matrix} ;1 \right]$$

et

$$G_2 = {}_3F_2 \left[\begin{matrix} 1, & p+q, & 2p+1/a \\ & p+1+1/a, & 2(p+q) \end{matrix} ;1 \right],$$

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

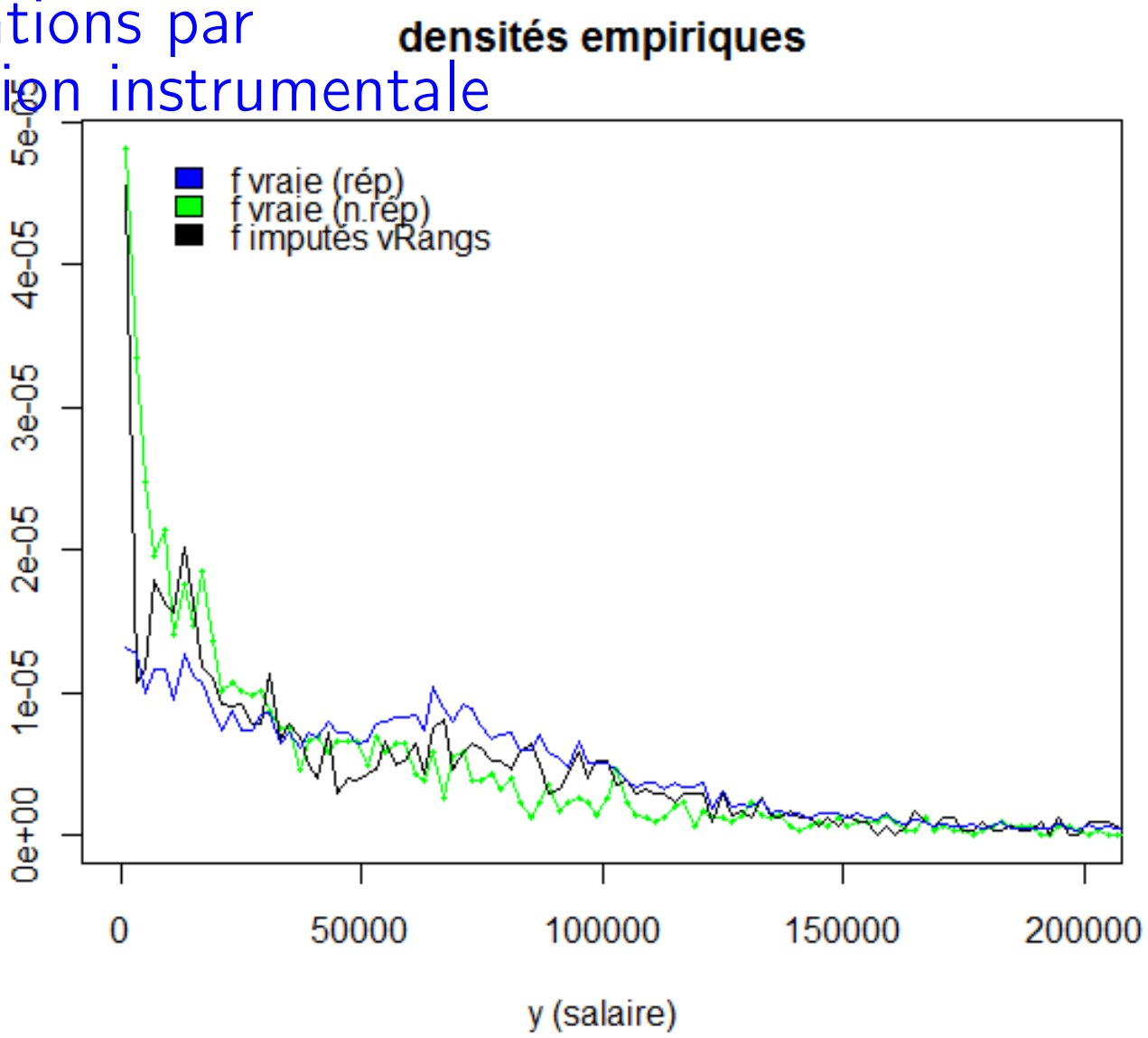
Calage généralisé

Comparaison des poids avant calage w_{1k}, à ceux obtenus par ajustement logisitique, raking ratio et exponentiel généralisé du poids.

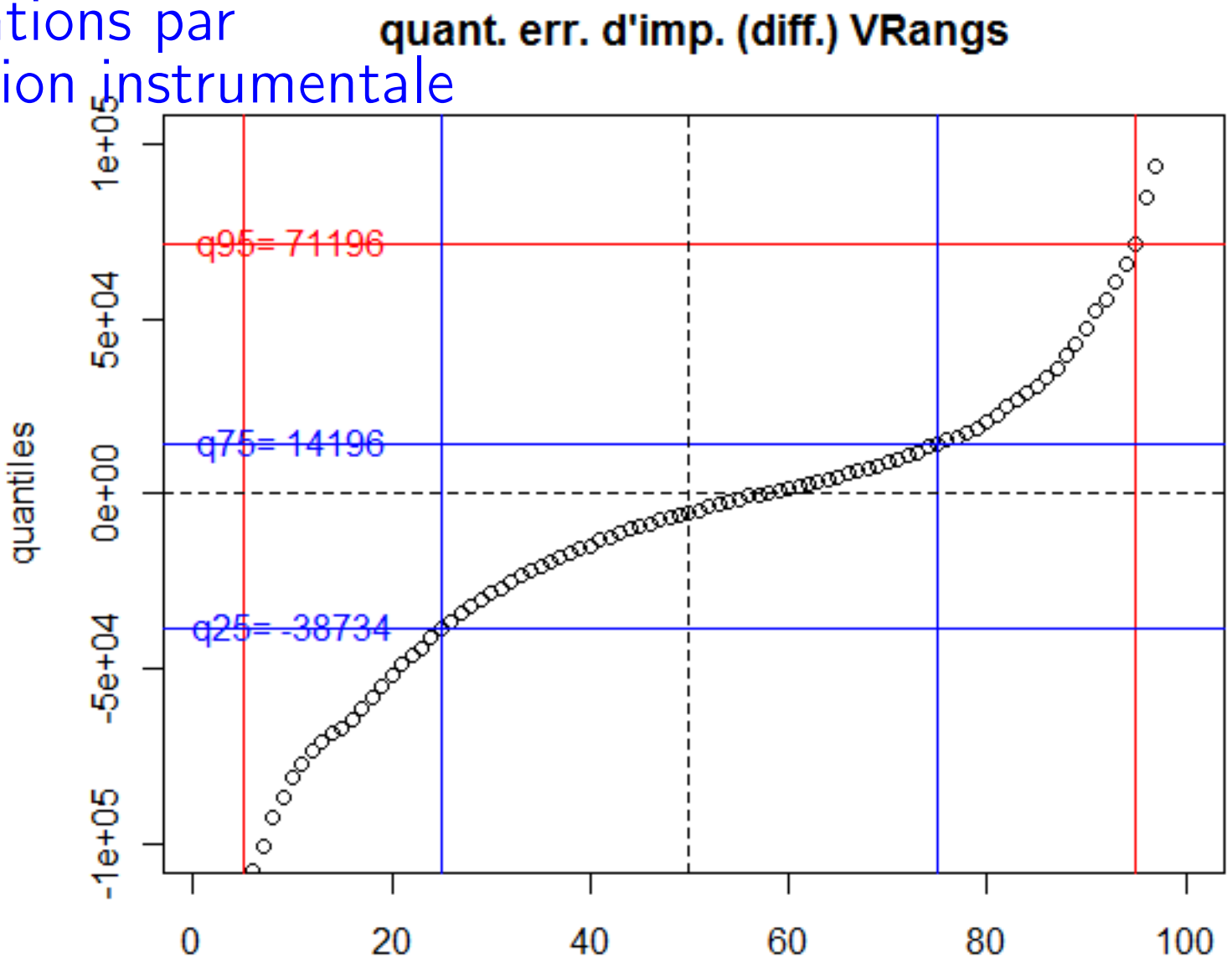
	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
w_{1k}	85.6	302.2	372.1	428.1	499.8	4583.0
$F^{logit}(\mathbf{z}_k'\boldsymbol{\lambda})$	1.00	1.02	1.07	1.31	1.19	11.25
w_{2k}^{logit}	85.8	340.7	439.4	550.1	616.3	5901.0
$F^{rak}(\mathbf{z}_k'\boldsymbol{\lambda})$	0.07	0.40	1.14	1.30	1.44	7.29
w_{2k}^{rak}	24.6	333.8	459.0	549.1	651.9	5461.0
$F^{trun}(\mathbf{z}_k'\boldsymbol{\lambda})$	0.10	0.93	1.20	1.30	1.56	4.24
w_{2k}^{trun}	25.6	327.4	472.5	548.8	671.9	5655.0

retour

Imputations par régression instrumentale



Imputations par régression instrumentale



Mécanismes de non réponse

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

- 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 : $P(\mathcal{R}|y) = P(\mathcal{R}|y_{obs}, y_{mis})$.

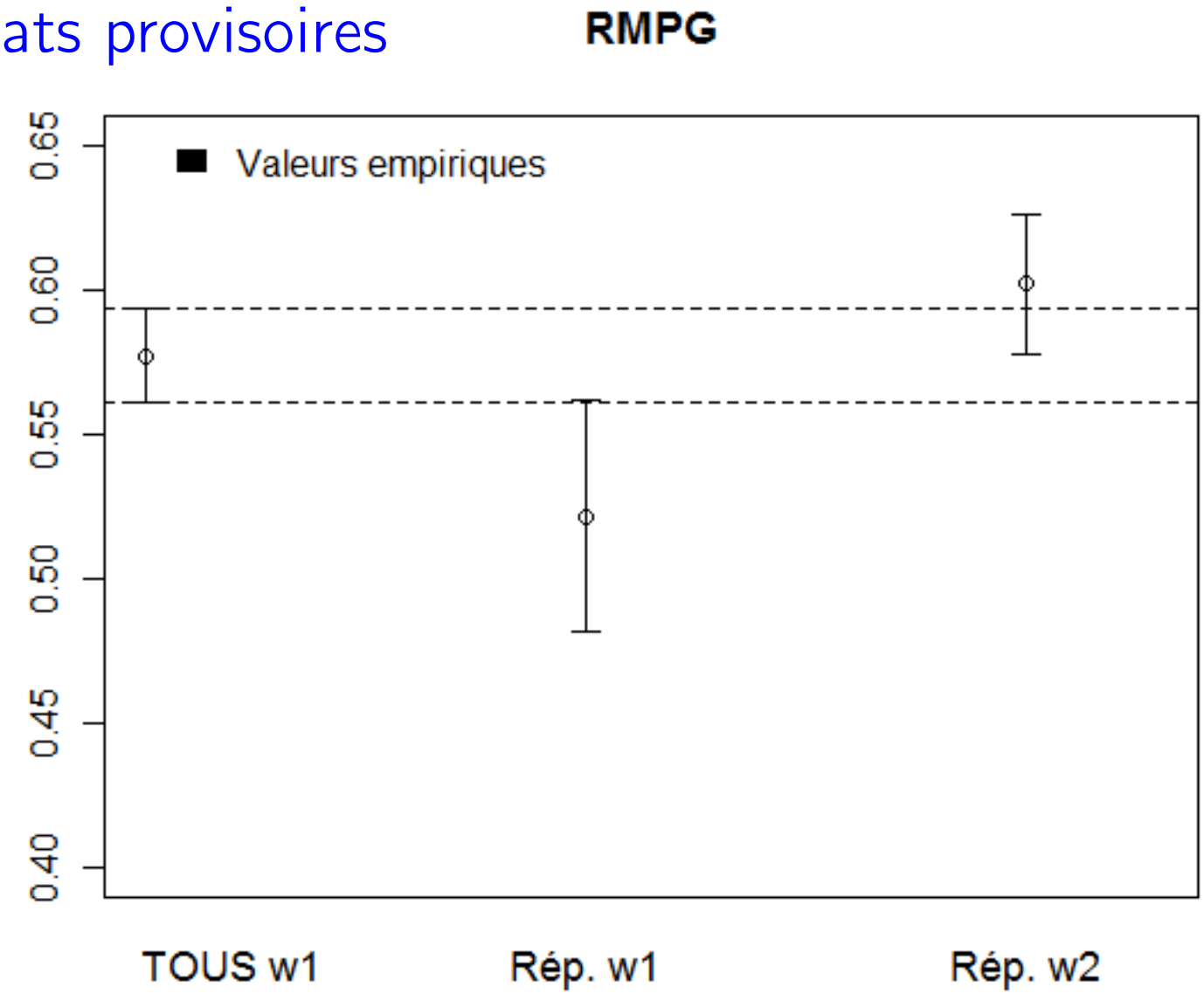
Extrait de l'arbre de segmentation.

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.

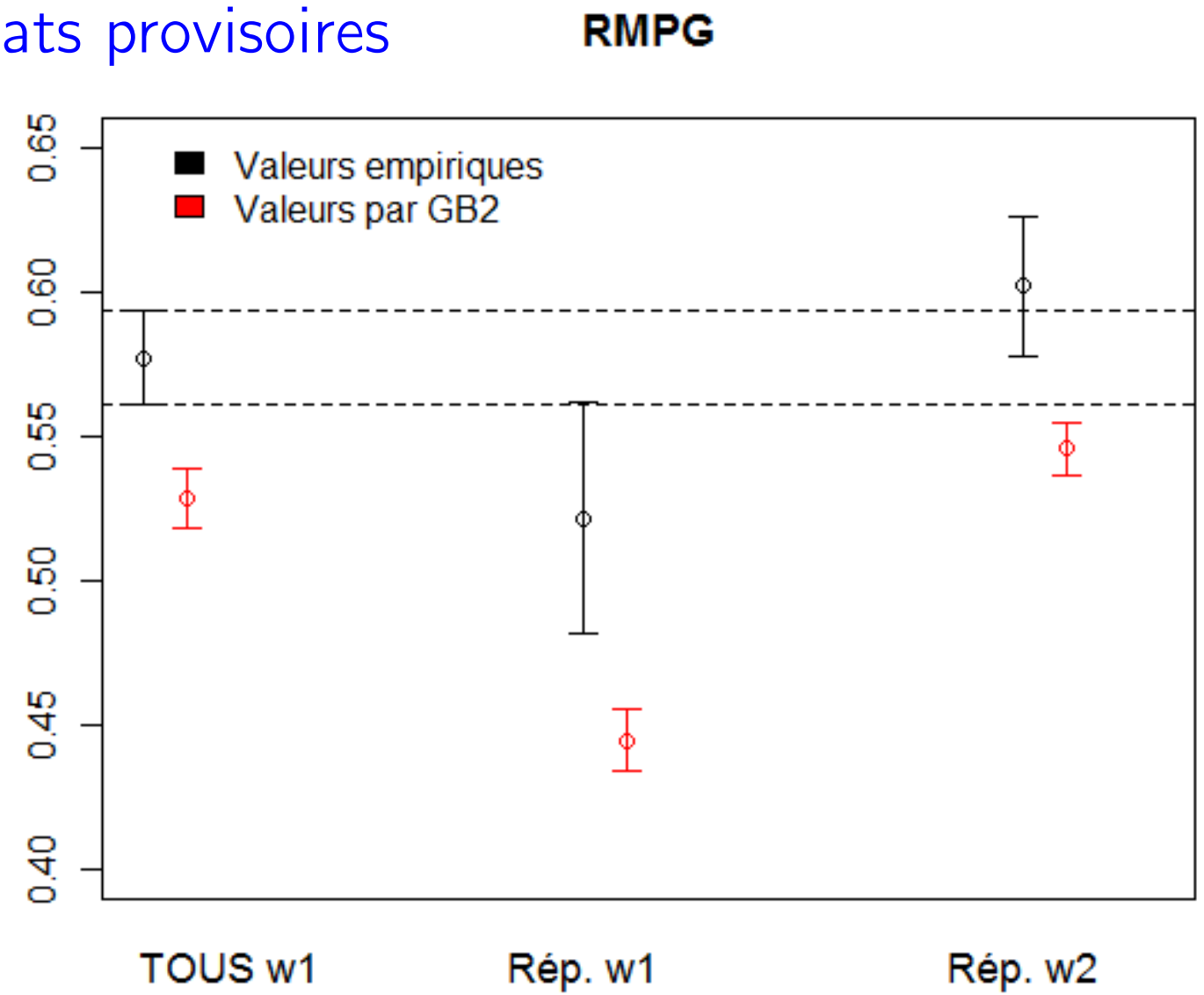
	$med_{y_{cdc}}^{GHR}$	$medNR_{y_{cdc}}^{GHR}$	tx_{rep}
$med_{y_{cdc}}^{GHR}$	1	0.77	0.39
$medNR_{y_{cdc}}^{GHR}$	0.77	1	0.32
tx_{rep}	0.39	0.32	1

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. retour ...

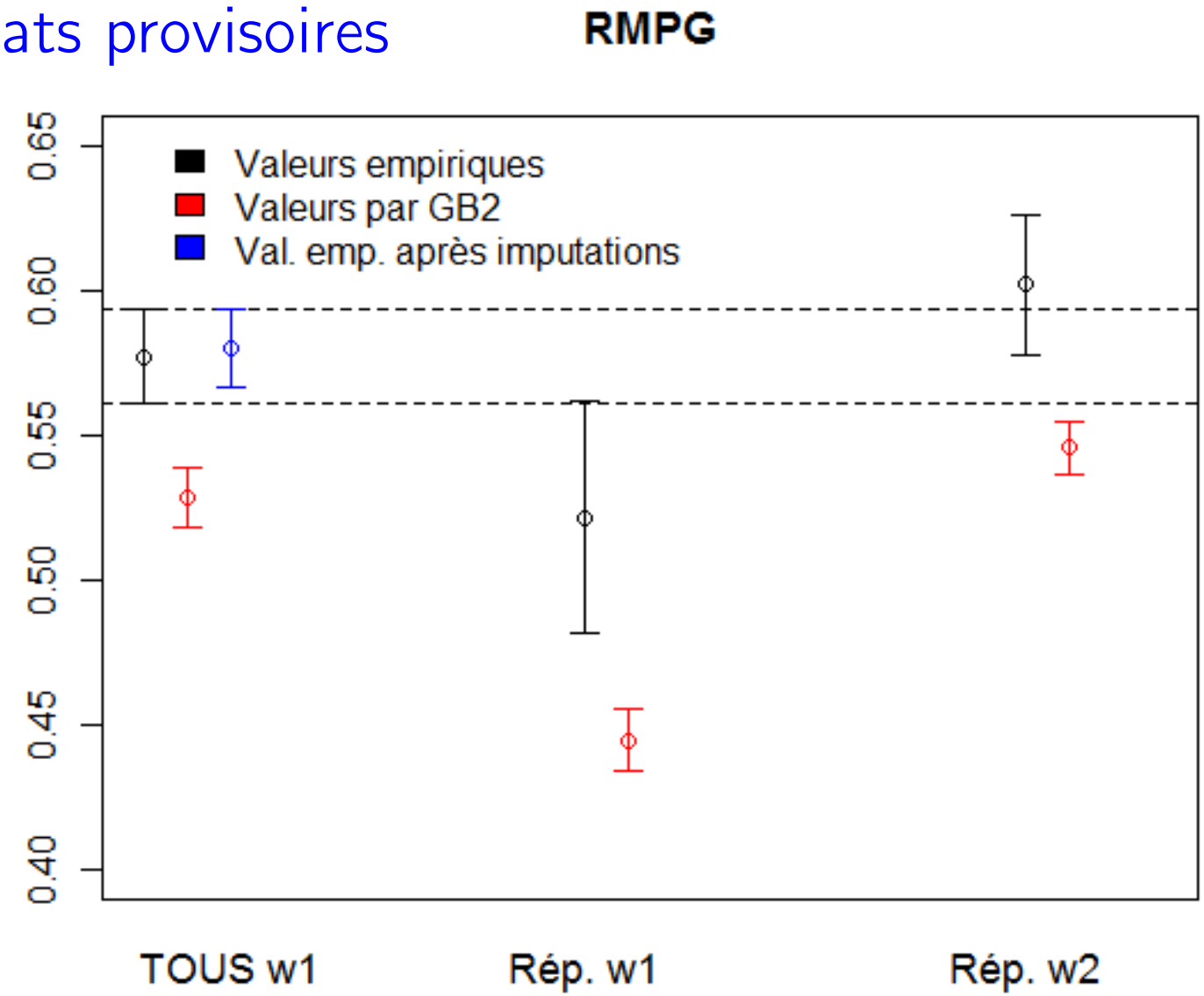
Résultats provisoires



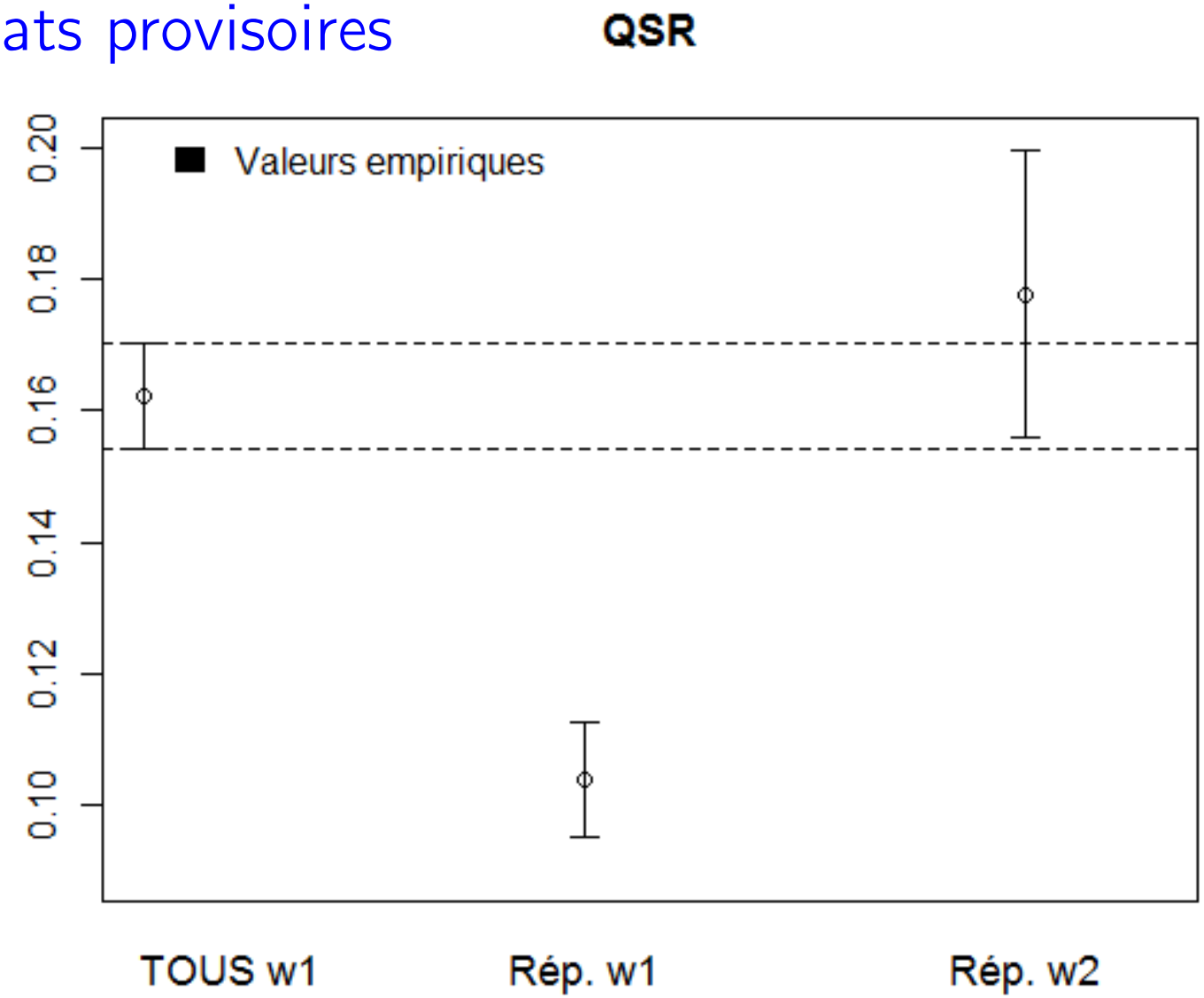
Résultats provisoires



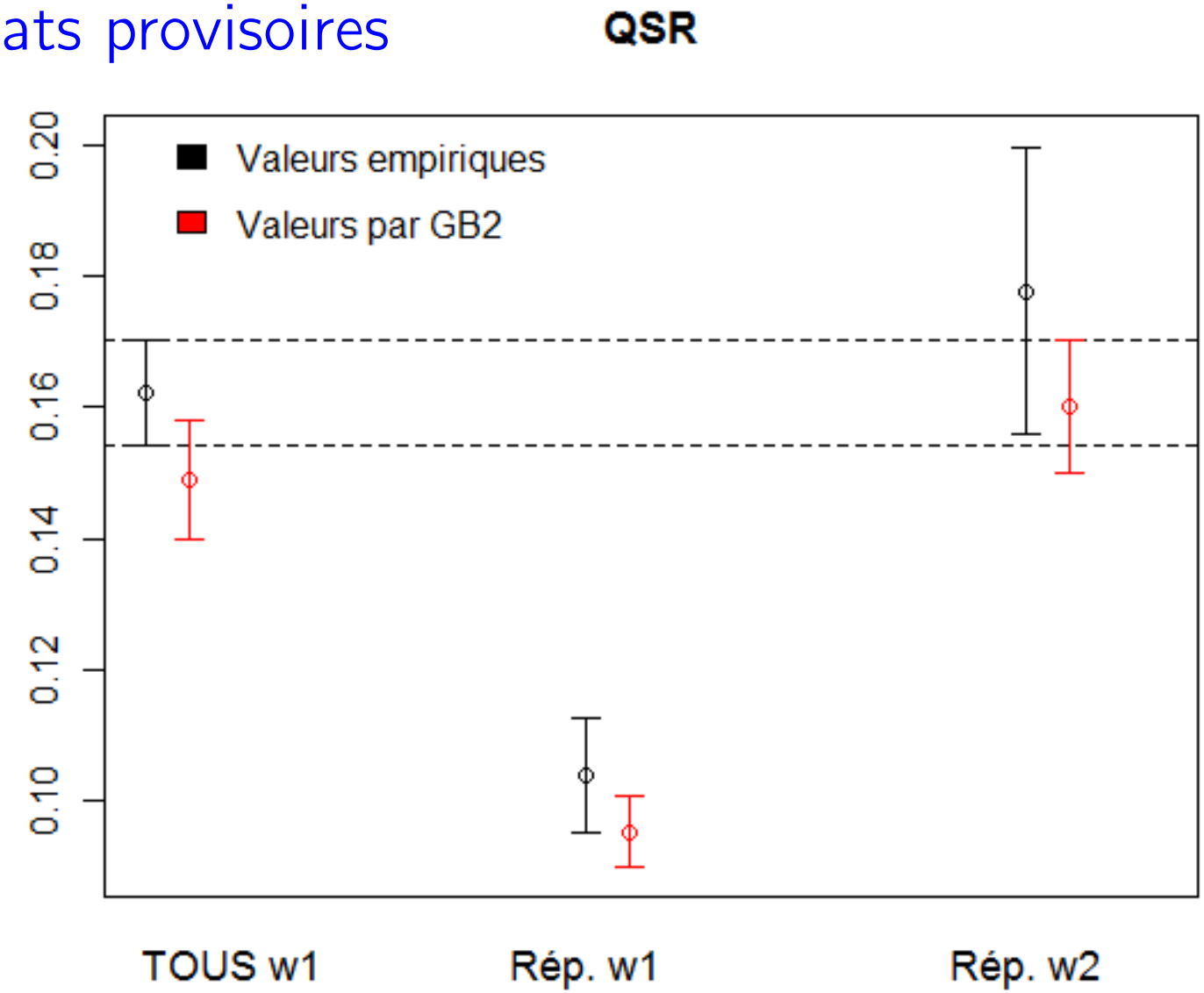
Résultats provisoires



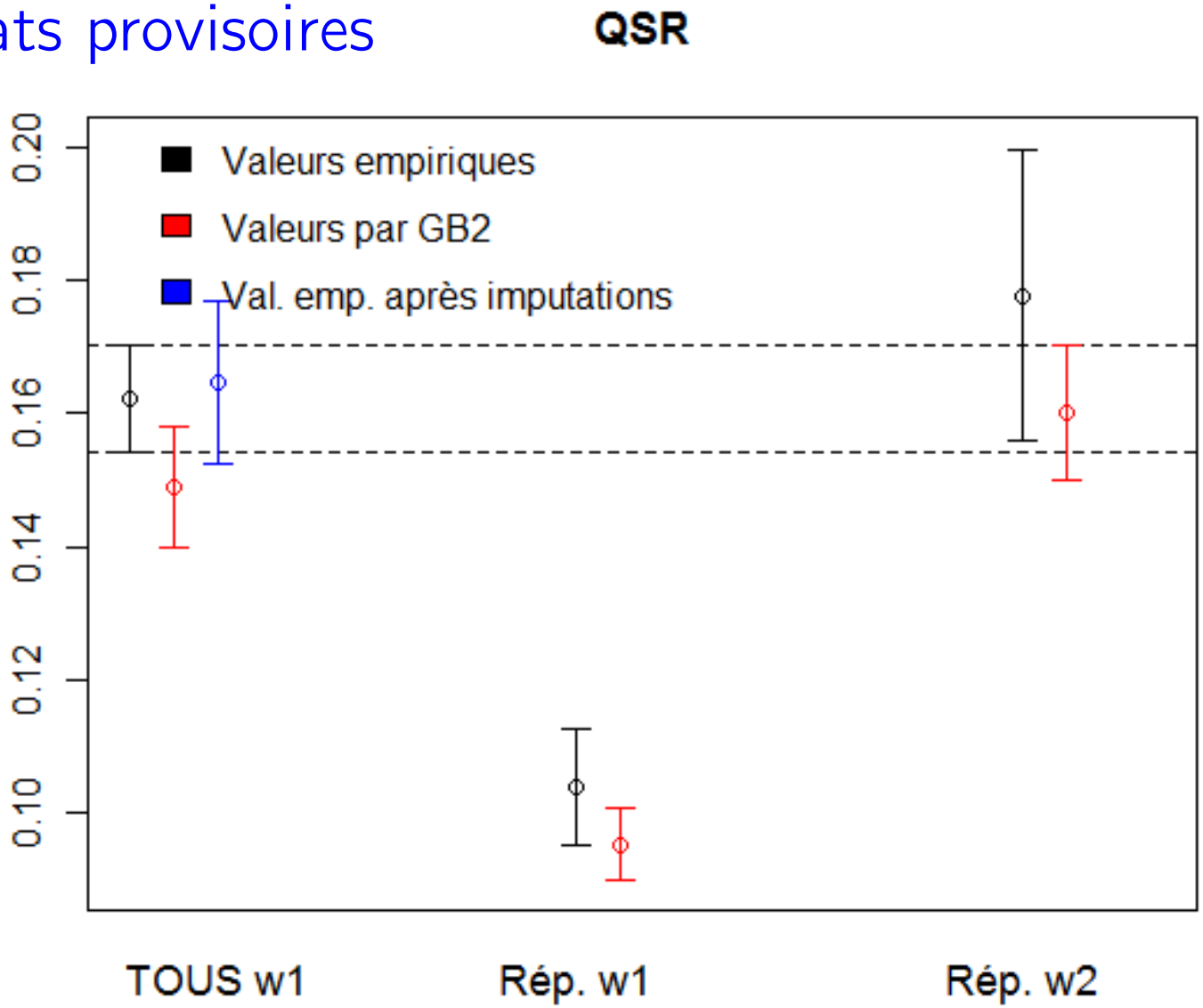
Résultats provisoires



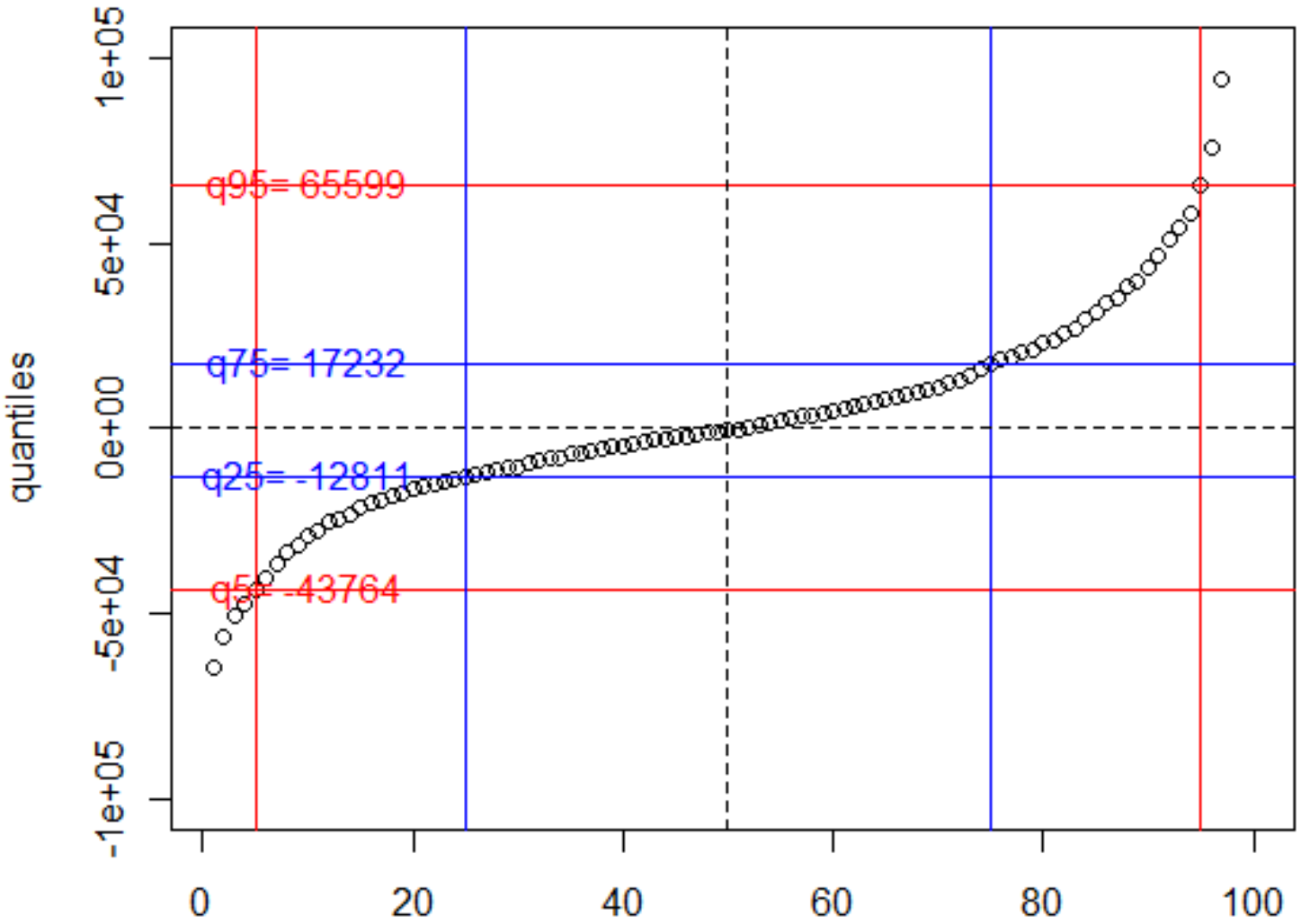
Résultats provisoires



Résultats provisoires



Résultats provisoires



Résultats provisoires densités empiriques