Statistics Belgium

Working Paper

La Direction générale Statistique et Information économique propose des informations statistiques impartiale. Les informations sont diffusées conformément à la loi, notamment pour ce qui concerne leur confidentialité.

Nous classons les statistiques en huit domaines :

- Généralités
- Territoire et environnement
- Population
- Société
- Économie et finances
- Agriculture et activités assimilées
- Industrie
- Services, commerce et transports

Éditeur responsable : N. DEMEESTER
© 2006, SPF ECONOMIE
DIRECTION GENERALE STATISTIQUE ET INFORMATION ECONOMIQUE | éditeur
B-1000 Bruxelles – 44 rue de Louvain

Statistics Belgium

Working Paper n° 11

Evaluation de trois logiciels de calage : g-Calib 2.0, Calmar 2 et Bascula 4.0

ETUDE REALISEE PAR

GUILLAUME CHAUVET
JEAN-CLAUDE DEVILLE
MOHAMMED EL HAJ TIRARI
JOSIANE LE GUENNEC

Année 2005

Table des matières

Introduction

Introduction

Pour pouvoir améliorer la précision des estimations des paramètres de la population, le calage est parmi les méthodes les plus utilisées en pratique. En partant d'un estimateur classique ne tenant pas compte d'une information auxiliaire, le calage consiste à modifier les poids de cet estimateur de telle sorte à respecter une propriété voulue mettant à profit l'information auxiliaire disponible.

L'objectif de cette étude est l'évaluation de g-Calib 2, qui est un software spécialisé dans les méthodes de calage, tout en le comparant avec deux autres softwares de calage : Calmar 2 et Bascula 4.0. Cette étude est le résultat d'une commande faite par l'Institut National de Statistique (*INS* – Belgique) dont l'objectif est de faire une expertise du software de calage g-Calib 2 développé à l'*INS*. L'étude est réalisée par une équipe de chercheurs de l'Université Libre de Bruxelles (Belgique) et de l'Ecole nationale de la statistique et de l'analyse de l'information (France).[1]

Notons que les trois softwares : g-Calib 2, Calmar 2 et Bascula 4.0 sont développés respectivement en *SPSS*, *SAS* et *Blaise*. Ces softwares sont gratuits.

Cadre théorique

En théorie d'échantillonnage, on s'intéresse à une population U composée de N unités (ou individus) notées par :

$$U = \{1,...,k,...,N\} .$$

[1] Demande n° 2004/CU F/0047 du SPF Economie. Le rapport a été soumis en mai 2005 ; la révision est terminée en juin 2006. Le contenu de la présente étude n'engage aucunement le SPF Economie, la responsabilité de cette étude revenant à ses seuls auteurs.

Dans cette population, on sélectionne un échantillon *s* de taille *n* en utilisant un plan de sondage $p(s)$. Pour tout $k \in U$, on note par π_k la probabilité d'inclusion dans l'échantillon *s*.

Le but d'un sondage est de décrire une variable d'intérêt ***y*** prenant des valeurs pour chaque unité *k* de la population. La valeur prise par la variable d'intérêt ***y*** pour la $k^{\text{ième}}$ unité d'observation de la population *U* est notée par y_k.

En général, on ne cherche pas à connaître la valeur y_k prise par chacune des unités de la population. L'intérêt se porte plutôt sur une fonction de ces valeurs y_k $(k \in U)$ qui constitue l'information que l'on cherche à acquérir.

Ainsi, l'objectif est d'estimer une fonction des valeurs de la variable d'intérêt ***y***, comme par exemple, le total donné par :

$$t_{\mathbf{y}} = \sum_{k \in U} y_k$$

L'estimateur classique qu'on utilise pour estimer le total $t_{\mathbf{y}}$ est l'estimateur d'Horvitz-Thompson :

$$\hat{t}_{\mathbf{y}\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

ce qui revient à affecter à chaque unité de l'échantillon un poids d_k égal à l'inverse de sa probabilité d'inclusion. Bien qu'il s'agisse d'un estimateur sans biais, cet estimateur a l'inconvénient de ne tenir compte d'aucune information auxiliaire qui pourrait améliorer la qualité des estimations obtenues.

En effet, en pratique, on peut disposer d'une information auxiliaire qui peut être la connaissance des valeurs d'une (ou de plusieurs variables) sur toutes les unités de la population ou en une fonction de ces valeurs. Cette information auxiliaire peut être utilisée pour améliorer la précision des estimations obtenues des paramètres de la population.

Dans le cadre du calage, on suppose qu'on dispose de *J* variables auxiliaires $\mathbf{x}_1, ..., \mathbf{x}_j, ..., \mathbf{x}_J$ dont les valeurs sont connues sur l'échantillon et dont on connaît les totaux sur la population :

$$t_{\mathbf{x}_j} = \sum_{k \in U} x_{jk} .$$

Ces variables auxiliaires peuvent être des variables catégorielles (calage sur marge). Le calage est une méthode dont l'objectif est d'améliorer la précision des estimateurs des paramètres de la population en mettant à profit cette information auxiliaire. Pour cela, on cherche à estimer le total $t_{\mathbf{y}}$ de la variable d'intérêt $\boldsymbol{y}$ à l'aide d'un estimateur de la forme :

$$\hat{t}_{\mathbf{y}w} = \sum_{k \in s} w_k y_k \ ,$$

où les poids w_k affectés aux individus sont « proches » (selon une fonction de distance à préciser) des poids d_k de l'estimateur d'Horvitz-Thompson, et vérifient les équations de calage

$$\sum_{k \in s} w_k x_{jk} = \sum_{k \in U} x_{jk} = t_{\mathbf{x}_j} \quad \text{pour tout } j = 1, ..., J \ .$$

On cherche donc un estimateur qui tient compte de l'information auxiliaire, puisqu'il « cale » l'échantillon sur les totaux des variables auxiliaires. De plus, on a de fortes chances d'obtenir un estimateur avec un faible biais puisque ces poids w_k sont en principe « peu différents » de ceux de l'estimateur d'Horvitz-Thompson.

Evaluation de g-Calib 2

Evaluation de g-Calib 2
(étude théorique)

Dans cette première partie, l'évaluation de g-Calib 2, tout en le comparant aux deux autres softwares de calage, va se faire uniquement sur le plan théorique en se limitant aux aspects suivants : la qualité de la documentation fournie avec g Calib 2 et la facilité d'obtenir celle-ci, les méthodes de calage implémentées dans g-Calib 2, et les niveaux de calage que g-Calib 2 permet de réaliser. Les autres aspects d'évaluation faisant partie du cahier des charges de cette étude seront évalués dans le rapport final.

1 La qualité de la documentation relative à g-Calib 2

La documentation relative à g-Calib 2 est assez détaillée avec un manuel d'utilisation très complet contenant toutes les informations théoriques et pratiques sur g-Calib 2. En effet, en plus des informations concernant l'installation et les fonctionnalités de g-Calib 2, ce manuel d'utilisation contient également une partie théorique bien détaillée permettant de comprendre les principes et la théorie sur lesquels se basent les méthodes de calage implémentées dans g-Calib 2. Cependant, on note que le recours systématique à des notations matricielles rend parfois les raisonnements assez difficiles à suivre. g-Calib 2 dispose aussi d'une base de données illustrative (sous format *MS-Excel*) permettant de générer des données pour pouvoir faire des études par simulations (Vanderhoeft, 2002).

Contrairement aux deux autres softwares, l'accès à la documentation relative à g-Calib 2 est facilité par le développement (par la direction générale de la Statistique et Information économique) d'une page Web sur le sujet du calage

(http://www.statbel.fgov.be/studies/cal_en.asp), à partir de laquelle les différents textes sur le calage et sur l'utilisation de g-Calib 2 sont rendus publics.

On note que Calmar 2 dispose également d'une documentation assez détaillée (Sautory et Le Guennec, 2003) mais l'accès à celle-ci n'est pas facile et une grande partie de cette documentation ne peut être accessible via internet. Bascula 4.0 propose un manuel de référence (reprenant, entre autres, quelques éléments théoriques sur la pondération) et un fichier d'aide consultable sur écran.

2 Les méthodes de calage

Pour mesurer la proximité entre les poids w_k et d_k, plusieurs fonctions de distance G ont été proposées. Chaque fonction de distance permet de construire une méthode de calage. Il est important de noter ici que, pour certaines méthodes de calage bornées, les fonctions de distance sont définies différemment dans les trois logiciels de calage (Voir annexe 1). En effet, pour ces fonctions le domaine de définition des bornes diffère d'un logiciel à un autre. Dans la partie pratique de cette étude, nous allons vérifier si le fait d'utiliser des domaines de définition différents pour ces bornes a un effet significatif sur les résultats obtenus avec ces différents softwares de calage.

On note que g-Calib 2 et Calmar 2 utilisent pratiquement les mêmes fonctions de distance (Calmar 2 contient une fonction de plus : *sinus hyperbolique*). Par contre, Bascula 4.0 est beaucoup plus restreint que les deux autres logiciels puisqu'il n'utilise que des cas particuliers de ces fonctions de distance.

Evaluation de g-Calib 2

Après avoir vérifié les différentes méthodes de calage implémentées dans les trois softwares, on peut conclure que, à l'exception de la méthode *sinus hyperbolique* (implémentée uniquement dans Calmar 2), on retrouve toutes les autres méthodes de calage usuelles dans g-Calib 2.

3 Les niveaux de calage

En théorie de calage, ce dernier peut se faire également quand on a plusieurs niveaux d'observation. En effet, par exemple, pour un plan de sondage à deux degrés, un échantillon d'unités primaires est sélectionné et dans chaque unité primaire sélectionnée, on sélectionne des unités secondaires. Si on dispose d'une information auxiliaire spécifique à chaque type d'unités sélectionnées, on peut effectuer également un calage des différentes unités observées, produisant des pondérations identiques pour les unités secondaires incluses dans la même unité primaire. Un tel calage peut inclure également plusieurs niveaux d'observation à condition qu'on dispose d'une information auxiliaire spécifique à chaque type d'unités sélectionnées.

3.1 Les niveaux de calage réalisés par g-Calib 2

Comme pour la plupart des softwares de calage, g-Calib 2 permet de réaliser des calages sur plusieurs niveaux d'observation. Ainsi, les possibilités offertes par g-Calib 2 sont :

- ***Un calage simple*** sur une table d'enquête ne comprenant qu'un seul niveau d'observation. Il s'agit d'un calage au niveau des unités de la population. C'est le calage le plus classique où le programme permet de réaliser un calage sur le total, dans la population, de l'information auxiliaire disponible.
- ***Calage au niveau des unités de la population tout en fournissant des poids identiques pour les unités appartenant à une même grappe*** : lorsque l'information auxiliaire est disponible sur les unités de la population et l'objectif est de produire des nouveaux poids w_g pour les unités de la population des grappes, le programme permet de réaliser un calage qui produit en plus des poids $w_{k,g}$ identiques pour les unités de la population appartenant à la même grappe U_g. En effet, après avoir agrégé l'information auxiliaire au niveau des grappes, le programme cale le total, sur les unités de la population des grappes, de l'information auxiliaire.

- ***Calage simultané au niveau des unités de la population et au niveau des grappes*** : lorsque qu'on dispose d'une information auxiliaire sur les unités de la population et d'une autre sur la population des grappes, le programme permet de réaliser un calage simultané sur ces deux types d'informations de telle sorte à produire des poids w_k identiques pour les unités de la population appartenant à la même grappe.

3.2 Les niveaux de calage réalisés par Calmar 2

A l'instar de g-Calib 2, Calmar 2 est également un software qui permet de réaliser des calages sur plusieurs niveaux d'observation. En effet, tous les types de niveaux de calage implémentés dans g-Calib 2 peuvent être réalisés par Calmar 2. Comparé à g-Calib 2, Calmar 2 permet de réaliser d'une manière automatique un ***Calage simultané entre trois niveaux d'observation emboîtés dans un sondage à deux degrés***. Ce dernier type de calage peut être réalisé lorsqu'on dispose d'une information auxiliaire sur les unités primaires, sur chaque unité secondaire appartenant à l'unité primaire de l'échantillon, et sur les unités secondaires sélectionnées au deuxième degré. Par exemple, sur les ménages, sur tous les individus de ménage, et sur les individus appartenant au champ du tirage Kish. On note que ce dernier type de niveaux de calage peut être également réalisé par g-Calib 2, mais pas d'une manière automatique.

3.3 Les niveaux de calage réalisés par Bascula 4.0

Un calage sur deux niveaux (unités de la population [individus] et grappes) est prévu dans Bascula 4.0, son fonctionnement nécessitant une variable additionnelle, à savoir $1/e_m$, où e_m est la taille de la grappe m.

3.4 Evaluation de g-Calib 2

Pour le calage sur plusieurs niveaux, g-Calib 2 est un software qui offre un large choix de possibilités de calage correspondant aux principaux niveaux d'observation qu'on peut rencontrer en pratique. De plus, d'autres types de niveaux de calage, comme le calage sur plus de deux niveaux d'observation,

peuvent être également réalisés par g-Calib 2, mais pas d'une manière automatique. On peut donc conclure que, de point de vue niveaux de calage, g Calib 2 paraît un software de calage complet qui peut être utilisé dans presque toutes les situations rencontrées en pratique.

4 D'autres possibilités spécifiques à g-Calib 2

On note que g-Calib 2 a l'avantage, par rapport aux deux autres softwares, d'inclure quelques options de plus dont l'objectif est de perfectionner les méthodes de calage implémentées dans le programme. Un bref aperçu de ces options est donné dans ce qui suit :

- Lorsque des strates peuvent être construites, le programme permet de réaliser un calage au niveau de ces strates en utilisant l'information auxiliaire disponible sur les unités de la population, c'est-à-dire en faisant des calages sur les totaux des strates de l'information auxiliaire. Le calage sur des strates a l'avantage de rendre g-Calib 2 plus performant puisque le calcul des nouveaux poids w_k se fait strate après strate, ce qui permet d'améliorer la gestion de la base de données.
- Possibilité de tenir compte, lors de l'opération de calage, d'un *paramètre d'échelle* ϕ permettant d'ajuster les poids de départ d_k (cette option est disponible également dans Calmar 2). Si ce paramètre d'échelle est inconnu, g-Calib 2 permet aussi de produire une estimation de ce paramètre (dans chaque strate de calage).
- Possibilité de pondérer dans les équations de calage par une variable q_k. Cette pondération permet de retrouver en particulier l'estimateur par le ratio.
- Dans le cas où on a affaire à une grande base de données, pour réduire le temps d'exécution des programmes calculant les nouveaux poids w_k, une technique appelée *collapsing* est implémentée dans g-Calib 2. En effet, on note que pour les unités de la population ayant des vecteurs $(\mathbf{x}_k, q_k)$

identiques, c'est-à-dire avec des valeurs relatives aux variables de calage qui sont égales, les rapports de poids $(w_k/d_k) = q_k F(\mathbf{x}_k'\boldsymbol{\lambda})$ sont également identiques. La technique de *collapsing* permet de ne considérer dans le calcul des rapports de poids w_k/d_k que les unités ayant des vecteurs $(\mathbf{x}_k, q_k)$ différents et d'attribuer par la suite le rapport de poids calculé w_k/d_k à toutes les unités restantes ayant le même vecteur $(\mathbf{x}_k, q_k)$, ce qui permet de réduire le temps du calcul de ces poids.

- Possibilité de normaliser les variables de calage, ce qui permet de ramener toutes les variables de calage à une même échelle. Bascula 4.0 choisit automatiquement les échelles des variables de calage, ceci dans le souci d'obtenir un meilleur comportement numérique.

5 Conclusion

Après avoir consulté la documentation relative à g-Calib 2, qui a l'avantage d'être très détaillée, g-Calib 2 paraît être un software qui tient compte de presque tous les aspects théoriques et pratiques de la théorie de calage. Ainsi, à l'exception de la méthode de *sinus hyperbolique*, g-Calib 2 permet de réaliser des calages en utilisant toutes les autres méthodes de calage usuelles. Le software permet également de réaliser des calages sur différents niveaux d'observation. Malheureusement, contrairement à Calmar 2 qui permet de réaliser un calage jusqu'à trois niveaux d'observation, g-Calib 2 se limite au calage à deux niveaux d'observation. Cependant, on note que le calage à trois niveaux d'observation peut être réalisé en utilisant g-Calib 2, mais pas d'une manière automatique.

De plus, contrairement aux deux autres softwares, g-Calib 2 dispose également de quelques options de plus (calage sur des strates, technique de *collapsing*, possibilité de pondérer les équations de calage, possibilité de normer les variables de calage) dont l'objectif est de perfectionner les méthodes de calage.

On peut donc conclure que, même si certaines opérations ne peuvent pas être réalisées de manière automatique (calage à plus de deux niveaux d'observation),

g-Calib 2 peut être considéré comme un software complet de calage qui permet de répondre à presque tous les problèmes de calage rencontrés en pratique.

Notons que toutes les conclusions émises dans ce rapport intermédiaire ne se basent que sur une évaluation théorique et qui restent toujours sous réserve de vérification. Pendant la deuxième partie de cette étude, l'évaluation de g-Calib 2 va se faire sur le plan pratique en utilisant des bases de données réelles et simulées.

Tests et évaluation des trois logiciels du calage

Tests et évaluation des trois logiciels du calage (étude pratique)

Dans cette deuxième partie du rapport, l'évaluation des logiciels g-Calib 2, Calmar 2 et Bascula 4.0 va se faire sur le plan pratique en considérant les aspects d'évaluation faisant partie du cahier des charges de cette étude. Tous les tests effectués ont été réalisés en utilisant les données suivantes :

Les répondants à l'enquête sur l'emploi en France, réalisée en 2001 par l'Insee, constituent la population de référence, soit 114 111 individus dans 58 189 logements.

L'échantillon simulé utilisé pour le calage comprend 8000 logements (fichier « echup ») et 15 696 individus de 15 ans ou plus (fichier « echant ») sélectionnés par sondage aléatoire simple de grappes.

Les variables auxiliaires retenues sont celles habituellement utilisées pour redresser cette enquête :

- Pour la population d'individus :
  - âge quinquennal de 15 à 70 ans, puis 70 ans et plus (variable catégorielle)
  - salaire net mensuel (variable numérique)
- Pour la population de ménages :
  - rang du sous-échantillon auquel appartient le ménage (3 modalités)
  - lieu de résidence : Paris/province
  - taille du ménage : personnes seules/autres
  - statut d'occupation du logement en 6 modalités, selon la nomenclature du recensement

Pour le calage par strate, la population des répondants à l'enquête emploi 2001 est stratifiée à l'aide de la variable TZPIU (Tranche de Zone de Peuplement

Industriel ou Urbain). Comme certains effectifs dans l'échantillon sont faibles pour les ZPIU rurales, on opère des regroupements. On obtient ainsi 6 strates :

- Modalités 0 à 3 de TZPIU (communes hors ZPIU ou de ZPIU de moins de 20 000 habitants) : strate 1
- Modalité 4 de TZPIU (ZPIU de 20 000 à 50 000 habitants) : strate 2
- Modalité 5 de TZPIU (ZPIU de 50 000 à 100 000 habitants) : strate 3
- Modalité 6 de TZPIU (ZPIU de 100 000 à 200 000 habitants) : strate 4
- Modalité 7 de TZPIU (ZPIU de 200 000 à 2 000 000 habitants) : strate 5
- Modalité 8 de TZPIU (ZPIU de paris) : strate 6

On dispose de la même information auxiliaire que précédemment. On réalise un calage simple avec l'information disponible au niveau individu.

Dans la population des répondants, on prélève un échantillon stratifié avec allocation proportionnelle en visant un échantillon de 8 000 ménages (fichier echupstrat). L'ensemble des individus de 15 ans et plus appartenant à ces ménages constitue l'échantillon d'individus (fichier echup_strat).

6 Evaluation de g-Calib 2

6.1 Aspects généraux de g-Calib 2

6.1.1 Installation du logiciel

La version 2 du logiciel g-Calib 2 a été testée dans la configuration suivante :

- système d'exploitation Windows NT
- SPSS version 12
- installation individualisée sur postes fixes sans accès partagé.

Lors de l'installation, faite par le service informatique, il est apparu qu'il semblait impossible d'avoir une utilisation partagée de g-Calib 2, c'est pourquoi on a renoncé à l'implanter de façon unique sur le serveur du réseau local. On a rencontré les difficultés suivantes :

- le nom du répertoire dans lequel est implanté le programme g-Calib 2 (« Installation Directory ») doit obligatoirement comprendre la racine : C:\program files\, ce qui n'est pas clairement spécifié dans la documentation. Lorsqu'on modifie le nom proposé par défaut à l'installation, l'exécution ultérieure d'une requête entraîne sa ré-initialisation automatique à une valeur par défaut et provoque un bug.
- le programme est adapté à la version en langue anglaise du logiciel SPSS. Dans la version française, certains paramètres renseignés par « Yes » doivent être modifiés en « Oui ». C'est pourquoi il a été plus rapide d'utiliser les fichiers de production SPSS de préférence à l'interface Visual Basic pour exécuter successivement un grand nombre de requêtes g-Calib 2.

6.1.2 Facilité d'utilisation

g-Calib 2 est écrit en SPSS. Il est compatible avec la version 12 de ce logiciel. L'appel de g-Calib 2 peut se faire sous Windows en cliquant sur une icône si l'on a créé un raccourci.

6.1.2.1 Logiciel

L'interface de saisie des paramètres nécessaires au calage d'une enquête est claire. Cependant, l'oubli d'un paramètre obligatoire ne donne pas lieu à l'affichage systématique d'un message d'erreur, ce qui peut être très pénalisant pour un utilisateur débutant.

Le modèle de calage est défini dans l'interface de saisie. Pour une équation de calage de la forme : $\sum_{k \in s} w_k x_{jk} = X_j$, où X_j désigne la modalité j d'une variable catégorielle X, l'utilisateur doit spécifier le nom x_j de la variable indicatrice qui désignera ce caractère dans la table de travail, écrire l'équation logique définissant le domaine correspondant dans la table échantillon, puis spécifier le total X_j dans la population. L'écriture de ces contraintes nécessite de connaître la syntaxe d'une instruction d'affectation dans SPSS, en particulier les règles de concordance des formats et de spécification de la valeur d'une variable caractère.

Chaque contrainte de calage fait l'objet d'une ligne de saisie. Il n'est pas possible de spécifier le nom d'une table contenant la liste des variables auxiliaires et leurs totaux dans la population.

Quant à Bascula 4.0, les variables indicatrices sont créées en interne et la modélisation des poids peut être définie sous forme de tables, forme pratique pour les modèles de grandes tailles.

6.1.2.2 Sauvegarde

Une demande peut être sauvegardée. Le programme sera obligatoirement implanté dans le répertoire de travail spécifié contenant les fichiers de données (table échantillon). Il peut ensuite être rappelé, modifié et ré-exécuté à condition de faire appel à des fichiers de données situés dans le même répertoire.

La table des poids de calage est également stockée dans le répertoire de travail spécifié.

g-Calib 2 crée en sortie, outre la table des poids, un fichier appelé « Temp_Design.sps » généré par le programme et un fichier « Temp.spp » contenant le programme de calage. Ce dernier peut ensuite être rappelé sous Windows, modifié, renommé pour un stockage définitif, et exécuté sans utiliser l'interface de saisie, à condition d'avoir conservé « Temp_Design.sps ». Dans le contexte particulier de l'ENSAI, cette dernière solution s'est révélée plus rapide.

6.1.3 Input

La table contenant l'échantillon doit être en format SPSS. Ce logiciel permet notamment la lecture des formats Excel, Lotus, Dbase, et SAS dans différentes versions (extensions : sd2, sd7, sas7bdat, xpt, et version 6 sous Unix), ainsi que celle de fichiers texte (.txt). Il faut avoir fait cette conversion avant d'exécuter une requête g-Calib 2.

Outre les variables de calage et la variable numérique de pondération initiale[2], la table de données doit contenir, de façon obligatoire :

- une variable identifiant les unités de l'échantillon, qui doit être de format numérique ;
- une variable identifiant la strate, même si le sondage n'est pas stratifié. Cette variable doit être numérique et codifiée de façon séquentielle à partir de 1, sans discontinuités.
- une variable de pondération à spécifier dans le paramètre *qpond* de l'interface de saisie. Dans le cas où l'on ne souhaiterait pas utiliser ce paramètre, cette variable sera la constante 1.

Dans le cas d'un sondage en grappes avec calage simultané, la table des unités secondaires doit contenir l'identifiant de l'unité primaire à laquelle elle appartient

[2] On note que SPSS ne reconnaît pas le nom d'une variable contenant un « blanc souligné » (underscore : _) quand elle a été créée sous SAS.

en plus de l'identifiant de l'unité secondaire. Les deux paramètres doivent alors être spécifiés. Le poids initial doit être présent dans les deux tables.

6.1.4 Fonctionnalités

6.1.4.1 Plan de sondage

Rappelons que g-Calib 2 fournit des poids redressés pour une enquête réalisée selon l'un des plans de sondage suivants (voir § 3.1) :

- Sondage à un seul degré d'observation : calage de l'échantillon sur les totaux dans la population de variables auxiliaires numériques ou catégorielles ;
- Sondage en grappes, avec calage sur des totaux dans la population des unités secondaires respectant l'égalité des poids entre unités d'une même grappe ;
- Sondage en grappes avec calage simultané sur des totaux dans la population des grappes et sur des totaux dans la population des unités secondaires.

Dans tous les cas, le calage peut être réalisé par strate, dans la totalité ou dans une sélection de strates, ou sur les seuls totaux dans la population entière. Avec une fonction de calage bornée, le calage appliquera les mêmes bornes dans chaque strate spécifiée.

6.1.4.2 Modèle de redressement

g-Calib 2 propose quatre fonctions de calage :

- fonction linéaire : $F(u) = 1 + u$
- fonction exponentielle : $F(u) = e^u$
- fonction linéaire tronquée : $F(u) = 1 + u$, avec $L <= F(u) <= U$
- fonction logit bornée : $F(u) = \dfrac{L(U-1) + U(1-L)\exp(Au)}{(U-1) + (1-L)\exp(Au)}$

avec $A = \frac{U - L}{(U - 1)(1 - L)}$ et $0 \leq L < 1 < U$

et résout les équations de calage par la méthode itérative de Newton, mais en utilisant la technique des matrices inverses généralisées.

Selon la fonction de calage et le type de variables auxiliaires utilisées, les poids fournis par g-Calib 2 redressent l'enquête par régression, par post-stratification, par raking-ratio, par le ratio.

6.1.4.3 Facteur d'échelle

Si l'on choisit un facteur d'échelle différent de 1, les poids initiaux sont tous multipliés par cette valeur avant le calage. Par défaut, le facteur d'échelle est calculé automatiquement dans g-Calib 2 comme le rapport entre le total $X_1 / \hat{X}_{1\pi}$, où X_1 est la première variable auxiliaire spécifiée dans la table des équations de calage, X_1 le total de cette variable dans la population et $\hat{X}_{1\pi} = \sum_{k \in s} d_k x_{1k}$ son estimation à partir de l'échantillon et des poids initiaux. On peut utiliser ce paramètre pour un redressement uniforme de la non-réponse, à condition de spécifier en première contrainte de calage une constante égale à 1 et la taille de la population.

6.1.4.4 Fusion des individus en classes

g-Calib 2 crée une table de travail contenant une observation par vecteur des variables de calage, résultat du regroupement (*collapsing*) des individus de l'échantillon présentant les mêmes modalités des variables auxiliaires. C'est cette table fusionnée qui est utilisée en entrée des itérations, avec une pondération tenant compte de l'effectif de la classe.

6.1.5 Output

6.1.5.1 Fichier des poids de calage

g-Calib 2 fournit en sortie un fichier de format SPSS (extension : .svo) contenant les poids de calage. Ce fichier contient deux groupes de variables.

Variables de l'utilisateur :

- identifiant de l'observation
- variable identifiant la strate
- poids initial
- variable contenant la q-pondération
- variables de calage (indicatrices des modalités dans le cas des variables catégorielles)

Variables créées par g-Calib 2:

- *Scale* : valeur du facteur d'échelle
- *Scawei* : poids initial multiplié par le facteur d'échelle
- *Calwei* : poids de calage
- *G_weig* : rapport entre les variables *calwei* et *scawei*. Lorsque le facteur d'échelle est choisi à 1, c'est le rapport entre poids de calage et poids initial.

Le fichier des poids est trié par strate et par identifiant de l'unité. Il peut donc se trouver dans un ordre différent du fichier échantillon spécifié en entrée.

Chaque exécution de g-Calib 2 crée une nouvelle table de poids. Si l'on spécifie le nom d'une table existante, celle-ci est écrasée.

6.1.5.2 Fichiers de programmes

g-Calib 2 crée dans le répertoire de l'utilisateur deux fichiers de syntaxe SPSS :

- *Temp_Design.sps* est généré par chaque exécution du programme.
- *Temp.spp* est généré par chaque exécution lancée par l'interface de saisie. Il contient les paramètres de la demande et les commandes d'exécution des modules du programme. Il peut être renommé sous un nom propre à l'utilisateur, ce qui permet de le rappeler sous Windows pour une éventuelle modification des paramètres et exécution.

6.1.5.3 Autres fichiers

g-Calib 2 conserve dans le répertoire de l'utilisateur des fichiers contenant les données de l'échantillon (*E-desmat.sav* pour le niveau unité secondaire, *C-desmat.sav* pour le niveau grappe), triés dans le même ordre que celui des poids de calage.

Il stocke aussi des fichiers contenant les totaux des variables de calage (fichiers *E-totals.sav* et *C-totals.sav*, équivalents des tables de marges dans Calmar 2) et des fichiers de travail intermédiaires (*Survey.sav*, *Bench.sav*) dont l'utilité pour l'utilisateur est hasardeuse. En cas de non-convergence ou de bug, ils n'apportent aucune information utile.

6.1.5.4 Editions

Même en choisissant les options limitant l'apparition du log, l'ergonomie de SPSS ne permet pas de séparer complètement le log et les éditions en sortie, ce qui complique la lecture du fichier de résultats.

On peut notamment regretter l'absence de titres facilitant l'interprétation et la recherche des résultats utiles. On trouve dans le listing produit :

- un rappel de quelques-uns des paramètres du calage (type et fonction de calage, numéros des strates, bornes, mais pas le nom des tables spécifiées)
- la liste des variables de calage
- le nombre d'observations résultant de la fusion des individus identiques
- la valeur du facteur d'échelle
- un tableau donnant, à chaque itération :
    - le rang de la matrice Φ'
    - la valeur de la fonction de distance
    - la valeur du critère d'arrêt (« change »)
    - le nombre de poids négatifs
- un tableau présentant les totaux des variables de calage estimés respectivement avec les poids initiaux, les poids initiaux multipliés par le facteur d'échelle, les poids de calage, ainsi que les totaux dans la population et la différence relative entre l'estimateur calé et la vraie valeur

- la valeur des multiplicateurs de Lagrange après la dernière itération
- le nombre d'itérations réalisées
- une table des quantiles des poids de calage et des rapports de poids calculés respectivement sur la table de contingence et sur celle des individus
- les « box-plots » de la distribution des poids de calage et des rapports de poids dans la table de contingence et dans celle des individus
- la durée d'exécution du programme

Les statistiques et graphiques « box-plots » sont fournis par strate.

6.1.6 Problèmes

Dans les tests réalisés (voir plus loin), on a rencontré des cas de non-convergence totale due au choix des bornes avec une méthode de calage bornée, et des cas de convergence imparfaite sur une partie des variables de calage.

Dans le premier cas un message d'erreur est édité en fin de listing (« LN out of range »).

Le second cas n'est identifiable que par l'analyse du tableau de comparaison des totaux réels aux totaux estimés. Le programme ne fournit pas de message pour attirer l'attention de l'utilisateur.

La technique des matrices inverses généralisées crée un problème avec la fonction linéaire tronquée. Il n'y a pas d'arrêt des itérations dans les cas de non-convergence. Avec cette fonction, la non-convergence est détectée par la non-inversibilité de la matrice Φ' lorsque le nombre de poids égaux aux bornes devient important. Les matrices g-inverses passent outre au problème et fournissent un résultat faux.

6.2 Résultats des tests réalisés

Les temps d'exécution indiqués sont ceux fournis par l'application g-Calib 2 dans ses éditions. Ce sont des durées réelles. Les programmes ayant été exécutés sur un poste relié à un réseau local, ces durées peuvent varier selon la charge du réseau, elles ne sont donc qu'approximatives.

6.2.1 Calage simple sur la population d'individus

On cale la table échantillon « echant » sur les effectifs de la population de 15 ans ou plus par âge quinquennal et sur la somme des salaires mensuels.

6.2.1.1 Résultats obtenus avec g-Calib 2 selon les différentes fonctions de calage

- Fonction linéaire : le calage converge en 52 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,96 et 1,49.
- Fonction exponentielle : le calage converge en 4 itérations et 54,5 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,96 et 1,60.
- Fonction logit pour des rapports de poids compris entre :
    - [0,97-1,6] : le calage converge en 6 itérations et 54,5 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,2] : le calage converge en 6 itérations et 55,2 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,06] : non convergence, avec le message "LN out of range".
- Fonction linéaire tronquée :
    - [0,96-1,6] : le calage converge en 2 itérations et 52,3 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,06] : le calage converge en 6 itérations et 53,6 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,03] : le calage converge imparfaitement en 5 itérations et 50,5 secondes, mais pas de message explicite dans les résultats édités. La différence relative maximum entre l'estimateur calé et le total réel est de 2,1 %.

- [0,98-1,02] : le calage converge imparfaitement en 6 itérations et 53,2 secondes, mais pas de message explicite dans les résultats édités. La différence relative entre l'estimateur calé et le total réel est comprise entre 0,1 % et 3 %. Les rapports de poids sont tous égaux à l'une des deux bornes, aucun n'est compris entre les deux.

6.2.1.2 Comparaison avec Calmar 2

Lorsque le calage converge parfaitement, g-Calib 2 et Calmar 2 fournissent des poids qu'on peut considérer égaux pour chaque individu. Les écarts maximums sont compris entre 10^{-8} et 10^{-9}, qu'on peut attribuer aux différences de précision de calcul entre les logiciels SPSS et SAS.

Calmar 2 signale une convergence imparfaite sur la variable numérique salaire, de l'ordre du centième en valeur absolue. Dans ses éditions, Calmar 2 calcule les écarts absolus entre total réel et total estimé, tandis que g-Calib 2 ne calcule que les écarts relatifs, en s'arrêtant aux différences atteignant 0,1 % de la valeur concernée.

Avec la fonction linéaire tronquée, Calmar 2 donne les mêmes résultats que g-Calib 2 sur l'intervalle [0,97-1,06] et sur les intervalles plus larges. Calmar 2 avec l'option « colin=non » (impliquant une inversion classique de matrice) ne converge pas sur l'intervalle [0,97-1,03] et sur tout intervalle plus strict. Avec l'option « colin=oui » entraînant l'inversion généralisée de la matrice Φ', Calmar 2 se comporte comme g-Calib 2 : il poursuit les itérations en fournissant un résultat imparfait et des rapports de poids tous égaux à l'une des deux bornes. La répartition du nombre d'unités entre borne inférieure et borne supérieure varie entre les deux logiciels.

6.2.2 Calage simple sur la population de ménages

On cale la table échantillon « echup » sur les effectifs de ménages selon les critères ci-dessus.

6.2.2.1 Résultats obtenus avec g-Calib 2 selon les différentes fonctions de calage

- Fonction linéaire : le calage converge en 52 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,96 et 1,08.
- Fonction exponentielle : le calage converge en 4 itérations et 54,5 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,96 et 1,08.
- Fonction logit pour des rapports de poids compris entre :
    - [0,97-1,07] : le calage converge en 7 itérations et 37 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,06] : le calage converge en 7 itérations et 33,5 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,98-1,07] : non convergence, avec le message "LN out of range".
    - [0,98-1,06] : non convergence, avec le message "LN out of range".
- Fonction linéaire tronquée :
    - [0,97-1,06] : le calage converge en 5 itérations et 32 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,07] : le calage converge en 5 itérations et 33,7 secondes, ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,98-1,06] : le calage converge imparfaitement
    - [0,98-1,07] : le calage converge imparfaitement

6.2.2.2 Comparaison avec Calmar 2

Lorsque le calage converge parfaitement, g-Calib 2 et Calmar 2 fournissent des poids qu'on peut considérer égaux pour chaque individu. Les écarts maximums sont compris entre 10^{-10} et 10^{-11}, qu'on peut attribuer aux différences de précision de calcul entre les logiciels SPSS et SAS.

6.2.3 Calage au niveau individu avec contrainte d'égalité des poids dans le ménage

Le modèle de calage est identique au précédent, mais on contraint les poids des individus d'un même ménage à être égaux.

6.2.3.1 Résultats obtenus avec g-Calib 2 selon les différentes fonctions de calage

- Fonction linéaire : le calage converge en 1 minute 4 secondes environ, ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,94 et 1,44.
- Fonction exponentielle : le calage converge en 1 minute 21 secondes environ et 4 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage. Les rapports de poids sont compris entre 0,94 et 1,51.
- Fonction logit pour des rapports de poids compris entre :
    - [0,94-1,5] : le calage converge en 50 secondes environ et 5 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,95-1,4] : le calage converge en 2 minutes 35 secondes et 5 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,96-1,1] : le calage converge en 1 minute 31 secondes et 5 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,1] : non convergence, avec le message "LN out of range"
- Fonction linéaire tronquée :
    - [0,96-1,1] : le calage converge en 48,7 secondes et 5 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,97-1,1] : le calage converge en 52,4 secondes et 7 itérations. Il ne signale pas d'écart entre le total réel et le total estimé par calage.
    - [0,98-1,05] : : le calage converge imparfaitement en 52 secondes environ et 8 itérations, mais sans message d'erreur. Les écarts entre le total réel et le total estimé par calage vont de 0,1% à 5,7 % de la valeur concernée.

6.2.3.2 Comparaison avec Calmar 2

g-Calib 2 utilise la méthode Lemaître-Dufour pour contraindre les poids individuels à l'égalité dans la grappe. En notant *i* l'indice de la grappe (le ménage), *k* celui de l'unité secondaire (l'individu), N_i la taille de la grappe *i* et d_i son poids de sondage, les équations de calage s'écrivent :

$$\sum_{i \in s} d_i^* F\left(\gamma' \bar{X}_i\right) \bar{X}_i = X$$
$$d_i^* = \sum_{k \in s_i} d_k = N_i d_i$$
$$\bar{X}_i = \frac{1}{N_i} \sum_{k \in s_i} x_k$$

Pour dépouiller l'enquête, il faut donc préalablement diviser le poids de calage fourni par g-Calib 2 par le nombre d'individus dans le ménage et utiliser cette nouvelle valeur en poids d'extrapolation. Le fichier de poids créé par g Calib 2 contient une observation par ménage et la taille de la grappe dans une variable appelée *clsize__*.

Calmar 2 réalise le calage directement sur un fichier de grappes constitué par sommation des variables auxiliaires des individus composant la grappe, avec les équations de calage :

$$\sum_{i \in s} d_i F\left(\lambda' X_i\right) X_i = X$$
$$X_i = \sum_{k \in s_i} x_k$$

Il fournit une table de poids avec une observation par individu.

Les poids individuels obtenus dans g-Calib 2 (après division par la taille de la grappe) sont donc sensiblement différents de ceux fournis par Calmar 2, avec des écarts absolus variant de quelques unités à 10^{-5}. Pour plus de 95 % des unités, les écarts sont compris entre 0,1 et 0,001.

Avec la fonction logit, Calmar 2 admet des bornes plus strictes que g-Calib 2 avant de diverger. Calmar 2 converge dans l'intervalle [0,97-1,08] mais pas dans [0,97-1,07].

Avec la fonction linéaire tronquée et les bornes [0,98-1,05], Calmar 2 (option « colin=non ») ne converge pas. g-Calib 2 fournit des poids sans s'interrompre pour non-convergence, alors que les estimateurs des variables de calage ne sont pas justes. Il n'y a pas de message explicite édité.

6.2.4 Calage simultané sur des totaux dans la population des ménages et dans celle des individus

On cale la table « echup » sur le nombre de ménages dans la population selon les critères indiqués dans l'introduction à la section 6 et la table « echant » sur les totaux, dans la population d'individus, des mêmes critères que précédemment. Les individus d'un même ménage ont en sortie des poids identiques.

6.2.4.1 Résultats obtenus avec g-Calib 2 selon les différentes fonctions de calage

- Fonction linéaire : le calage converge en 1 minute 43 secondes et ne signale aucun écart entre le total et son estimateur par calage. Les rapports de poids sont compris entre 0,87 et 1,37.
- Fonction exponentielle : le calage converge en 4 itérations et 1 minute 4 secondes. Il ne signale aucun écart entre le total et son estimateur par calage. Les rapports de poids sont compris entre 0,88 et 1,42.
- Fonction logit :
    - [0,90-1,08] : le calage converge en 5 itérations et 1 minute 9 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
    - [0,93-1,07] : le calage converge en 6 itérations et 1 minute 7 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
    - [0,94-1,07] : non convergence, avec le message "LN out of range"

- Fonction linéaire tronquée :
    - [0,93-1,07] : le calage converge en 6 itérations et 1 minute 15 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
    - [0,94-1,07] : le calage converge en 6 itérations et 1 minute 36 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
    - [0,95-1,05] : le calage converge très imparfaitement en 21 itérations et 1 minute 36 secondes. Les écarts entre le total d'une variable de calage et son estimateur sont compris entre 0,3 % et 5,2 % de la valeur concernée. Les rapports de poids sont tous égaux à l'une des deux bornes.

6.2.4.2 Comparaison avec Calmar 2

g-Calib 2 fournit en sortie une seule table de poids de niveau grappe (ménage), tandis que Calmar 2 fournit deux tables, l'une au niveau grappe et l'autre au niveau unité secondaire.

L'algorithme est le même dans les deux logiciels et conduit aux mêmes poids lorsque le calage converge. Les différences entre les poids individuels sont au maximum compris entre 10^{-8} et 10^{-9}.

Avec la méthode logit, g-Calib 2 refuse plus vite la convergence, tandis que Calmar 2 accepte des bornes plus strictes. Calmar 2 diverge à partir des bornes [0,945-1,055] tandis que g-Calib 2 s'arrête avec les bornes [0,94-1,07].

Avec la fonction linéaire tronquée, g-Calib 2 continue à fournir des poids même en cas d'impossibilité de calage. C'est le cas avec les bornes [0,95-1,05], limites à partir desquelles Calmar 2 (et l'option « colin=non ») diverge.

6.2.5 Calage sur la population d'individus dans chaque strate

On a tiré un second échantillon d'individus (fichier « echantstrat ») selon un sondage en grappes avec stratification des grappes (logements) par zone géographique (variable *tzpiu* à 6 modalités), et sondage aléatoire simple dans les strates. On a obtenu 8 001 ménages et 15 737 individus.

L'échantillon d'individus est calé sur les effectifs de la population de 15 ans ou plus par âge quinquennal et sur la somme des salaires mensuels dans chaque strate.

6.2.5.1 Résultats obtenus avec g-Calib 2 selon les différentes fonctions de calage

g-Calib 2 permet un calage dans l'ensemble des strates en une seule session ou séparément sur une sélection de strates. Dans le premier cas, avec les méthodes logit et linéaire tronquée, les bornes imposées aux rapports de poids sont les mêmes dans toutes les strates. Le deuxième procédé permet de faire varier les intervalles des rapports de poids d'une strate à l'autre. On a donc testé un calage global sur les 6 strates puis un calage séparé par strate avec les méthodes bornées.

- Fonction linéaire : le calage converge en 2 minutes 10 secondes et ne signale aucun écart entre les totaux réels et les totaux estimés. Les rapports de poids sont compris entre les bornes suivantes :

Strate	Minimum	Maximum
1	0,76	1,28
2	0,81	1,12
3	0,76	1,18
4	0,89	1,10
5	0,82	1,11
6	0,88	1,32
Ensemble	0,76	1,32

- Fonction exponentielle : : le calage converge en 4 itérations dans chaque strate et 1 minute 52 secondes. Il ne signale aucun écart entre les totaux réels et les totaux estimés. Les rapports de poids sont compris entre les bornes suivantes :

Strate	Minimum	Maximum
1	0,77	1,29
2	0,81	1,12
3	0,77	1,19
4	0,89	1,11
5	0,84	1,11
6	0,88	1,37
Ensemble	0,77	1,37

- Fonction logit et calage unique sur toutes les strates avec les bornes :

  [0,77-1,27] : le calage converge en 1 minute 38 secondes.

Strate	Nombre d'itérations	Ecart entre total vrai et total estimé
1	8	0,2 % sur la variable *j3*
2	5	aucun
3	4	aucun
4	4	aucun
5	4	aucun
6	4	aucun

  [0,77-1,26] : le calage ne converge pas. Message « LN out of range ».

  [0,78-1,27] : le calage ne converge pas. Message « LN out of range ».

- Fonction logit et calage séparé par strate

<u>Strate 1</u>

  - [0,77-1,28] : le calage converge en 7 itérations et 32,6 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
  - [0,77-1,27] : le calage converge en 8 itérations et 1 minute. Il ne signale aucun écart entre le total et son estimateur par calage.
  - [0,77-1,26] : le calage ne converge pas. Message « LN out of range ».
  - [0,78-1,27] : le calage ne converge pas. Message « LN out of range ».

<u>Strate 2</u>

  - [0,81-1,11] : le calage converge en 7 itérations et 39,2 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
  - [0,81-1,10] : le calage converge imparfaitement en 8 itérations et 41,1 secondes. Il y a un écart de 3 % entre le total de la variable *ad4* et son estimateur par calage.

- [0,81-1,09] : le calage ne converge pas. Message « LN out of range ».
- [0,80-1,10] : le calage ne converge pas. Message « LN out of range ».

Strate 3

- [0,76-1,18] : le calage converge en 7 itérations et 29,6 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,76-1,17] : le calage ne converge pas. Message « LN out of range ».
- [0,90-1,18] : le calage converge en 7 itérations et 30 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,91-1,18] : le calage ne converge pas. Message « LN out of range ».

Strate 4

- [0,89-1,09] : le calage converge en 8 itérations et 31,1 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,89-1,08] : le calage converge imparfaitement en 5 itérations et 31 secondes. On a un écart de 0,86 % entre le total d'une variable catégorielle (*v2*) et son estimateur par calage.
- [0,89-1,07] : le calage ne converge pas. Message « LN out of range ».
- [0,90-1,08] : le calage ne converge pas. Message « LN out of range ».

Strate 5

- [0,84-1,10] : le calage converge en 7 itérations et 39,2 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,84-1,09] : le calage ne converge pas. Message « LN out of range ».
- [0,92-1,10] : le calage converge en 7 itérations et 38 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,93-1,10] : le calage ne converge pas. Message « LN out of range ».

Strate 6

- [0,88-1,10] : le calage converge en 7 itérations et 31,4 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,88-1,09] : le calage ne converge pas. Message « LN out of range ».
- [0,90-1,10] : le calage converge en 8 itérations et 32,5 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,91-1,10] : le calage ne converge pas. Message « LN out of range ».

- Fonction linéaire tronquée et calage unique sur toutes les strates avec les bornes :

[0,77-1,28] : le calage converge en 1 minute 32 secondes.

Strate	Nombre d'itérations	Ecart entre total vrai et total estimé
1	3	aucun
2	2	aucun
3	3	aucun
4	2	aucun
5	2	Aucun
6	3	Aucun

[0,77-1,27] : le calage converge en 1 minute 34 secondes, imparfaitement en strate 1.

Strate	Nombre d'itérations	Ecart entre total vrai et total estimé
1	5	0,2 % sur la variable *j3*
2	2	aucun
3	3	aucun
4	2	aucun
5	2	aucun
6	3	aucun

[0,80-1,28] : le calage converge en 1 minute 32 secondes environ.

Strate	Nombre d'itérations	Ecart entre total vrai et total estimé
1	4	aucun
2	2	aucun
3	3	aucun
4	2	aucun
5	2	aucun
6	3	aucun

- Fonction linéaire tronquée et calage séparé par strate

Strate 1

- [0,76-1,28] : le calage converge en 3 itérations et 40,6 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,77-1,28] : le calage converge en 3 itérations et 28 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,77-1,27] : le calage converge imparfaitement.
- [0,76-1,27] : le calage converge imparfaitement.

Strate 2

- [0,81-1,11] : le calage converge en 3 itérations et 32,8 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,81-1,10] : le calage converge imparfaitement.
- [0,82-1,11] : le calage converge imparfaitement.

Strate 3

- [0,77-1,18] : le calage converge en 3 itérations et 31,4 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,77-1,17] : le calage converge imparfaitement.
- [0,90-1,18] : le calage converge en 5 itérations et 31,5 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,91-1,18] : le calage converge imparfaitement.

Strate 4

- [0,89-1,09] : le calage converge en 3 itérations et 31,9 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,89-1,08] : le calage converge imparfaitement.
- [0,90-1,09] : le calage converge en 4 itérations et 27,9 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,91-1,09] : le calage converge imparfaitement.

Strate 5

- [0,82-1,10] : le calage converge en 5 itérations et 39,9 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,82-1,09] : le calage converge imparfaitement.
- [0,93-1,10] : le calage converge en 5 itérations et 36,4 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,94-1,10] : le calage converge imparfaitement.

Strate 6

- [0,88-1,10] : le calage converge en 5 itérations et 28,8 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,88-1,09] : le calage converge imparfaitement.
- [0,90-1,10] : le calage converge en 7 itérations et 32,1 secondes. Il ne signale aucun écart entre le total et son estimateur par calage.
- [0,91-1,10] : le calage converge imparfaitement.

6.3 Contrôle des paramètres dans g-Calib 2

A condition d'utiliser l'interface de saisie, une logique est imposée aux paramètres spécifiés par l'utilisateur :

- on ne peut spécifier que le nom d'un fichier existant, puisque celui-ci est sélectionné dans l'arborescence via Explorer
- dans la table des variables nécessaires au calage, on ne peut spécifier que le nom d'une variable existante, puisque celle-ci est sélectionnée dans le dictionnaire du fichier
- on ne peut faire de calage simultané que si l'on a spécifié une table « individus » et une table « grappes » dans les champs de saisie
- en cas de calage simultané, les identifiants de la grappe et de l'unité secondaire sont nécessairement des variables distinctes
- on ne peut pas faire de calage simultané si la variable spécifiée en identifiant de la grappe n'est pas la même dans les deux tables échantillon
- la fonction de calage est choisie dans la liste des fonctions proposées
- le paramètre ECHELLE est forcément numérique : tout autre caractère qu'un chiffre n'est pas accepté
- on ne peut progresser dans la saisie du modèle de calage et dans l'exécution que lorsque tous les paramètres obligatoires ont été spécifiés. Il faut cependant se reporter à la documentation pour savoir quels sont les paramètres obligatoires, le défaut de saisie n'entraîne généralement aucun message en clair.

En dehors de cela, g-Calib 2 ne semble pas comporter de programme systématique de contrôle de la cohérence des paramètres de l'utilisateur. Les bugs

résultant d'une erreur de spécification entraînent des messages d'erreur de syntaxe dans le fichier de résultats SPSS, mais il faut connaître le programme pour les élucider. L'application n'affiche pas de messages en clair permettant d'identifier l'erreur de l'utilisateur.

Par ailleurs, l'interface de saisie ne permet pas la correction du nom d'une variable dans la liste des « *special variables* », sauf par remplacement par une autre variable non déjà utilisée dans le dictionnaire du fichier.

Les erreurs suivantes ont été testées :

- Omission du nom de la table contenant les poids de calage : si on conserve l'option par défaut « none » de l'interface de saisie, la table est créée sous le nom *none*. Si on laisse le champ à blanc, le programme se plante, sans message programmé.
- Omission de la borne inférieure des rapports de poids, avec la fonction logit (la valeur par défaut est mise à blanc) : le programme se plante, sans message programmé.
- Poids initiaux négatifs dans le fichier d'enquête : le calage se réalise. Il n'y a aucun message dans le fichier de résultats, en particulier aucune observation éliminée du calage. Avec la fonction linéaire, le tableau récapitulatif des itérations ne donne aucun poids négatif en sortie, alors que dans la table des poids, le poids de calage est égal au poids initial (donc négatif) pour les observations à poids initial négatif.
- Spécification de l'identifiant de la grappe au lieu de celui de l'unité secondaire, dans un calage simple sur la population des unités secondaires : le calage se réalise, les poids sont les mêmes que ceux obtenus en spécifiant le bon identifiant, aucun message.
- Inversion des identifiants grappe et unité secondaire dans un calage en grappes sans données de niveau grappe : le calage se réalise et donne des poids égaux à celui d'un calage simple. Les grappes sont traitées comme étant toutes de taille 1.
- Calage en grappes sans données sur les grappes, avec des poids initiaux différents entre unités d'une même grappe : avec la fonction linéaire, le

calage se réalise et donne des poids distincts de ceux obtenus avec une spécification correcte. Aucun message.

- Variable de pondération initiale non numérique : le programme se plante, avec une erreur de syntaxe SPSS, mais pas de message programmé.
- Spécification d'une variable de calage qui n'existe pas dans le fichier d'enquête : le programme se plante avec une erreur de syntaxe SPSS, mais pas de message programmé.

7 Evaluation de Calmar 2

7.1 Aspects généraux de Calmar 2

7.1.1 Facilité d'utilisation

7.1.1.1 Logiciel

La macro Calmar 2 est programmée en langage SAS ; elle est compatible avec la version 8 de SAS. L'exécution se fait depuis la fenêtre Editor de SAS, en appelant en début de programme la version compilée de la macro, ou de façon automatique si la macro est intégrée à la version de SAS disponible (comme c'est le cas à l'INSEE et à l'ENSAI).

7.1.1.2 Saisie des paramètres

La saisie des paramètres se fait dans la fenêtre Editor de SAS, lors de l'appel de la macro, de la façon suivante :

%CALMAR 2 (PARAM1 = valeur1 , PARAM2 = valeur2 , …)

où valeur1 est la valeur que l'on souhaite donner au paramètre PARAM1, …

Des connaissances SAS de base sont nécessaires pour l'utiliser :

- Savoir allouer une librairie (celle où vont se trouver les bases utilisées pour le calage)
- Savoir manipuler et visualiser des tables SAS (la table des répondants et celle des marges)

La macro est d'un maniement simple pour un utilisateur habituel de SAS : un certain nombre de paramètres ont des valeurs par défaut, ce qui limite les saisies à effectuer par l'utilisateur. De même, certains paramètres ne sont à renseigner que pour des opérations de calage complexes (comme le calage à plusieurs niveaux ou le traitement de la non-réponse) ce qui limite les pré-requis nécessaires sur le calage et ses applications.

7.1.1.3 Sauvegarde

Après écriture, un programme de calage peut facilement être enregistré et réutilisé ou modifié.

La table des poids de calage est stockée dans le répertoire de travail spécifié dans les paramètres de la macro.

7.1.2 Input

Ce sont de façon obligatoire :

- La table des données (paramètre DATAMEN), qui contient au moins les individus de l'échantillon, ainsi que pour chacun :
  - la valeur correspondante de chaque variable auxiliaire de calage
  - la valeur de la pondération initiale
- La table des marges (paramètre MARMEN), qui spécifie :
  - les variables auxiliaires utilisées pour le calage
  - leur nombre de modalités (0 pour une variable numérique)
  - les valeurs des marges

Une phase de préparation des données est nécessaire :

- La table des marges contient autant d'observations que de variables utilisées pour le calage ; pour chacune d'elles sont spécifiées son nom, son nombre de modalités (0 pour une variable numérique) et les valeurs des marges de ces modalités
- Pas de contrainte sur les variables de calage catégorielles (la macro les recodifie dans un codage numérique séquentiel)
- Les variables de calage quantitatives doivent être de type numérique (ce point est vérifié par Calmar 2)
- Les variables de pondération doivent être numériques (ce point est vérifié par Calmar 2 sur demande avec le paramètre CONT)
- Les variables de calage ne doivent pas comporter de valeurs manquantes ; si c'est le cas, l'observation correspondante est éliminée par Calmar 2

- Les variables de pondération ne doivent pas comporter de valeurs manquantes, négatives ou nulles ; si c'est le cas, l'observation correspondante est éliminée par Calmar 2.

Dans le cas d'un sondage par grappes avec calage simultané, on doit également spécifier :

- la table des unités secondaires (paramètre DATAIND). Elle doit contenir :
  - les variables de calage
  - une variable identifiant l'unité secondaire (paramètre IDENT2)
  - la variable identifiant l'unité primaire à laquelle se rattache l'unité secondaire (paramètre IDENT).
- La table des marges (paramètre MARIND), qui spécifie :
  - les variables auxiliaires utilisées pour le calage
  - leur nombre de modalités (0 pour une variable numérique)
  - les valeurs des marges

Dans le cas d'un sondage à deux degrés avec trois niveaux d'observation, on doit en plus des tables précédentes (table DATAMEN des unités primaires sélectionnées et table des marges associées MARMEN, table DATAIND des unités secondaires correspondantes et table des marges associées MARIND), spécifier :

- la table des unités secondaires sélectionnées au $2^{ème}$ degré de tirage (paramètre DATAKISH). Elle doit contenir :
  - les variables de calage
  - la variable identifiant l'unité secondaire, variable déjà présente dans DATAIND (paramètre IDENT2)
  - la variable identifiant l'unité primaire à laquelle elle se rattache, variable déjà présente dans DATAIND et DATAMEN (paramètre IDENT)
  - la variable de pondération de l'unité secondaire dans l'unité primaire, i.e. l'inverse de sa probabilité conditionnelle de tirage de degré 2. Elle doit être de type numérique.
- La table des marges (paramètre MARKISH), qui spécifie :
  - les variables auxiliaires utilisées pour le calage
  - leur nombre de modalités (0 pour une variable numérique)
  - les valeurs des marges

7.1.3 Fonctionnalités

7.1.3.1 Plan de sondage

La macro Calmar 2 permet de calculer des poids redressés pour une enquête réalisée selon un des plans de sondage suivants :

- Sondage à un degré d'observation : on cale l'échantillon sur les totaux dans la population de variables auxiliaires numériques ou catégorielles
- Sondage par grappes : on cale l'échantillon sur
  - (Facultatif) les totaux dans la population des unités primaires de variables auxiliaires numériques ou catégorielles
  - les totaux dans la population des unités secondaires de variables auxiliaires numériques ou catégorielles, en assurant l'égalité des poids pour les unités d'une même grappe

  Si on n'utilise pas de données sur les grappes, il faut obligatoirement renseigner les paramètres suivants :
  - EGALPOI=OUI (impose l'égalité des poids pour les unités secondaires d'une même grappe)
  - POPMEN (nombre d'unités primaires dans la population)
- Sondage à deux degrés : on cale l'échantillon sur
  - les totaux dans la population des unités primaires de variables auxiliaires numériques ou catégorielles
  - (Facultatif) les totaux dans la population des unités secondaires de variables auxiliaires numériques ou catégorielles
  - les totaux, dans la population des unités secondaires éligibles au second degré, de variables auxiliaires numériques ou catégorielles

7.1.3.2 Modèle de redressement

Calmar 2 permet d'effectuer un redressement par calage sur des variables auxiliaires, selon la méthode mise au point par Jean-Claude Deville et Carl-Eric Särndal (1992). Elle comprend les 4 fonctions de calage proposées dans l'article fondateur :

- Fonction linéaire : $F(u) = 1 + u$
- Fonction exponentielle : $F(u) = e^u$
- Fonction linéaire tronquée : $F(u) = 1 + u$ pour $L \le F\left(u\right) \le U$
- Fonction logit : $F(u) = \dfrac{L(U-1) + U(1-L)e^{Au}}{(U-1) + (1-L)e^{Au}}$ avec $A = \dfrac{U-L}{(U-1)(1-L)}$

ainsi qu'une proposée plus récemment :

- Fonction sinus hyperbolique :

$$F(u) = \frac{1}{2}\left[\frac{1}{\alpha} Log\left(2\alpha\, u + \sqrt{4\alpha^2 u^2 + 1}\right) + \sqrt{\frac{1}{\alpha^2}\left(Log\left(2\alpha\, u + \sqrt{4\alpha^2 u^2 + 1}\right)\right)^2 + 4}\right]$$

avec $\alpha > 0$

7.1.3.3 Traitement de la non réponse

Calmar 2 permet également d'utiliser la technique de calage généralisé (Deville, 2002) pour faire un redressement de la non-réponse par calage à l'aide de variables instrumentales, connues uniquement sur l'échantillon de répondants. Si on souhaite utiliser cette technique :

- Le paramètre NONREP doit être codé à OUI
- Les différentes tables de marges (MARMEN et éventuellement MARIND et MARKISH) doivent contenir :
  - comme observations : les variables instrumentales, en plus des variables de calage
  - un indicateur du type de variable (variable instrumentale ou variable de calage)
- Les différentes tables de données (DATAMEN et éventuellement DATAIND et DATAKISH) doivent contenir les variables instrumentales
- Le vecteur des variables instrumentales doit avoir la même dimension que le vecteur des variables de calage, et le même nombre de variables catégorielles

7.1.4 Output

7.1.4.1 Contrôles effectués

La macro permet d'effectuer un nombre important de contrôles de cohérence sur les tables, les variables en entrée et les paramètres. L'étendue des contrôles est gérée par le paramètre CONT.

7.1.4.2 Fichier des poids de calage

La macro Calmar 2 fournit en sortie des tables SAS contenant les poids de calage.

Pour un calage simple correspondant à un degré d'observation, on obtient une table DATAPOI contenant les observations non éliminées de la table DATAMEN ainsi que la(les) variable(s) :

- Pondération finale
- (Eventuellement) Identifiant de chaque unité

Dans le cas d'un sondage par grappes avec calage simultané, on obtient en plus de la table précédente une table DATAPOI2 contenant en observations les unités secondaires non éliminées de la table DATAIND ainsi que les variables :

- Pondération finale associée aux unités secondaires
- Identifiant de chaque unité secondaire
- Identifiant de la grappe associée à chaque unité secondaire

Dans le cas d'un sondage à deux degrés avec trois niveaux d'observation, on obtient en plus des tables précédentes une table DATAPOI3 contenant en observations les unités secondaires non éliminées de la table DATAKISH ainsi que les variables :

- Pondération finale associée aux unités secondaires échantillonnées au second degré
- Identifiant de chaque unité secondaire
- Identifiant de la grappe associée à chaque unité secondaire

Si le paramètre MISAJOUR vaut NON, la macro réinitialise chacune de ces tables à chaque itération. Si ce paramètre vaut OUI, les variables de pondération sont ajoutées dans les différentes tables, ce qui permet d'empiler des jeux de pondération obtenus selon des méthodes différentes et de les comparer plus facilement.

7.1.4.3 Editions

Elle donne également un bilan des différentes étapes du calage. L'affichage des sorties suivantes est modulé par le paramètre EDITION :

- Tableau contenant les valeurs des paramètres donnés à la macro
- Tableau comparant les marges dans la population à leurs estimations avec les pondérations initiales
- Tableau comparant les marges dans la population à leurs estimations avec les pondérations après calage
- Tableau donnant la valeur du critère d'arrêt de l'algorithme itératif, et le nombre de poids négatifs après chaque itération
- Tableau donnant le vecteur de Lagrange après chaque itération
- Un bilan du calage (nom de la table en entrée, nombre d'observations et nombre d'observations éliminées, …)

D'autres sorties peuvent être ajoutées :

- Si EDITPOI=OUI, un tableau donne les valeurs des différents rapports de poids obtenus
- Si STAT=OUI, on obtient les sorties d'une PROC UNIVARIATE sur les variables rapport de poids et pondération finale, ainsi qu'un tableau donnant le rapport de poids moyen par modalité de chaque variable catégorielle
- Si CONTPOI=OUI, on obtient les sorties d'une PROC CONTENTS sur la table contenant les poids finaux

7.1.5 Problèmes

Ce sont des problèmes inhérents à la technique du calage :

- Le calage peut ne pas être réalisé
- L'algorithme peut ne pas converger avec le nombre d'itérations maximal fixé
- L'algorithme peut ne pas converger

Ces problèmes peuvent notamment se produire en cas de redressement pour non-réponse, si les variables de calage et les variables instrumentales ne sont pas assez corrélées entre elles, ou en cas d'utilisation d'une fonction de calage tronquée, si les bornes imposées aux rapports de poids sont trop contraignantes.

7.1.6 Documentation

Le manuel de Calmar 2 est très clair et illustré de nombreux exemples. Il faudrait cependant y rajouter un manuel d'utilisateur de Calmar 2_Guide, la version interactive de la macro.

Cependant, ni le logiciel ni son manuel d'utilisation ne sont encore disponibles en ligne.

7.2 Résultats obtenus avec Calmar 2

Les limites du logiciel sont essentiellement celles de SAS. A noter qu'une variable de calage catégorielle et de type caractère ne doit pas avoir plus de 999 modalités. Les résultats des tests effectués sont :

7.2.1 Calage simple au niveau individu

On va caler la table « echant.sas » à l'aide de l'information disponible au niveau des individus (effectifs de la population des individus pour la pyramide des âges quinquennaux et le salaire total).

- Avec la fonction linéaire, on obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,96 et 1,50.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,96 et 1,60.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,96-1,50] : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,96-1,30] : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,96-1,10] : le calage est réalisé en 4 itérations. On observe des différences de l'ordre du centième entre les marges réelles et estimées des variables catégorielles, et une différence de l'ordre de l'unité pour la variable Salaire
  - [0,96-1,05] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,96-1,30] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,96-1,10] : le calage est réalisé en 4 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,96-1,05] : le calage ne peut être réalisé
- Avec la fonction sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,96 et 1,18
  - 100 : le calage est réalisé en 9 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,96 et 1,06
  - 200 : le calage ne peut être réalisé (l'algorithme ne converge pas)

7.2.2 Calage simple au niveau ménage

On va caler la table « echup.sas » à l'aide des variables auxiliaires disponibles au niveau des ménages.

- Avec la fonction linéaire, le calage est réalisé exactement. Les rapports de poids sont compris entre 0,96 et 1,08.
- Avec la fonction exponentielle, le calage est réalisé exactement, en 3 itérations. Les rapports de poids sont compris entre 0,96 et 1,08.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,94-1,1] : le calage est exactement réalisé, en 4 itérations
  - [0,95-1,09] : le calage est réalisé en 4 itérations. On observe une différence inférieure au centième entre un total réel et un total estimé d'une variable de calage.
  - [0,96-1,08] : le calage est réalisé en 4 itérations. On observe des différences inférieures au centième entre des totaux réels et estimés de variables de calage.
  - [0,97-1,07] : le calage est exactement réalisé en 7 itérations.
  - [0,98-1,06] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,94-1,1] : le calage est exactement réalisé, en 2 itérations
  - [0,95-1,09] : le calage est exactement réalisé, en 2 itérations
  - [0,96-1,08] : le calage est exactement réalisé, en 3 itérations
  - [0,97-1,07] : le calage est exactement réalisé, en 5 itérations.
  - [0,98-1,06] : le calage ne peut être réalisé
- Avec la fonction sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est exactement réalisé, en 4 itérations. On obtient des rapports de poids compris entre 0,96 et 1,07
  - 100 : le calage est exactement réalisé, en 8 itérations. On observe des différences inférieures au centième entre des totaux réels et estimés de variables de calage. On obtient des rapports de poids compris entre 0,97 et 1,05
  - 200 : le calage ne peut être réalisé

7.2.3 Calage au niveau individu avec contrainte d'égalité des poids dans le ménage

On procède comme en section 7.2.1, mais en imposant l'égalité des poids pour les individus d'un même ménage.

- Avec la fonction linéaire, on obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,90 et 1,42.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,90 et 1,49.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,90-1,49] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du dixième entre le total réel et estimé de la variable Salaire
  - [0,90-1,30] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 4 itérations. On observe une différence inférieure à l'unité entre le total réel et estimé de la variable Salaire
  - [0,95-1,05] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,90-1,42] : le calage est réalisé en 2 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,30] : le calage est réalisé en 3 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 4 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,95-1,05] : le calage ne peut être réalisé
- Avec la méthode du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 4 itérations. On observe une différence inférieure au dixième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,92 et 1,17

- 50 : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,94 et 1,08
- 100 : le calage ne peut être réalisé (l'algorithme ne converge pas)

7.2.4 Calage simultané au niveau ménage et individu (sondage par grappes)

On cale la table « echup.sas » sur les variables auxiliaires disponibles au niveau ménage, et la table « echant.sas » sur les variables auxiliaires disponibles au niveau individu, en imposant l'égalité des poids pour les individus d'un même ménage.

- Avec la fonction linéaire, on obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,87 et 1,37.
- Avec la fonction exponentielle, le calage est réalisé en 3 itérations. On obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,88 et 1,42.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,88-1,42] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 15 secondes et 4 itérations. On observe des différences de l'ordre du centième entre les marges réelles et estimées des variables catégorielles, et une différence supérieure à l'unité pour la variable Salaire
  - [0,96-1,06] : le calage est réalisé en 11 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,97-1,05] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,87-1,37] : le calage est réalisé en 2 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,20] : le calage est réalisé en 3 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire

  - [0,96-1,06] : le calage est réalisé en 11 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,97-1,06] : le calage ne peut être réalisé

- Avec la méthode du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 4 itérations. On observe une différence de l'ordre de l'unité entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,91 et 1,17
  - 50 : le calage est réalisé en 8 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,93 et 1,09
  - 100 : le calage ne peut être réalisé (l'algorithme ne converge pas)

7.2.5 Calage simultané dans un sondage à deux degrés avec deux niveaux d'observation

On cale la table « echup.sas » sur les variables auxiliaires disponibles au niveau ménage, et la table « echant2d.sas » sur les variables auxiliaires disponibles au niveau individu.

- Avec la fonction linéaire, on obtient une différence inférieure au dixième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,86 et 1,41.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence inférieure au dixième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,87 et 1,47.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,87-1,47] : le calage est réalisé en 4 itérations. On observe une différence inférieure au dixième entre le total réel et estimé de la variable Salaire
  - [0,90-1,30] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 5 itérations. On observe une différence inférieure au dixième entre le total réel et estimé de la variable Salaire
  - [0,95-1,05] : le calage ne peut être réalisé

- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,86-1,41] : le calage est réalisé en 2 itérations. On observe une différence inférieure à l'unité entre le total réel et estimé de la variable Salaire
  - [0,90-1,30] : le calage est réalisé en 3 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,95-1,05] : le calage ne peut être réalisé
- Avec la méthode du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 4 itérations. On observe une différence inférieure à l'unité entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,90 et 1,18
  - 50 : le calage est réalisé en 8 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,93 et 1,09
  - 100 : le calage ne peut être réalisé (l'algorithme ne converge pas)

7.2.6 Calage sur la population d'individus dans chaque strate

Strate 1

- Avec la fonction linéaire, on obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,76 et 1,28.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,77 et 1,29.
- Avec la méthode logit, pour des rapports de poids compris entre :
  - [0,77-1,29] : le calage est réalisé en 6 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,77-1,28] : le calage est réalisé en 6 itérations. On observe une différence inférieure à l'unité entre le total réel et estimé de la variable Salaire
  - [0,77-1,27] : le calage ne peut être réalisé

- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,76-1,28] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,76-1,27] : le calage ne peut être réalisé
- Avec la fonction sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,82 et 1,27
  - 40 : le calage est réalisé en 12 itérations. On observe une différence de plusieurs dizaines d'unités entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,83 et 1,27
  - 50° : le calage ne peut être réalisé

Strate 2

- Avec la fonction linéaire, on obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,81 et 1,12.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,81 et 1,12.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,81-1,12] : le calage est réalisé en 6 itérations. On observe une différence de l'ordre du dixième entre le total réel et estimé de la variable Salaire
  - [0,81-1,11] : le calage est réalisé en 6 itérations. On observe une différence de l'ordre du dixième entre le total réel et estimé de la variable Salaire
  - [0,81-1,10] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,81-1,12] : le calage est réalisé en 2 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,81-1,11] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,81-1,10] : le calage ne peut être réalisé

- Avec la fonction du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 6 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,81 et 1,11
  - 30 : le calage est réalisé en 10 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,81 et 1,10
  - 40 : le calage ne peut être réalisé

Strate 3

- Avec la fonction linéaire, on obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,76 et 1,18.
- Avec la fonction exponentielle, le calage est réalisé en 4 itérations. On obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,77 et 1,19.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,90-1,18] : le calage est réalisé en 6 itérations. On observe une différence de l'ordre du dixième entre le total réel et estimé de la variable Salaire
  - [0,76-1,17] : le calage est réalisé en 6 itérations. On observe une différence de plusieurs dizaines d'unités entre le total réel et estimé de la variable Salaire
  - [0,76-1,16] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,77-1,19] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,18] : le calage est réalisé en 5 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,77-1,17] : le calage ne peut être réalisé
- Avec la fonction du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 6 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,85 et 1,18

- 50 : le calage est réalisé en 11 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,90 et 1,17
- 100 : le calage ne peut être réalisé

Strate 4

- Avec la fonction linéaire, on obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,89 et 1,10.
- Avec la fonction exponentielle, le calage est réalisé en 3 itérations. On obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,89 et 1,11.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,89-1,11] : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,89-1,09] : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,89-1,08] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,89-1,10] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,89-1,09] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,89-1,08] : le calage ne peut être réalisé
- Avec la fonction du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 5 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,90 et 1,10
  - 50 : le calage est réalisé en 9 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,90 et 1,09
  - 100° : le calage ne peut être réalisé

Strate 5

- Avec la fonction linéaire, on obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,82 et 1,11.
- Avec la fonction exponentielle, le calage est réalisé en 3 itérations. On obtient une différence inférieure au centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,84 et 1,11.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,93-1,11] : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du dixième entre le total réel et estimé de la variable Salaire
  - [0,93-1,10] : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,93-1,09] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,93-1,11] : le calage est réalisé en 4 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,93-1,10] : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire
  - [0,93-1,09] : le calage ne peut être réalisé
- Avec la fonction du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 5 itérations. On observe une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,89 et 1,10
  - 50 : le calage est réalisé en 8 itérations. On observe une différence de l'ordre de l'unité entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,94 et 1,10
  - 100° : le calage ne peut être réalisé

Strate 6

- Avec la fonction linéaire, on obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,88 et 1,32.
- Avec la fonction exponentielle, le calage est réalisé en 3 itérations. On obtient une différence de l'ordre du centième entre le total réel et estimé de la variable Salaire. Les rapports de poids sont compris entre 0,88 et 1,37.
- Avec la fonction logit, pour des rapports de poids compris entre :
  - [0,88-1,30] : le calage est réalisé en 5 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,88-1,10] : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,09] : le calage ne peut être réalisé
- Avec la fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,88-1,32] : le calage est réalisé en 3 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,88-1,20] : le calage est réalisé en 5 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,10] : le calage est réalisé en 7 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire
  - [0,90-1,09] : le calage ne peut être réalisé
- Avec la fonction du sinus hyperbolique, pour un coefficient égal à :
  - 10 : le calage est réalisé en 5 itérations. On observe une différence inférieure au centième entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,89 et 1,18
  - 50 : le calage est réalisé en 9 itérations. On observe une différence de l'ordre de l'unité entre le total réel et estimé de la variable Salaire. On obtient des rapports de poids compris entre 0,90 et 1,11
  - 100° : le calage ne peut être réalisé

8 Evaluation de Bascula 4.0

8.1 Aspects généraux de bascula 4.0

8.1.1 Facilité d'utilisation

Bascula est une composante de Blaise, qui est un système de traitement de données d'enquêtes assisté par ordinateur. Bascula se présente sous forme d'outil interactif, articulé sur une configuration contenant l'information nécessaire pour exécuter un calage. Celle-ci peut être lancée en partant de zéro et en entrant progressivement toutes les informations nécessaires. Il est également possible de sélectionner une configuration déjà existante dans Bascula afin de poursuivre le traitement. Bascula est aussi disponible sous forme de composante logicielle, appelée Interface de programme d'application ou *API* (*Application Programming Interface*) de Bascula.

Le fait que Bascula soit axé sur Blaise peut engendrer un travail non négligeable au niveau de la préparation de la base de données avant de pouvoir exécuter un calage. En effet, Bascula ne gère que des bases de données générées par Blaise ou sous format *Ascii*. Le problème se pose lorsque la base de données n'est pas sous format Blaise, ce qui est souvent le cas en pratique. Le seul moyen qui reste pour récupérer une telle base de données est de l'enregistrer sous format *Ascii*. Cependant, le format *Ascii* ne permet pas de sauvegarder les libellés et les formats des variables. Ainsi, avant de pouvoir utiliser une base de données sous Bascula, il faut redéfinir les libellées et les formats des variables, ce qui n'est souvent pas une simple tâche surtout quand on dispose d'une base de données contenant un grand nombre de variables. Ce problème ne se pose pas avec Calmar 2 et g-Calib 2 qui sont respectivement des macros de SAS et SPSS (qui offrent un large choix de formats de fichiers).

8.1.2 Input

Pour pouvoir exécuter une session de pondération avec Bascula 4.0, on doit disposer des fichiers suivants :

- La table des données, qui contient pour chaque individu de l'échantillon, la valeur correspondante de chaque variable auxiliaire et le poids de sondage initial, la variable identifiant des grappes (pour le calage par grappe). De préférence, la base de données doit contenir au moins deux variables servant à sauvegarder les poids finaux et les rapports de poids.
- La table des marges contenant les totaux sur la population des variables auxiliaires utilisées pour le calage. Ces totaux peuvent être également saisis manuellement sans recourir à cette table de marges.

Une phase de préparation des données est nécessaire :

- Si la base de données n'est pas sous format Blaise, il faut redéfinir les libellés, type et format des variables de la base de données.
- L'identifiant des grappes doit être de type numérique *integer* (ayant des valeurs entières).
- Les variables de calage quantitatives doivent être déclarées sous le format *integer* ou *real*.
- Les variables de calage catégorielles doivent être déclarées sous le format *enumerated*.
- La variable contenant les valeurs des probabilités d'inclusion doit être de type *integer* ou *real*. Cette variable ne doit pas comporter de valeurs manquantes ou négatives ; si c'est le cas, la session de pondération ne peut être réalisée.
- Les variables servant à sauvegarder les poids finaux et les rapports de poids doivent être de type *real*.
- Les variables de calage ne doivent pas comporter de valeurs manquantes ; si c'est le cas, la session de pondération ne peut être réalisée.
- Pour les variables de type *real* comportant des décimales, il faut utiliser « . » au lieu de « , » pour séparer la partie entière des décimaux.

8.1.3 Fonctionnalités

Dans Bascula 4.0, le calcul des nouveaux poids w_k se fait en utilisant quatre méthodes de redressement : la *poststratification*, *l'estimation par ratio*, la *pondération linéaire*, et la *pondération multiplicative*.

Pour la méthode de redressement par *pondération linéaire*, les poids w_k sont calculés en utilisant l'estimateur par la régression ce qui revient à utiliser la méthode de *calage linéaire* implémentée dans g-Calib 2 et Calmar 2. Les méthodes de redressement par la *poststratification* et par l'*estimation par ratio* ne sont que des cas particuliers de la méthode de *calage linéaire*. La *poststratification* correspond au cas d'un calage sur une seule variable catégorielle et l'*estimation par le ratio* correspond au cas où on redresse sur une seule variable quantitative repondérée par $\sigma_k^2 = 1/x_k$.

Comme la méthode de redressement par la *pondération linéaire* peut fournir de poids w_k négatifs, Bascula 4.0 permet de définir pour le rapport de poids une borne inférieure L et une borne supérieure U $\left(\text{avec } 0 < L < 1 < U\right)$, mais l'algorithme est différent de celui qui est utilisé dans Calmar 2 et g-Calib 2.

Enfin, pour la méthode par la *pondération multiplicative*, les poids w_k sont calculés en utilisant un algorithme classique de redressement proposé par Deming et Stefan (1940) appelé *iterative proportional fitting* (IPF). Cet algorithme n'est qu'un cas particulier de la méthode de calage par *raking ratio* implémentée dans g-Calib 2 et Calmar 2 correspondant au calage utilisant la fonction exponentielle avec des variables auxiliaires catégorielles. Pour le calage par grappe, Bascula utilise la méthode de redressement par la *régression linéaire* (*linear consistent weighting*) qui permet de calculer les nouveaux w_k tout en contraignant l'égalité des poids pour les individus d'une même grappe.

Sous Bascula 4.0, le choix de la méthode de redressement dépend du type des variables de calage utilisées dans le redressement. Souvent, les possibilités du choix offertes par Bascula sont limitées :

- Si le modèle de pondération comporte une seule variable de calage de type catégorielle, le choix se limite au redressement par *poststratification*.
- Si le modèle de pondération comporte une seule variable de calage de type quantitative, le programme impose de choisir entre le redressement par l'*estimation par ratio* et le redressement par la régression.
- Si le modèle de pondération comporte plus d'une seule variable de calage qui sont toutes de type catégoriel, le programme impose de choisir entre le redressement par la *régression linéaire* et le redressement par *raking ratio*.
- Pour le calage par grappe, le programme impose d'utiliser le redressement par la *régression linéaire* (*linear consistent weighting*).

Pour les autres types de variables, Bascula impose d'utiliser le redressement par la *pondération linéaire*. L'option qui permet de choisir, pour les rapports de poids, une borne inférieure L et une borne supérieure U $\left(\text{avec } 0 < L < 1 < U\right)$ n'est active qu'avec le redressement par la *pondération linéaire* et le redressement par la *pondération linéaire consistante*.

Remarque : Ces restrictions proviennent du fait que, selon le contexte, certaines méthodes sont désactivées, celles-ci débouchant sur des résultats semblables aux méthodes imposées, mais avec davantage de calculs. Par exemple, lorsque le modèle ne comporte qu'une seule variable catégorielle, la *poststratification* livre un résultat identique à *IPF* ou à la *pondération linéaire*. Du point de vue du calcul numérique, la *poststratification* est dans ce cas la méthode la plus performante.

Bascula ne permet d'utiliser que des calages sur un seul niveau d'observation avec l'option de contraindre l'égalité de poids pour les unités d'une même grappe pour un sondage par grappe.

On note que contrairement à Calmar et g-Calib 2, Bascula permet de calculer une estimation de la variance en utilisant deux méthodes d'estimation de la variance : la méthode par linéarisation de Taylor et la méthode connue sous le nom de *demi-échantillons équilibrés* (*Balanced repeated replication*).

8.1.4 Output

Avant de lancer une requête de calage, Bascula permet d'effectuer certains contrôles de cohérence sur les tables et les variables en entrée (voir page 41 du manuel). Le bilan des différentes étapes avec quelques statistiques sur les pondérations (page 69 de Nieuwenbroek et Boonstra, 2002) sont sauvegardés dans un fichier dont l'extension est « .blg ». Les valeurs des poids finaux et celles des rapports de poids sont sauvegardés dans un fichier dont l'extension est « .wga ». Si ce fichier existe déjà, le nouveau fichier de pondération calculé par Bascula écrase l'ancien, ce qui ne permet pas d'empiler des jeux de pondération obtenus selon des méthodes différentes. Le seul moyen pour pouvoir sauvegarder les poids obtenus avec les différentes méthodes est de prévoir dans la base de données autant de variables servant à sauvegarder ces pondérations que les modèles de calage envisagés.

8.1.5 Problèmes

Le problème le plus sérieux qu'on peut rencontrer est propre à la technique du calage. Le calage ne peut pas être réalisé d'une manière parfaite dans les cas suivants :

- L'algorithme peut ne pas converger avec le nombre d'itérations maximal fixé ;
- L'algorithme peut ne pas converger.

Ce problème peut notamment se produire en cas d'utilisation d'une fonction de distance tronquée, si les bornes imposées aux rapports de poids sont trop contraignantes.

8.1.6 Documentation

Contrairement aux manuels d'utilisation de g-Calib 2 et de Calmar 2, celui de Bascula 4.0 n'est pas assez clair et ne contient pas beaucoup d'explications où les variables utilisées ne sont pas bien définies. C'est le cas par exemple des

définitions des variables : *Inclusion weight* et *correction weight* (voir pages 11-12 de Nieuwenbroek et Boonstra, 2002).

8.2 Résultats obtenus avec Bascula 4.0 selon les différentes fonctions de calage

Les limites de Bascula sont les suivantes :

- Une variable catégorielle ne doit pas avoir plus de 2000 modalités.
- Les valeurs des probabilités d'inclusion ne doivent pas être négatives. Lors des calculs, les observations correspondantes aux probabilités d'inclusion nulles sont ignorées par Bascula.
- Pour le calage par grappe, les individus de la même grappe doivent se suivre dans la base de données et leurs probabilités d'inclusion doivent être égales.
- Au maximum 200 échantillons peuvent être sélectionnés (pour l'estimation de la variance par réplication d'échantillons).

Les résultats obtenus avec Bascula 4.0 sont :

8.2.1 Calage simple au niveau ménage

- Avec la méthode linéaire, le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,96 et 1,08.
- Avec la méthode du raking ratio, le calage est réalisé exactement. Les rapports de poids sont compris entre 0,96 et 1,08.
- Avec la méthode linéaire tronquée, pour des rapports de poids compris entre :
    - [0,97-1,08] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,97 et 1,06.
    - [0,98-1,08] : convergence imparfaite après 15 itérations (message : Weight restriction not succeeded within 15 iterations). Bascula fournit comme même les poids finaux (Les contraintes de calage sont satisfaites). L'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,97 et 1,06

- [0,97-1,07] : convergence imparfaite après 15 itérations. Bascula fournit comme même les poids finaux. L'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,97 et 1,051

Ainsi, les bornes limites sont **[*L ; U*]=[0,97 ;1,08]**

8.2.2 Calage simple au niveau individu

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,96 et 1,5.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,96-1,06] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,96 et 1,06.
  - [0,97-1,06] : convergence imparfaite après 15 itérations. Bascula fournit comme même les poids finaux. L'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,97 et 1,06.
  - [0,96-1,05] : convergence imparfaite après 15 itérations. Bascula fournit comme même les poids finaux. L'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,96 et 1,06.

Ainsi, les bornes limites sont **[*L ; U*]=[0,96 ;1,06]**

8.2.3 Calage simple au niveau individu dans un sondage en grappe

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,94 et 1,5.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,95-1,05] : convergence imparfaite après 15 itérations. L'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,94 et 1,08.
  - [0,95-1,06] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,95 et 1,06.
  - [0,96-1,06] : convergence imparfaite après 15 itérations. Bascula fournit comme même les poids finaux. L'estimateur calé des totaux des variables et

les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,95 et 1,064.

Ainsi, les bornes limites sont **[*L ; U*]=[0,95 ;1,06]**

8.2.4 Calage stratifié

Strate 1

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,76 et 1,3.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,84-1,31] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,85 et 1,304
  - [0,84-1,30] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,85 et 1,304
  - [0,85-1,31] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,85 et 1,304

Ainsi, les bornes limites sont **[*L ; U*]=[0,84 ;1,31]**

Strate 2

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,81 et 1,116.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,81-1,10] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,81 et 1,10
  - [0,82-1,10] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,81 et 1,10
  - [0,81-1,09] : convergence imparfaite après 20 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,81 et 1,10

Ainsi, les bornes limites sont **[*L ; U*]=[0,81 ;1,10]**

Strate 3

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,76 et 1,185.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
    - [0,91-1,19] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,9180 et 1,182
    - [0,91-1,18] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,91 et 1,181
    - [0,92-1,19] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,91 et 1,182

Ainsi, les bornes limites sont **[*L ; U*]=[0,91 ;1,19]**

Strate 4

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,89 et 1,101.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
    - [0,91-1,11] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,91 et 1,106
    - [0,90-1,11] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,91 et 1,106
    - [0,91-1,10] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,91 et 1,106

Ainsi, les bornes limites sont **[*L ; U*]=[0,91 ;1,11]**

Strate 5

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,82 et 1,109.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
    - [0,93-1,10] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,93 et 1,092

- [0,94-1,10] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,93 et 1,094
- [0,93-1,09] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,93 et 1,091

Ainsi, les bornes limites sont **[*L ; U*]=[0,93 ;1,10]**

Strate 6

- Fonction linéaire : le calage est réalisé exactement, pas de poids négatifs. Les rapports de poids sont compris entre 0,88 et 1,325.
- Fonction linéaire tronquée, pour des rapports de poids compris entre :
  - [0,90-1,11] : le calage est réalisé exactement. Les rapports de poids sont compris entre 0,90 et 1,108
  - [0,91-1,11] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,89 et 1,107
  - [0,90-1,10] : convergence imparfaite après 15 itérations. Les poids finaux sont fournis et l'estimateur calé des totaux des variables et les vrais totaux sont égaux. Les rapports de poids sont compris entre 0,89 et 1,106

Ainsi, les bornes limites sont **[*L ; U*]=[0,90 ;1,11]**.

Comparaison Calmar2, g-Calib 2 et Bascula 4.0

9 Comparaison Calmar 2, g-Calib 2 et Bascula 4.0

Tableau 9.1

tableau synoptique des intervalles des rapports de poids obtenus avec les trois logiciels

Méthode de calage	Calmar 2	g-Calib 2	Bascula 4.0
Calage au niveau ménage			
Méthode linéaire	[0,96 ;1,08]	[0,96 ;1,08]	[0,96 ;1,08]
Méthode exponentielle	[0,96 ;1,08]	[0,96 ;1,08]	[0,96 ;1,08]
Méthode linéaire tronquée	[0,97 ;1,07]	[0,97 ;1,06]	[0,97 ;1,08]
Méthode logit	[0,97 ;1,07]	[0,97 ;1,06]	--
Calage au niveau individus			
Méthode linéaire	[0,96 ;1,50]	[0,96 ;1,49]	[0,96 ;1,50]
Méthode exponentielle	[0,96 ;1,60]	[0,96 ;1,60]	--
Méthode linéaire tronquée	[0,96 ;1,10]	[0,97 ;1,06]	[0,96 ;1,06]
Méthode logit	[0,96 ;1,10]	[0,97 ;1,20]	--
Calage avec égalité des poids dans la grappe		(Lemaître-Dufour)	(Lemaître-Dufour)
Méthode linéaire	[0,90 ;1,42]	[0,94 ;1,44]	[0,94 ;1,50]
Méthode exponentielle	[0,90 ;1,49]	[0,94 ;1,51]	--
Méthode linéaire tronquée	[0,90 ;1,10]	[0,97 ;1,10]	[0,95 ;1,06]
Méthode logit	[0,90 ;1,10]	[0,96 ;1,10]	--
Calage en grappes simultané			
Méthode linéaire	[0,87 ;1,37]	[0,87 ;1,37]	--
Méthode exponentielle	[0,88 ;1,42]	[0,88 ;1,42]	--
Méthode linéaire tronquée	[0,90 ;1,10]	[0,94 ;1,07]	--
Méthode logit	[0,92 ;1,08]	[0,93 ;1,07]	--

Méthode de calage	Calmar 2	g-Calib 2	Bascula 4.0
Calage stratifié			
Strate 1			
Méthode linéaire	[0,76 ;1,28]	[0,76 ;1,28]	[0,76 ;1,30]
Méthode exponentielle	[0,77 ;1,29]	[0,77 ;1,29]	--
Méthode linéaire tronquée	[0,76 ;1,28]	[0,77 ;1,28]	[0,84 ;1,31]
Méthode logit	[0,77 ;1,28]	[0,77 ;1,27]	--
Strate 2			
Méthode linéaire	[0,81 ;1,12]	[0,81 ;1,12]	[0,81 ;1,12]
Méthode exponentielle	[0,81 ;1,12]	[0,81 ;1,12]	--
Méthode linéaire tronquée	[0,81 ;1,11]	[0,81 ;1,11]	[0,81 ;1,10]
Méthode logit	[0,81 ;1,11]	[0,81 ;1,10]	--
Strate 3			
Méthode linéaire	[0,76 ;1,18]	[0,76 ;1,18]	[0,76 ;1,18]
Méthode exponentielle	[0,77 ;1,19]	[0,77 ;1,19]	--
Méthode linéaire tronquée	[0,77 ;1,18]	[0,90 ;1,18]	[0,91 ;1,19]
Méthode logit	[0,76 ;1,17]	[0,90 ;1,18]	--
Strate 4			
Méthode linéaire	[0,89 ;1,10]	[0,89 ;1,10]	[0,89 ;1,10]
Méthode exponentielle	[0,89 ;1,11]	[0,89 ;1,11]	--
Méthode linéaire tronquée	[0,89 ;1,09]	[0,90 ;1,09]	[0,91 ;1,11]
Méthode logit	[0,89 ;1,09]	[0,89 ;1,08]	--
Strate 5			
Méthode linéaire	[0,89 ;1,11]	[0,82 ;1,11]	[0,82 ;1,11]
Méthode exponentielle	[0,84 ;1,11]	[0,84 ;1,11]	--
Méthode linéaire tronquée	[0,82 ;1,10]	[0,93 ;1,10]	[0,93 ;1,10]
Méthode logit	[0,84 ;1,10]	[0,92 ;1,10]	--

Strate 6			
Méthode linéaire	[0,88 ;1,32]	[0,88 ;1,32]	[0,88 ;1,32]
Méthode exponentielle	[0,88 ;1,37]	[0,88 ;1,37]	--
Méthode linéaire tronquée	[0,88 ;1,10]	[0,90 ;1,10]	[0,90 ;1,11]
Méthode logit	[0,88 ;1,10]	[0,90 ;1,10]	--

Le tableau 9.1 montre que, pour ces trois logiciels de calage, quels que soient le modèle et la méthode de calage considérés, les bornes des intervalles pour les rapports de poids sont égales dans presque tous les cas. Ces bornes sont différentes uniquement dans le cas de calage par grappe avec égalité des poids dans la grappe. Cette différence est due au fait que g-Calib 2 et Bascula 4.0 utilisent la méthode Lemaître-Dufour, ce qui n'est pas le cas pour Calmar 2.

Tableau 9.2
tableau synoptique des différences entre les trois logiciels

	Calmar 2	g-Calib 2	Bascula 4.0
Modèles de calage associés aux plans de sondage			
• Sondage à un degré et calage simple	OUI	OUI	OUI
• Sondage en grappes, calage sur la population des unités secondaires avec égalité des poids dans la grappe	OUI	OUI	OUI
• Sondage en grappes et calage simultané sur des totaux de grappes et d'unités secondaires	OUI	OUI	NON
• Sondage à deux degrés, calage simultané sur des totaux d'unités primaires et d'unités secondaires (deux niveaux d'observation)	OUI	NON	NON
• Sondage à deux degrés, calage simultané sur des totaux d'unités primaires et d'unités secondaires avec trois niveaux d'observation	OUI	NON	NON
• Sondage stratifié et calage séparé par strate	NON	OUI	NON
Redressement de la non-réponse utilisant des variables instrumentales relevées sur les répondants seulement	OUI	NON	NON
Fonctions de calage	• Les 4 fonctions fondamentales • Sinus hyperbolique	Les 4 fonctions fondamentales	4 méthodes de redressement

	Calmar 2	g-Calib 2	Bascula 4.0
Algorithme de calage :			
• Inversion de matrice	Inversion normale par défaut, généralisée optionnelle	Inversion généralisée systématique	Inversion généralisée systématique
• Calage en grappes sans données de niveau grappe	Calage des totaux de grappes	Calage des moyennes de grappes (Lemaître-Dufour)	Calage des moyennes de grappes (Lemaître-Dufour)
Précision des calculs par défaut :			
• Critère d'arrêt	10^{-4}	10^{-5}	10^{-8}
• Nombre maximum d'itérations	15	20	15
Initialisation des paramètres par défaut :			
• PONDQK	1	non initialisé	1
• ECHELLE	1	$X_1 / \hat{X}_{1\pi}$, X_1 étant la 1ère variable de calage listée	
• Bornes des rapports de poids (méthodes logit et linéaire tronquée)	non initialisées	[0 - 1,5]	non initialisées
Paramètres obligatoires			
• PONDQK	NON (=1 par défaut)	OUI	NON
• STRATE	n'existe pas	OUI	n'existe pas
dans un sondage en grappes :			
• POIDS de la grappe	OUI	OUI	OUI
• POIDS de l'unité secondaire	NON	OUI	NON
Mode de spécification des contraintes de calage	Table SAS de format imposé	Interface de saisie	Interface de saisie
Mode de spécification des totaux de calage	Valeur absolue ou pourcentage	Valeur absolue	Valeur absolue

	Calmar 2	g-Calib 2	Bascula 4.0
Table des poids de calage	• Mise à jour d'une table existante par ajout de variable sur option	• Table créée à chaque requête	• Table créée à chaque requête
	• Variables identifiant et poids de calage	• Variables identifiant, poids de calage, rapport de poids (g_k), facteur d'échelle, poids initial multiplié par le facteur d'échelle, poids initial et variables de calage	• Poids de calage, rapports de poids (g_k), poids initial
Avec un sondage en grappes sans données de niveau grappes	Table de niveau unités secondaires	Table de niveau grappes	Table de niveau grappes
Avec un sondage en grappes et un calage simultané	Deux tables de poids, niveau grappes et niveau unités secondaires	Une table de poids de niveau grappes	Une table de poids de niveau grappes
Ordre de tri	Celui de la table échantillon en entrée	Strate, identifiant	
Autre table en sortie	Observations éliminées (sur option)		
Editions			
• Etendue	Modulable	Fixe	Fixe
• Valeur des coefficients λ	A chaque itération	A la dernière itération	
• Tableau comparatif des totaux X et $\hat{X}_{cal}$ des variables de calage	• En valeur absolue et en % pour les variables catégorielles	• En valeur absolue	• En valeur absolue
	• $X - \hat{X}_{cal}$ calculé en valeur absolue	• $X - \hat{X}_{cal}$ calculé en % de X	• $X - \hat{X}_{cal}$ calculé en valeur absolue

	Calmar 2	g-Calib 2	Bascula 4.0
• Sauvegarde • Données individuelles	En format SAS	Format SPSS par défaut et exportation en format HTML sur option Editées partiellement sur option	En format Ascii ou format Blaise
Fichier Log	Distinct du fichier de résultats	Confondu avec le fichier de résultats	Distinct du fichier de résultats
Articulation avec d'autres programmes de traitement de la table échantillon	Exécution de CALMAR 2 possible à l'intérieur d'un programme SAS ou d'une macro SAS de l'utilisateur	**Sur option** : • lancement d'un programme SPSS avant calage (traitement des données) • lancement d'un programme SPSS après calage (estimation d'un paramètre) • à partir de l'interface de saisie	Exécution (interactive) de Bascula possible à l'intérieur Blaise et depuis Manipula (Blaise script). Possible aussi depuis divers programmes tournant sous Windows (VBA, Delphi, C++, S-Plus).

Conclusion générale

10 Conclusion générale

Notre travail portait sur la comparaison de trois logiciels permettant de redresser des échantillons par calage sur une information auxiliaire : Bascula 4.0 (programmé en langage Pascal (Delphi)), Calmar 2 (programmé en langage SAS) et g-Calib 2 (transformations des données et calculs programmés en langage SPSS ; l'interface programmé en langage Visual Basic). Nous avons pour cela effectué un certain nombre de tests, présentés dans le corps du rapport. Dans la mesure du possible, ces tests ont été communs à tous les logiciels.

Les tests menés sur Bascula 4.0 ont fait apparaître un certain nombre de problèmes. Le manuel de l'utilisateur est peu détaillé, ce qui rend sa prise en main difficile. La phase de préparation des données en entrée et de récupération des poids calés en sortie est assez longue. D'autre part, le logiciel est moins général que ses deux concurrents car toutes les fonctionnalités du calage proposées par Deville et Särndal (1992) ne sont pas proposées, et suivant les variables de calage utilisées, le choix de la méthode de redressement est imposé ou tout au moins limité par le logiciel[3] (voir à ce propos la remarque du point 8.1.3).

Les manuels de g-Calib 2 et Calmar 2 sont tous les deux très complets, et enrichis de nombreux exemples. Le manuel de g-Calib 2 souffre quelque peu d'un recours systématique au langage matriciel qui ne facilite pas forcément la compréhension ; d'autre part, il faudrait fusionner les manuels des versions 1 et 2 pour plus de clarté. Les deux logiciels permettent de redresser des poids par calage à l'aide des fonctions de distance proposées par Deville et Särndal dans l'article fondateur. Calmar 2 propose également une fonction de distance proposée

[3] Bascula 4.0 présente tout de même l'avantage de permettre une estimation de la variance, ce qui n'est pas possible avec Calmar 2 et g-Calib 2. Le manuel précise quels plans de sondages permettent une estimation de la variance. Cependant, il n'est pas indiqué comment cette estimation prend en compte la non-réponse et éventuellement plusieurs degrés de tirage. D'autre part, une des méthodes d'estimation de variance se base sur des réplications d'échantillons, et le nombre maximum de simulations possibles est faible (égal à 200).

plus récemment, celle du sinus hyperbolique (Roy et Vanheuverzwyn, 2001). g-Calib 2 permet de mettre en œuvre la technique de « collapsing » permettant de regrouper les unités identiques au sens des variables de calage afin de limiter les calculs : cette méthode est surtout intéressante dans le cas d'un calage sur des variables catégorielles uniquement. g-Calib 2 permet de réaliser un calage par strates. Calmar 2 permet d'utiliser la technique de calage généralisé (Deville, 2000) pour redresser la non-réponse.

La phase d'installation de g-Calib 2 a été délicate. En particulier, nous avons rencontré des problèmes de compatibilité avec la version française de SPSS. D'autre part, il ne paraît pas possible d'utiliser g-Calib 2 en réseau. Une fois ces problèmes résolus, la prise en main du logiciel est facilitée par une interface conviviale. Trois principaux points de vigilance nous paraissent importants :

- Il n'existe pas, comme c'est le cas dans Calmar 2, de contrôles sur la saisie et la cohérence des paramètres.
- Des problèmes se posent avec la fonction linéaire bornée, qui renvoie parfois des poids calés totalement aberrants. Ces problèmes semblent dus à une utilisation systématique des matrices g-inverses.
- Les sorties données par la Macro sont difficilement lisibles et mériteraient d'être mieux classifiées et commentées.

Il nous paraît essentiel d'améliorer ces points afin de permettre une large utilisation du logiciel. Les autres améliorations à envisager seraient :

- La possibilité de prendre en compte des poids de second degré différents de 1 pour un calage à deux degrés.
- Le redressement de la non-réponse par calage généralisé.

Bibliographie

11 Bibliographie

DEMING, W.E. et STEPHAN, F.F. [1940], On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **81,** 427-444.

DEVILLE, J.-C. [2002], La correction de la non-réponse par calage généralisé. *Journées de Méthodologie Statistique*, Insee.

DEVILLE, J.-C. et SÄRNDAL, C.-E. [1992], Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87,** 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E. et SAUTORY, O. [1993], Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88,** 1013-1020.

NIEUWENBROEK, N. et BOONSTRA, H.J. [2002], Bascula 4.0 Reference manuel, Statistics Netherlands.

ROY, G. et VANHEUVERZWYN, A. [2001], Redressement sur la macro Calmar : Applications et pistes d'amélioration. Traitements des fichiers d'enquête, *Presses universitaires de Grenoble*, pp. 31-46.

SAUTORY, O. et LE GUENNEC, J. [2003], La macro Calmar 2 : Redressement d'un échantillon par calage sur marges, Document de travail de la direction des statistiques démographiques et social, *INSEE*.

SAUTORY, O [1991], Redressement d'échantillons auprès des ménages par calage sur marges, Document de travail de la direction des statistiques démographiques et social, *n°F9103*, *INSEE*.

VANDERHOEFT, C. [2002], g-Calib Release 1.0 – Generalised Calibration under *SPSS*, Statistics Belgium.

VANDERHOEFT, C. [2003], g-Calib Release 2.0 – Supplement to the Manual for Release 1.0, Statistics Belgium.

Informations

La Direction générale Statistique et Information économique relève du SPF Economie, PME, Classes moyennes et Energie. Une de nos missions est de répondre aux besoins des autorités, des entreprises et des citoyens par une information chiffrée sur la situation réelle du pays dans différents domaines d'actualité

Où trouver l'information statistique et économique?

Sur nos sites Internet http://statbel.fgov.be (statistiques) et http://economie.fgov.be (économie)

Dans cinq grandes villes du pays, la Direction générale Statistique et Information économique met à la disposition du public :

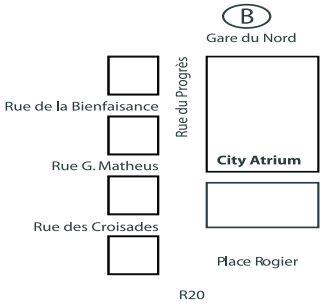
- ◊ Des annuaires et des publications spécialisées ainsi qu'une sélection de disquettes et de cédéroms.
- ◊ Une salle de lecture où il est possible de consulter nos publications, ainsi que celles d'autres ministères ou d'institutions belges et internationales.

Toutes nos bibliothèques sont accessibles les jours ouvrables de 8h30 à 16h30 (Bxl) ou de 9h à 12h et de 13h à 16h (autres).

Bruxelles City Atrium C

Rue du Progrès 50, 1210 Bruxelles
tél. 02/277.55.03 – 02/277.55.04 fax 02/277.55.19
e-mail : info@statbel.economie.fgov.be

Train (B) :	Gare du Nord
Métro (M) :	ligne 2, station Rogier
Trams :	3, 52, 55, 56, 81, 90 arrêts Rogier ou Nord
Bus STIB :	38, 58, 61 arrêts Rogier ou Nord
Bus De Lijn :	318, 351, 358, 410, 526, 554 arrêt Nord



Anvers

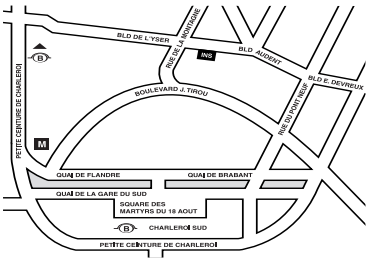
Italiëlei 124 - bus 85, 2000 Antwerpen
tél. 03/229.07.00 fax 03/233.28.30
e-mail : info.antwerpen@statbel.economie.fgov.be

Train (B) :	Centraal Station
Métro (M) :	arrêt Opera
Tram-Bus :	accès facile (Fr. Rooseveltplaats)

Charleroi

Tour Biarent, Bd Audent 14/5, 6000 Charleroi
tél. 071/27.44.14 fax 071/27.44.19
e-mail : info.charleroi@statbel.economie.fgov.be

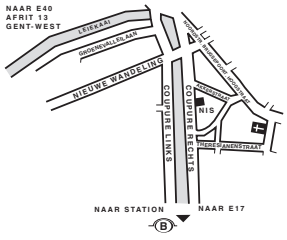
Train (B) :	Charleroi Sud, 20 min depuis la gare (Place Buisset, Rue du Collège, Place Charles II, Boulevard Tirou, rue de la Montagne)
Bus :	arrêt Tirou
Autoroute :	petite ceinture de Charleroi - sortie Gare du Sud
Parking (P) :	payant face à l'INS



Gand

Coupure rechts 620, 9000 Gent
tél. 09/267.27.00 fax 09/267.27.29
e-mail : info.gent@statbel.economie.fgov.be

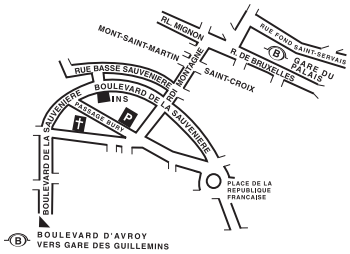
Train (B):	Gent St. Pieters
Tram-Bus :	40, 43 arrêt Theresianenstraat
Autoroute:	accès aisé par autoroute E40 (sortie N° 13 - Gent - West/Drongen)
Parking (P):	au long de la "Coupure Rechts"



Liège

Bd de la Sauvenière 73-75, 4000 Liège
tél. 04/223.84.11 fax 04/222.49.94
e-mail : info.liege@statbel.economie.fgov.be

Train (B):	Gare des Guillemins ou Gare du Palais
Tram-Bus :	(Guillemins) 1 et 4 arrêt Sauvenière
Parking (P):	Neujean (à 20 m - même trottoir) Mercure (en face)



Nous diffusons de nombreux produits qui donnent une image chiffrée de la réalité socio-économique belge. Ces produits, repris dans notre catalogue, sont disponibles auprès de nos centres régionaux ou auprès de notre service de Documentation - vente de Bruxelles. Notre catalogue vous sera envoyé sur simple demande. (voir adresses ci-contre).
Vous trouverez également un extrait de nos données, ainsi que la liste de nos publications sur notre site Internet : http://statbel.fgov.be

Publications générales

Communiqué hebdomadaire

Chaque semaine, nous vous donnons la primeur des dernières statistiques disponibles dans les domaines suivants : Territoire et environnement ; Population ; Société ; Économie et finances ; Agriculture ; Industrie ; Services, commerce et transport.

Chiffres-clés

Cette petite publication explore notre territoire sous ses aspects les plus divers : le climat, l'environnement, la population, la vie sociale, l'économie, les finances, l'agriculture, l'industrie, le transport, la société de l'information… Chiffres-clés 2004 est une brochure gratuite de 50 pages, en couleurs, de format réduit. Vous y trouverez une sélection de la rubrique Statistiques de notre site Internet brossant une vue singulière de l'information statistique disponible en Belgique. Les tableaux sont éclairés par des graphiques et des cartogrammes.

Quelques autres publications

Publications générales

Annuaire de statistiques régionales

Territoire et environnement

Statistique de l'occupation du sol (**disquette**)
Aperçu Environnement - *Annuel*

Population

Mouvement de la population – *Annuel*
Perspectives de population 2000/2050

Société

Enquête sur les budgets des ménages – *Annuel*
Causes de décès - *Annuel*

Économie et finances

Vente de biens immobiliers – *Annuel*
La conjoncture - *Mensuel*

Agriculture

Recensement agricole et horticole
au 15 mai - *Annuel*

Industrie

Production industrielle et construction - *Mensuel*

Commerce, services et transports

Statistiques mensuelles du transport – *Mensuel*
Commerce intérieur - *Annuel*

Achevé d'imprimer
par l'imprimerie de la
Direction générale Statistique
et Information économique
B-1000 Bruxelles

Novembre 2006