Calage sur bornes minimales et choix des bornes de calage

Emmanuel Gros [1] & Antoine Rebecq [2]

[1] *Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE* – emmanuel.gros@insee.fr
[2] *Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE* – antoine.rebecq@insee.fr

Résumé. Le calage sur marges est largement utilisé en sondages afin de réduire la variance des estimations. Afin de se prémunir contre le risque d'augmentation du biais et de la variance pour certains sous-domaines de taille réduite, les méthodes « bornées » logit ou linéaire tronquée sont en général privilégiées, car elles permettent de limiter l'ampleur des ajustements de poids induits par le calage. On peut imaginer chercher les bornes conduisant à une étendue minimale des facteurs de calage pour ces méthodes bornées : on parle de « calage sur bornes minimales ». Dans ce papier, on propose un algorithme linéaire de résolution du programme de calage sur bornes minimales de taille $\mathcal{O}(n^2)$.

On effectue ensuite une étude par simulations afin de quantifier l'impact du choix des bornes de calage sur la qualité des estimateurs calés. On compare en particulier deux stratégies de choix des bornes de calage :

- le calage sur bornes minimales, qui limite au maximum l'étendue des facteurs de calage mais conduit en général à de fortes accumulations aux bornes ;
- le calage sur des bornes plus lâches, permettant d'assurer un contrôle suffisant des déformations maximales de poids induites par le calage tout en évitant les phénomènes d'accumulation aux bornes.

Il s'avère que le calage sur bornes minimales conduit à des estimateurs calés moins efficaces, et ce d'autant plus que le phénomène d'accumulation de rapports de poids aux bornes est prononcé.

Mots-clés. Calage sur marges, méthodes de calage bornées, calage sur bornes minimales, choix des bornes de calage.

Introduction

Le calage sur marges est l'une des méthodes de redressement les plus populaires et les plus utilisées en statistique d'enquête. Il consiste à remplacer les pondérations initiales des unités de l'échantillon par des pondérations calées aussi proches que possible des pondérations initiales au sens d'une certaine distance, et permettant d'assurer la cohérence entre les estimations issues de l'enquête et les totaux connus sur la population entière pour un certain nombre de variables auxiliaires. Le calage sur marges conduit en outre à un

estimateur plus efficace que l'estimateur non calé pour les variables d'intérêt bien corrélées aux variables auxiliaires.

Le choix de la fonction de distance à utiliser constitue une question classique que se pose tout utilisateur de la procédure de calage sur marges. Dans leur article fondateur [3], Deville et Särndal soulignent que, toutes les fonctions de distance conduisant à des estimateurs calés asymptotiquement équivalents à l'estimateur par la régression, le choix de la fonction de distance n'a au final qu'un impact très marginal sur les propriétés statistiques de l'estimateur calé pour des estimations portant sur des domaines de taille moyenne ou grande. Ils notent cependant que les distances classiques du khi-deux et de l'entropie relative – correspondant aux méthodes linéaire et raking ratio – présentent un certain nombre d'inconvénients :

- elles peuvent conduire à des poids calés négatifs (pour la méthode linéaire) ou positifs mais inférieurs à 1 (pour les méthodes linéaire et exponentielles), potentiellement problématiques lors de la diffusion du fichier de données individuelles contenant les poids calés ;
- elles peuvent également induire des déformations de poids très importantes à la hausse et donc engendrer l'apparition de poids extrêmement élevés. Or de tels poids sont susceptibles de créer des unités influentes et donc de conduire à des estimations peu robustes sur certaines sous-populations de taille réduite.

C'est afin de pallier ces problèmes que Deville et Särdnal introduisent deux fonctions de distance – correspondant aux méthodes logit et linéaire tronquée – conduisant à des fonctions de calage bornées, permettant de limiter l'étendue des corrections de poids induites par le calage. Depuis, un consensus s'est établi quant au choix de la méthode de calage, qui conduit en général à privilégier l'utilisation d'une méthode bornée – logit ou linéaire tronquée, sans préférence pour l'une ou l'autre de ces méthodes – par rapport aux méthodes non bornées – linéaire ou raking ratio.

Si un consensus s'est fait jour quant au choix de la méthode de calage, la question des bornes de calage L et U à retenir lorsque l'on procède à un calage via une méthode bornée n'a en revanche été que très peu abordée à notre connaissance dans la littérature, et ne semble en tout cas pas tranchée :

- certains praticiens (cf. [4] par exemple) conseillent de limiter au maximum l'étendue des corrections de poids, afin de se prémunir autant que faire se peut contre de potentiels problèmes lors d'estimations portant sur des domaines de taille réduite ;
- pour d'autres (cf. [8] par exemple) en revanche, les bornes L et U doivent certes être choisies de manière à éviter les déformations excessives de poids induites par le calage mais sans pour autant être trop « strictes »[1]. En effet, le choix de bornes L et U très « serrées » autour de 1 conduit en général à de fortes accumulations

1. Il s'agit de l'approche retenue à l'Insee.

de rapports de poids à ces bornes, qui nous éloignent de la philosophie initiale du calage sur marges consistant à modifier « le moins possible » les poids initiaux tout en respectant les équations de calage. En outre, dans une telle configuration, on peut imaginer que des estimations portant sur certains domaines particuliers – pour lesquels tous les poids initiaux auraient été ajustés à la hausse d'un facteur U par le calage par exemple – s'avèrent problématiques.

Il semble impossible de trancher entre ces deux approches d'un point de vue théorique, puisqu'on ne dispose que de propriétés asymptotiques pour les estimateurs calés et qu'on s'intéresse justement ici à des estimations portant sur des domaines de taille réduite.

Dans cet article, nous nous intéressons dans un premier temps à la première approche de « calage sur bornes minimales ». Cette approche, déjà explorée par le passé (cf. [11] et [4] par exemple), consiste à chercher, via une procédure d'optimisation numérique, les bornes L^* et U^* limitant au maximum l'étendue des facteurs de calage, *i.e.* telle que $U^* - L^*$ soit minimale. On présente ici ce programme ainsi qu'un algorithme optimisé de résolution, implémenté dans le package R Icarus ([7]).

Puis, dans un second temps, nous procédons à une étude par simulations afin d'essayer de quantifier l'impact du choix des bornes de calage sur la qualité des estimateurs calés, pour des estimations portant sur des domaines de taille moyenne ou réduite.

1 Le calage sur bornes minimales

1.1 Notations et programme

On se donne une population $\mathcal{U}$ de taille N et un échantillon s de taille n, sélectionné par un plan de sondage p, de probabilités d'inclusions simples $(\pi_k)_{k\in\mathcal{U}}$. On note $d_k = 1/\pi_k$ les poids de sondages associés à l'estimateur Horvitz-Thompson. On dispose de J variables auxiliaires X_j, dont les totaux sur la population T_{X_j} sont connus par une source externe à l'enquête. Les poids calés w_k issus d'un calage sur ces J variables auxiliaires sont solutions du problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w_k} \sum\limits_{k\in s} d_k G(\dfrac{w_k}{d_k}) \\ \text{s.c. } \sum_{k\in s} w_k \mathbf{x}_k = T_{\mathbf{X}} \end{cases} \tag{1}$$

où G est la fonction de distance associée à la méthode de calage – linéaire, raking ratio, logit ou linéaire tronquée – retenue.

L'objectif est ici, dans le cadre d'une méthode de calage bornée, de déterminer les bornes L^* et U^* assurant une étendue minimale des **facteurs de calage** $g_k = w_k/d_k$. En notant $\widetilde{X}_s$ la matrice dont les coefficients (k, j) sont égaux à $(d_k \cdot X_{jk})_{1\leq k\leq n, 1\leq j\leq J}$, t le vecteur des totaux T_{X_j} et g le vecteur des facteurs de calage, les équations de calage de 1

s'écrivent sous la forme :

$$\widetilde{X}_s' g = t$$

et les bornes L^* et U^* sont alors solutions du programme de minimisation suivant :

$$\begin{cases} \min\limits_{g \in \mathbb{R}^n} \left(\max\limits_{k \in [[1,n]]} g_k - \min\limits_{k \in [[1,n]]} g_k \right) \\ \qquad \text{s. c. } \widetilde{X}_s' g = t \ ; \ g \geq 0 \end{cases} \tag{2}$$

La contrainte $g \geq 0$ est imposée par la méthode du simplexe utilisée pour la résolution, mais elle n'est pas gênante en pratique. En effet, la méthode linéaire est peu mise en œuvre en production car elle peut justement conduire à des poids calés négatifs. On peut donc supposer sans perte d'efficacité que $L^* > 0$.

Notons que la distance de calage n'apparaît pas dans ce programme d'optimisation. Il s'agit dans un premier temps de trouver L^* et U^* telles que $U^* - L^*$ soit minimale tout en respectant les contraintes de calage. Une fois trouvée une telle solution g_{min}, on pose $L^* = \min g$ et $U^* = \max g$ et on effectue le calage avec une méthode bornée prenant ces bornes en paramètres. L'introduction de la distance à cette étape permettra de discriminer entre plusieurs éventuelles solutions du programme de calage sur bornes minimales. Si la solution au programme est unique, g_{min} sera la solution choisie par le calage sur marges.

On peut montrer (via une écriture matricielle non détaillée ici) que le programme 2 peut s'écrire sous la forme d'un programme d'optimisation linéaire de taille $n \times (n-1) \times (n+1)$. On considère également deux autres programmes :

$$\begin{cases} \min\limits_{g \in \mathbb{R}^n, a \in \mathbb{R}} \max\limits_{k \in [[1,n]]} |g_k - a| \\ \text{s. c. } \widetilde{X}_s' g = t \ ; \ g \geq 0 \end{cases} \tag{3}$$

$$\begin{cases} \min\limits_{g \in \mathbb{R}^n} \max\limits_{k \in [[1,n]]} |g_k - 1| \\ \text{s. c. } \widetilde{X}_s' g = t \ ; \ g \geq 0 \end{cases} \tag{4}$$

Là encore, ces deux programmes peuvent s'écrire sous la forme de programmes d'optimisation linéaires, de tailles respectives $2n \times (n+2)$ et $2n \times (n+1)$.

1.2 Résolution du programme

On note respectivement S_2, S_3 et S_4 les ensembles solution des programmes 2, 3 et 4. Le théorème 1 caractérise les solutions de ces programmes :

Théorème 1. *On suppose que la contrainte de calage $\widetilde{X}_s' g = t$ possède au moins une solution. Alors on a :*

— $S_2 = S_3$
— $S_4 \subset S_2$

Tous ces programmes linéaires sont solubles par la méthode du simplexe de Dantzig ([2]). Résoudre le programme sur bornes minimales implique donc de résoudre par cette méthode le programme 3, de dimension inférieure au programme 2 (utilisation de la mémoire en $\mathcal{O}(n^2)$ contre $\mathcal{O}(n^3)$).

Le package R Icarus, développé dans l'optique de proposer des fonctions de calage adaptées au cadre de la production statistique (en reprenant largement l'interface de la macro SAS Calmar, [9]), contient une implémentation du calage sur bornes minimales. Ce package s'appuie sur l'implémentation du simplexe proposée par le solveur linéaire *RGlpk* ([10]). Elle possède l'avantage de se reposer sur la gestion efficace des matrices creuses du package *slam* ([6]). En pratique en statistique publique, il arrive fréquemment que les variables de calage utilisées soient catégorielles : la matrice $\widetilde{X}_s$ possède donc généralement de très nombreux coefficients nuls.

Le programme 4 ne donne qu'une solution sous-optimale au problème, mais contrairement aux programmes précédents, il est soluble par dichotomie. Ainsi, pour des problèmes de taille très importante, l'algorithme du simplexe en $\mathcal{O}(n^2)$ peut imposer une charge mémoire très importante et un temps d'exécution relativement long. Bien que la convergence de la méthode par dichotomie soit théoriquement lente, elle est potentiellement plus rapide que le simplexe lorsqu'il s'agit de problèmes de grande taille. Afin de ne pas générer de surcharge mémoire sur des machines qui sont potentiellement des ordinateurs personnels, la résolution par dichotomie du programme 4 est utilisée quand un calage sur bornes serrées est lancé dans Icarus avec une matrice de calage contenant plus de $5 \cdot 10^8$ enregistrements.

2 Impact du choix des bornes de calage sur la qualité des estimateurs calés

2.1 Données et protocole utilisés pour l'étude par simulations

Pour réaliser notre étude par simulations, nous nous sommes appuyés sur les données du dispositif Esane [2] portant sur l'année 2013, en nous restreignant au champ du commerce (section G de la NAF rév. 2) hors exhaustif [3], soit une population $\mathcal{U}$ composée de 580 180 entreprises.

2. Le dispositif Esane (pour Élaboration des Statistiques ANnuelles d'Entreprises) combine l'enquête statistique ESA (pour Enquête Sectorielle Annuelle) et l'exploitation des déclarations fiscales des entreprises. Pour plus de détails sur le sujet, voir [1].

3. L'exhaustif de l'ESA est composé des plus grandes entreprises de chaque secteur, qui sont incluses d'office dans l'échantillon. Les grandes entreprises de chaque secteur sont définies comme celles dont

En effet, le dispositif Esane nous permet de disposer, outre de la base de sondage de l'enquête ESA, d'un grand nombre de variables fiscales disponibles pour l'ensemble des unités du champ. Ces variables fiscales vont donc d'une part pouvoir être utilisées comme variables auxiliaires dans un calage sur marges, et d'autre part permettre l'évaluation de la qualité des différents estimateurs calés via le le calcul d'erreurs quadratiques moyennes.

Afin de quantifier l'impact du choix des bornes de calage sur la qualité des estimateurs calés, nous avons procédé comme suit :

- nous avons sélectionné K=40 000 échantillons de taille 1 000 au sein de $\mathcal{U}$ par sondage aléatoire simple stratifié avec allocation proportionnelle ;
- chaque échantillon a ensuite été calé selon trois scénarios de calage – raking ratio, méthode logit avec bornes L=0,5 et U=2 (qui permet un contrôle correct des déformations maximales de poids tout en induisant *a priori* peu d'accumulation de rapports de poids aux bornes) et méthodes logit avec bornes minimales (qui devrait *a priori* conduire a de fortes accumulations de rapports de poids aux bornes) sur les marges suivantes :
  - structures par secteur d'activité (au niveau 3 caractères de la NAF rév.2), tranches d'effectif (0 salarié / 1 salarié / 2 à 5 salariés / 6 à 10 salariés / plus de 10 salariés) et ZEAT (zone d'études et d'aménagement du territoire) ;
  - totaux de chiffre d'affaires, de valeur ajoutée, d'actif total et de passif total.
- Pour une variable d'intérêt Y donnée, on calcule ensuite les estimateurs calés pour chaque échantillon et chaque scénario de calage :

$$\hat{T}^k_{Y,w^{RR}} = \sum_{i \in s_k} w_i^{RR} y_i, k = 1, ..., K$$

$$\hat{T}^k_{Y,w^{logit\ [0,5-2]}} = \sum_{i \in s_k} w_i^{logit\ [0,5-2]} y_i, k = 1, ..., K$$

$$\hat{T}^k_{Y,w^{logit\ min}} = \sum_{i \in s_k} w_i^{logit\ min} y_i, k = 1, ..., K$$

- enfin, on évalue la qualité des 3 estimateurs calés $\hat{T}_{Y,w^{exp}}$, $\hat{T}_{Y,w^{logit\ [0,5-2]}}$ et $\hat{T}_{Y,w^{logit\ min}}$ à l'aide de la racine carrée de leurs erreurs quadratiques moyennes relatives Monte-Carlo :

$$RRMSE(\hat{T}_{Y,w}) = 100 \frac{\sqrt{K^{-1} \times \sum_{k=1}^{K} (\hat{T}^k_{Y,w} - T_Y)^2}}{T_Y}$$

l'effectif salarié, le chiffre d'affaire ou le total de bilan dépassent certains seuils, variables en fonction des secteurs. Dans les secteurs du commerce, l'exhaustif est composé en première approximation des entreprises de plus de 20 salariés ou présentant un chiffre d'affaires supérieur à 38 M€ou un total de bilan supérieur à 75 M€.

2.2 Résultats

Notons tout d'abord qu'en termes d'impact du calage sur les pondérations, les résultats observés sur les simulations pour chacun des trois scénarios envisagés sont conformes aux résultats attendus :

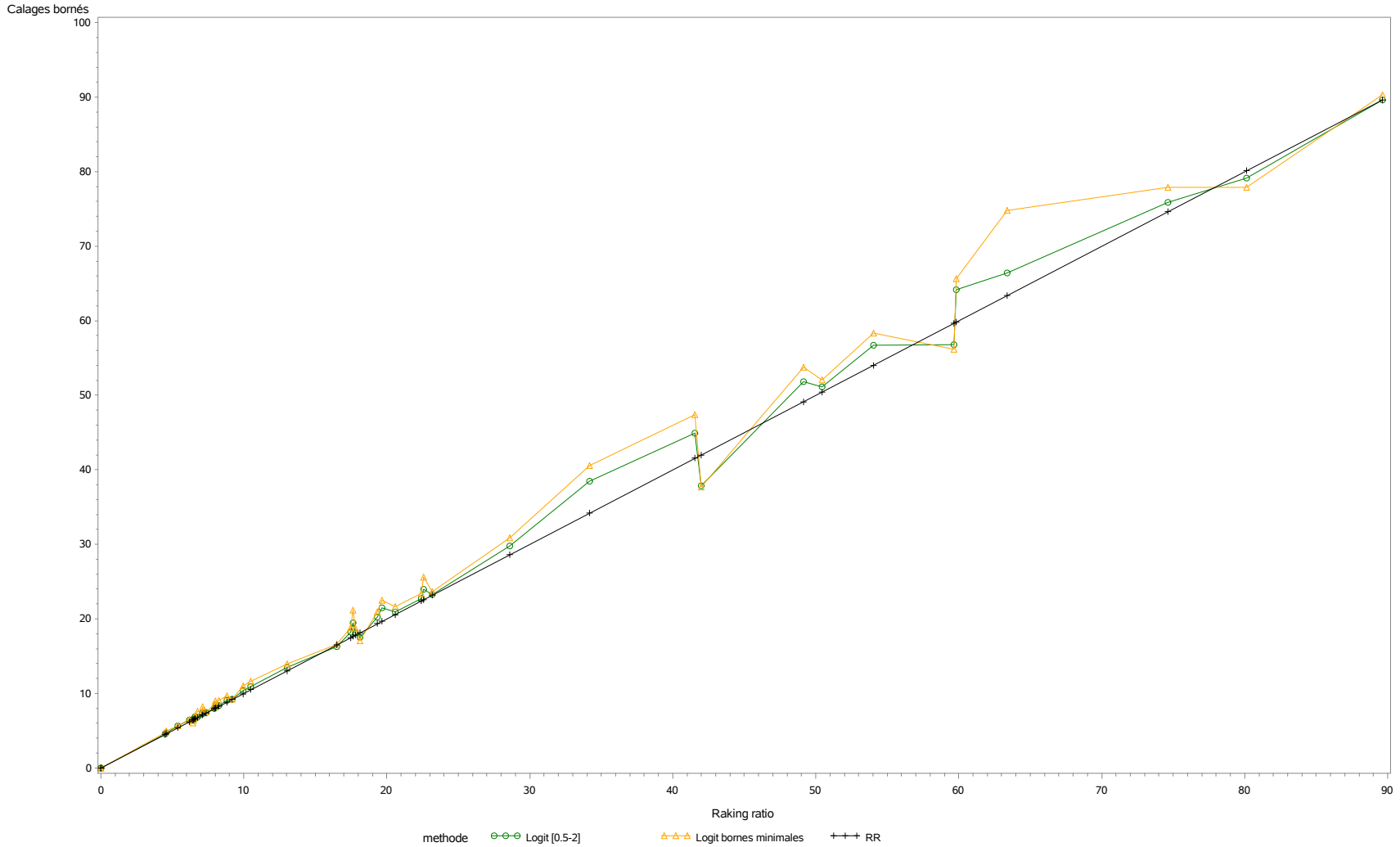
- le calage par raking ratio conduit, quel que soit l'échantillon considéré, à une distribution des rapports de poids présentant une allure log-normale centrée aux alentours de 1, sans point d'accumulation mais avec des déformations maximales potentiellement extrêmes, certains facteurs de calage pouvant aller à la baisse jusqu'à $2,42 \cdot 10^{-16}$ et à la hausse jusqu'à 80 ;
- le calage par méthode logit avec des bornes L=0,5 et U=2 conduit, quel que soit l'échantillon considéré, à une distribution des rapports de poids présentant une allure relativement log-normale centrée aux alentours de 1 avec peu d'accumulations aux bornes (moins de 10 % des rapports de poids collés à l'une ou l'autre borne) ;
- enfin, le calage par méthode logit sur bornes minimales conduit généralement à de fortes accumulations de rapports de poids aux bornes : plus de 50 % des échantillons présentent au moins 50 % des rapports de poids collés à l'une ou l'autre des deux bornes.

Les graphiques en figures 1 et 2 permettent de comparer les RRMSE des trois estimateurs calés, pour des estimations portant sur 16 variables d'intérêt [4] ventilées selon les trois grands secteurs d'activité du commerce commerce automobile (117 unités parmi les 1 000 unités de chaque échantillon), commerce de gros (331 unités parmi les 1 000 unités de chaque échantillon) et commerce de détail (552 unités parmi les 1 000 unités de chaque échantillon). Le graphique en figure 2 est un « zoom » du graphique du haut sur la portion [0 - 30] de l'axe des abscisses.

On constate que le calage sur bornes minimales – induisant de fortes accumulations de rapports de poids aux bornes – conduit à un estimateur moins efficace que celui obtenu avec un calage avec des bornes L=0,5 et U=2 beaucoup moins serrées – permettant un contrôle correct des déformations maximales de poids tout en générant peu d'accumulation de rapports de poids aux bornes – lui même légèrement moins efficace que l'estimateur du raking ratio. Ce dernier présente toutefois l'inconvénient de générer des poids potentiellement extrêmes, susceptibles de créer des unités influentes pour certaines variables, raison pour laquelle on lui préfère en général une méthode bornée.
On observe des résultats similaires (présentés dans [5]) pour les estimations ventilées par tranches d'effectif (les tailles de ces domaines parmi les 1 000 unités de chaque échantillon variant de 119 à 415), ainsi que pour les estimations portant sur l'ensemble du champ.

4. Nombre d'entreprises, chiffre d'affaires, ventes de marchandises, ventes de bien, ventes de services, total des achats, valeur ajoutée, effectifs salariés, masse salariale, excédent brut d'exploitation, résultat comptable, total des amortissements, total des provisions, actif total, passif total et investissement corporel brut.

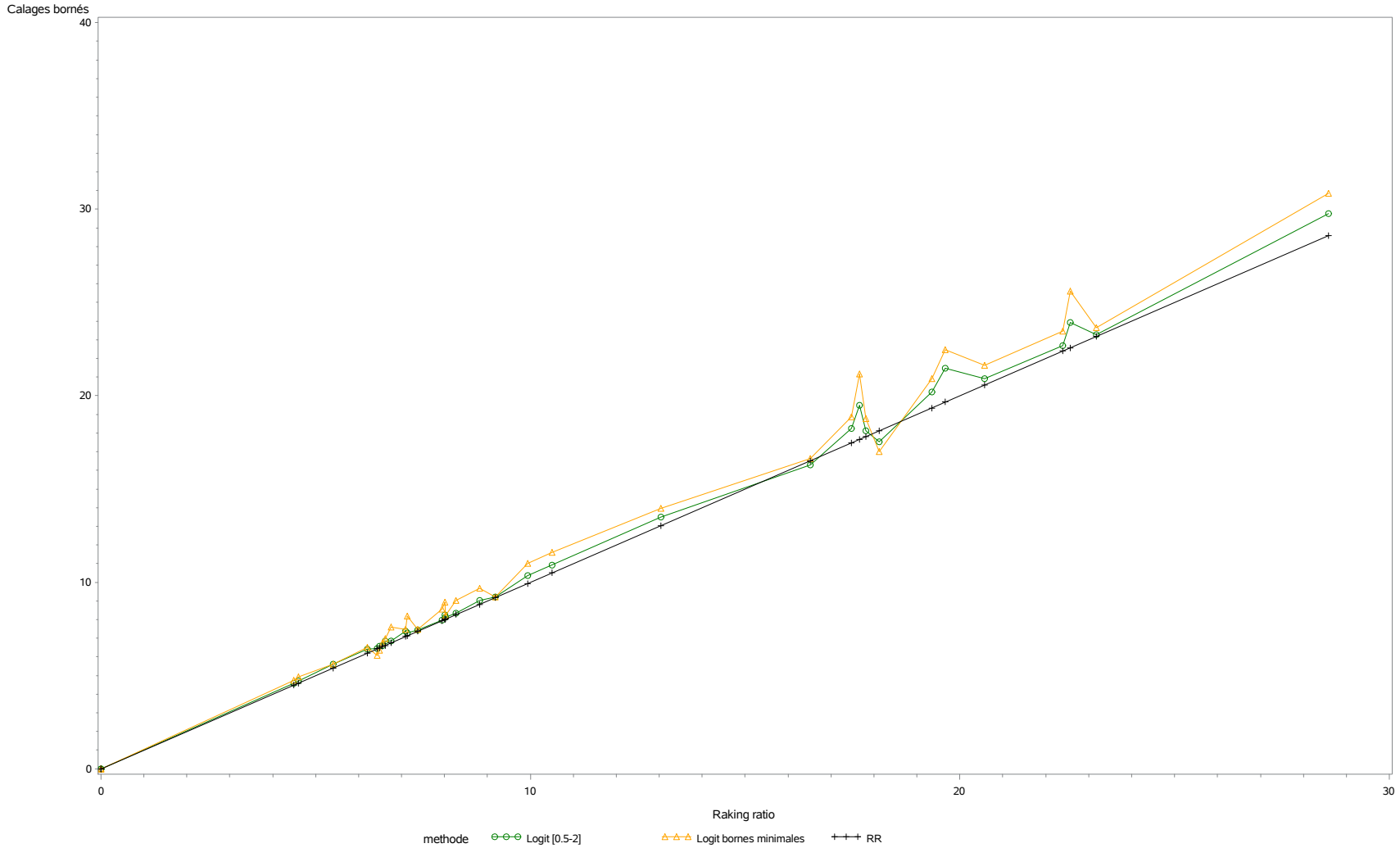
FIGURE 1 – RRMSE (en %) des estimateurs par division



∞

Note de lecture : chaque point représente le RRMSE de l'estimateur du total d'une variable donnée sur un domaine donné pour un scénario de calage donné.

FIGURE 2 – RRMSE (en %) des estimateurs par division (zoom du graphique en figure 1)



Note de lecture : chaque point représente le RRMSE de l'estimateur du total d'une variable donnée sur un domaine donné pour un scénario de calage donné.

Par ailleurs, cette inefficacité de l'estimateur calé sur bornes minimales par rapport à un estimateur calé sur des bornes plus lâches augmente avec l'intensité de la « saturation » aux bornes de calage : plus l'accumulation de rapport de poids aux bornes est importante, moins l'estimateur calé est efficace.

C'est ce que l'on observe lorsque l'on calcule les RRMSE de l'estimateur calé sur bornes minimales sur des sous-ensembles des 40 000 échantillons définis en fonction de l'intensité de la saturation aux bornes de calage : échantillons pour lesquels moins de 50 % des rapports de poids sont collés à l'une ou l'autre des bornes (∼50 % des échantillons), échantillons pour lesquels plus de 50 % des rapports de poids sont collés à l'une ou l'autre des bornes (∼50 % des échantillons), échantillons pour lesquels plus de 70 % des rapports de poids sont collés à l'une ou l'autre des bornes (sous-ensemble du sous-ensemble précédent, environ 10 % des échantillons). Ces RRMSE sont représentés sur les graphiques en figures 3 et 4, en sus des RRMSE des estimateurs déjà présentés en figures 1 et 2.
Là encore, on observe des résultats similaires (présentés dans [5]) pour les estimations ventilées par tranches d'effectifs ainsi que pour les estimations portant sur l'ensemble du champ.

Il semble donc que le fait de choisir des bornes de calage trop serrées nuise à la qualité de l'estimateur calé, en particulier lorsque ces bornes de calage serrées conduisent à de fortes accumulations de rapports de poids à ces bornes.

Conclusion

Après avoir présenté la problématique du calage sur bornes minimales, nous avons proposé un algorithme linéaire de résolution de ce programme de taille $\mathcal{O}(n^2)$, implémenté dans le package R Icarus.

Ce programme de calage sur bornes minimales nous a permis de réaliser une étude par simulations, visant à quantifier l'impact du choix des bornes de calage sur la qualité des estimateurs calés. L'objectif était en particulier de trancher entre deux approches fréquemment utilisées en pratique pour choisir les bornes de calage :

- procéder à un calage sur bornes minimales, quitte à observer de fortes accumulations de rapports de poids aux bornes ;
- retenir des bornes L et U permettant d'assurer un contrôle suffisant des déformations maximales de poids induites par le calage tout en évitant les phénomènes d'accumulation aux bornes.

Les résultats de ces simulations tendent à invalider la première approche au bénéfice de la seconde, qui semble donc constituer une bonne heuristique permettant d'assurer un choix judicieux des bornes de calage.

FIGURE 3 – RRMSE (en %) des estimateurs par division

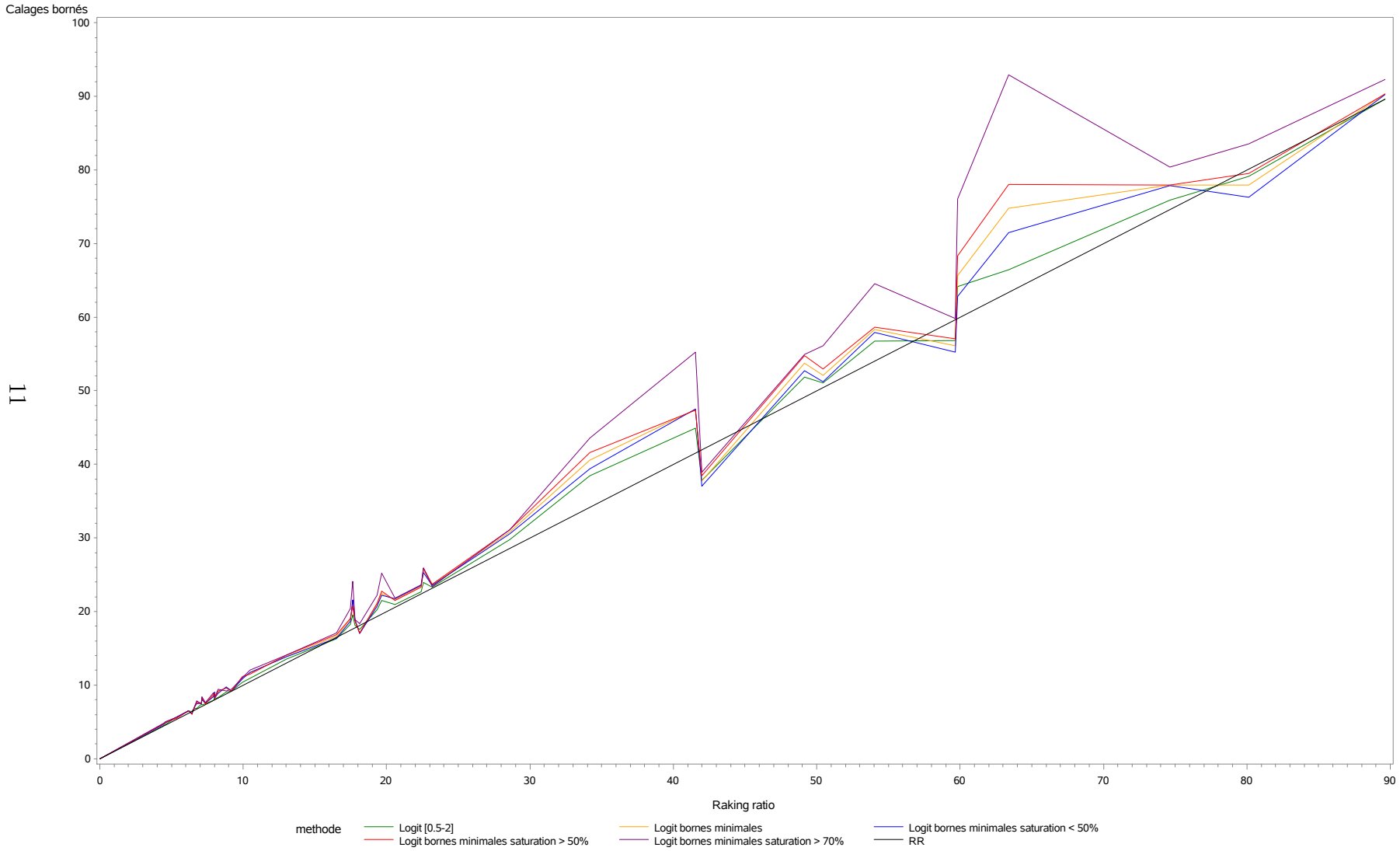
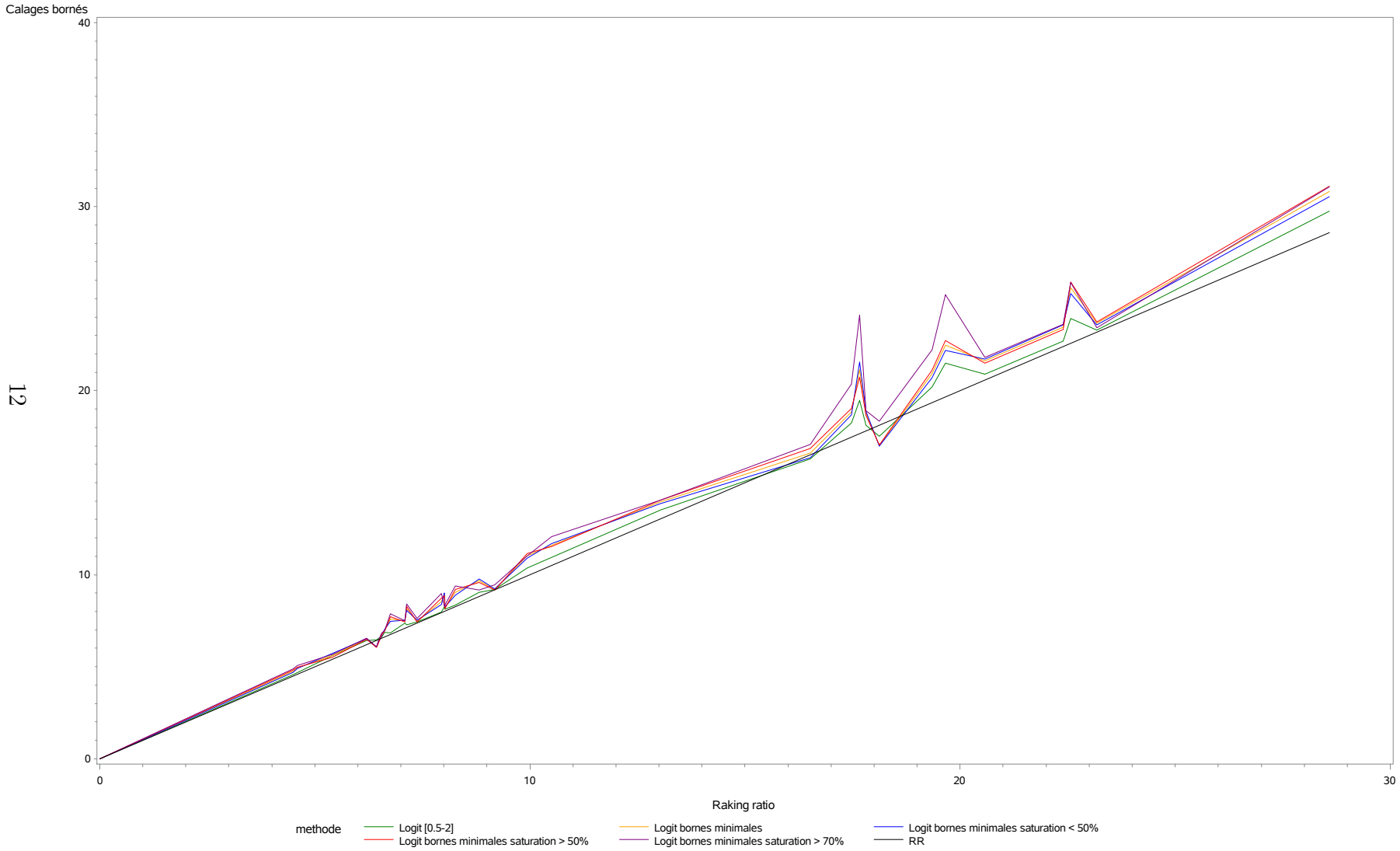


FIGURE 4 – RRMSE (en %) des estimateurs par division (zoom du graphique en figure 3)



Références

[1] Philippe Brion. Esane, le dispositif rénové de production des statistiques structurelles d'entreprises. *Courrier des statistiques N 130, Insee*, 2011.

[2] George B Dantzig, Alex Orden, Philip Wolfe, et al. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2) :183–195, 1955.

[3] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418) :376–382, 1992.

[4] Monique Graf. Calage serré des poids d'enquête. *Actes 8eme colloque francophone Sondages*.

[5] Emmanuel Gros. Les différentes méthodes de calage – le choix des « bornes » de calage. *Présentation au Séminaire de Méthodologie Statistique du département des méthodes statistiques de l'Insee*, 2016.

[6] K Hornik, D Meyer, and Ch Buchta. slam : Sparse lightweight arrays and matrices. *R package version 0.1-32, URL http ://CRAN. R-project. org/package= slam*, 2015.

[7] Antoine Rebecq. Icarus : an r package for calibration in survey sampling. R package version 0.2.0, 2016.

[8] Gildas Roy and Aurélie Vanheuverzwyn. Redressement par la macro calmar : applications et pistes d'amélioration. *Traitements des fichiers d'enquête*, pages 31–46, 2001.

[9] Olivier Sautory. La macro calmar. redressement d'un échantillon par calage sur marges. *Document F9310, DSDS, INSEE*, 1993.

[10] Stefan Theussl, Kurt Hornik, Christian Buchta, Andrew Makhorin, Timothy A Davis, Niklas Sorensson, Mark Adler, Jean-loup Gailly, and Maintainer Stefan Theussl. Package 'rglpk'. 2015.

[11] Camille Vanderhoeft. Generalized calibration at statistic belgium. *Technical Report 3, Statistics Belgium WorkingPaper*, 2001.