CALAGE DU SOUS-ECHANTILLON ANNUEL DE L'EEC SUR LES RÉSULTATS TRIMESTRIELS

Frédéric OURADOU[1] (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Résumé

L'enquête emploi en continu (EEC) est une enquête trimestrielle en panel. Chaque trimestre, six vagues sont interrogées : une vague en 1re interrogation, dite « entrante », la vague entrante du trimestre précédent, en 2e interrogation… jusqu'à la vague « sortante », en 6e interrogation. Certaines questions de l'enquête ne sont posées qu'en vague 1, d'autres seulement en vagues 1 et 6. Pour les variables correspondantes, les statistiques annuelles sont donc fondées sur un sous-échantillon de l'EEC, constitué de la compilation sur l'année des vagues pour lesquelles on dispose de ces variables. Un règlement d'Eurostat relatif à l'enquête Emploi impose, dans ce cas, qu'une « *cohérence entre les totaux annuels des sous-échantillons et les moyennes annuelles des échantillons complets* (soit) *assurée pour l'emploi, le chômage et l'inactivité par sexe et* (par) *tranches d'âge*». La contribution expose la méthode utilisée pour garantir cette cohérence.

Pour le calage des totaux du « sous-échantillon annuel » sur les moyennes annuelles des totaux trimestriels, dit plus simplement calage annuel, on peut souhaiter utiliser des marges de calage supplémentaires (niveau individu ou ménage), en plus de celles requises par Eurostat. La cohérence du calage peut être annuelle ou bien trimestrielle, et dans ce dernier cas, les totaux correspondant à la vague 1, ou aux vagues 1 et 6 d'un trimestre donné doivent correspondre aux totaux du trimestre sur l'ensemble des six vagues.

Différents scénarios de calage sont réalisés, ces scénarios différant selon les variables de calage, selon la méthode de calage, *Raking ratio* ou *Logit*, et selon la taille de l'échantillon à caler.

Les différents scénarios sont comparés selon plusieurs critères : nombre d'itérations du logiciel Calmar pour arriver à la convergence, statistiques de distribution des poids finals et des rapports de poids, troncature des poids extrêmes, effet du calage sur certaines variables dites « de contrôle ».

On testera notamment la robustesse du calage vis-à-vis du nombre de variables de calage, de la fenêtre des rapports de poids dans le cas de la méthode *Logit*, et on s'intéressera au comportement de Calmar en présence d'observations présentant des valeurs manquantes pour certaines variables de calage.

[1] frederic.ouradou@insee.fr

Abstract

The Labour Force Survey (LFS) is a quarterly "panel" survey: every household in the sample is surveyed for six quarters, called "waves". Both annual and quarterly LFS statistics are disseminated. Annual statistics have to be assessed by the collection of waves 1 or waves 1 and 6. An Eurostat regulation about LFS requires that « *consistency between annual sub-sample totals and full-sample annual averages shall be ensured for employment, unemployment and inactive population by sex and* (*by*) *age groups.* ».

This paper explains how this requirement is fulfiled. Different calibration scenarios are compiled: they differ from each other by the list of calibration variables, by the calibration method, *Raking ratio* or *Logit*, and by the size of the sample to calibrate. The quality of the calibration is examined through different criteria, such as the distribution of the calibrated weights or the influence of calibration on "control" variables.

Mot-clé

Calage

Introduction

L'enquête emploi en continu (EEC) est une enquête trimestrielle en panel. Chaque trimestre, six vagues sont interrogées : une vague en 1re interrogation, dite « entrante », la vague entrante du trimestre précédent, en 2e interrogation… jusqu'à la vague « sortante », en 6e interrogation.

Certaines questions de l'enquête ne sont posées qu'en vague 1, d'autres seulement en vagues 1 et 6. Pour les variables correspondantes, les statistiques annuelles sont donc fondées sur un sous-échantillon de l'EEC, constitué de la compilation sur l'année des vagues pour lesquelles on dispose de ces variables.

L'Insee transmet à Eurostat des résultats trimestriels de l'EEC mais aussi des résultats annuels. Ces derniers sont évalués à partir d'un "sous-échantillon" constitué soit de la seule vague 1, soit des vagues 1 et 6 des enquêtes trimestrielles. Les résultats annuels "spontanés" peuvent donc différer des résultats trimestriels.

Un règlement européen impose « *la cohérence entre les totaux annuels des sous-échantillons et les moyennes annuelles des échantillons complets doit être assurée pour l'emploi, le chômage et l'inactivité par sexe et pour les tranches d'âge suivantes : 15-24 ans, 25-34 ans, 35-44 ans, 45-54 ans et 55 ans ou plus* ».

Il faut donc caler les totaux annuels du sous-échantillon sur les moyennes annuelles de l'échantillon complet pour le croisement indiqué des variables sexe, âge et activité.

L'objet de ce document est de préciser la méthode de calage, qui sera appliquée à l'année 2013. La partie I présente la méthode de calage et ses options, les parties II à IV présentent chacune différents scénarios de calage : la partie II examine l'influence sur le calage d'un certain nombre de paramètres ; la partie III regarde plus particulièrement l'influence du nombre de variables de calage ; la partie IV celle de la fenêtre des rapports de poids pour la méthode *Logit*.

I. La méthode de calage annuel

Les traitements post-collecte de l'EEC dans les DOM, et notamment le calcul des poids, font l'objet d'un traitement séparé de celui effectué pour la métropole. Les travaux évoqués dans ce document portent sur le seul **échantillon métropolitain de l'EEC**.

1. La base trimestrielle des données individuelles

Pour les travaux de calage, on part de la base dite « *z* », base trimestrielle « finale » des données individuelles de l'EEC. Cette base permet de compiler les résultats bruts, c'est-à-dire non corrigés des variations saisonnières, de l'EEC. C'est la base des **individus** des **ménages répondants**, y compris les individus non répondants de 15 ans et plus ainsi que les individus non éligibles à un questionnaire individuel (individus âgés de moins de 15 ans, individus de 15 ans et plus hors champ de l'EEC), et y compris les individus des ménages constitués d'inactifs de 65 ans ou plus en vague intermédiaire[2] (res = 16). Les individus non répondants de 15 ans et plus ont un poids (variable *extri14*) nul et ne sont pas pris en compte lors du calage. Ils sont néanmoins récupérés dans le fichier final.

Les poids figurant dans la base *z* sont des poids post-calage trimestriel, et à ce titre corrigés de la non-réponse. Cette base contient différents jeux de poids trimestriels :

- *extri14* est le poids correspondant à la base des individus, 14 désignant le « millésime » des poids[3];
- *extrilog14* est le poids correspondant à la base des logements ;
- *extrid14* est le poids correspondant au sous-échantillon des individus de la vague 1. On a *extrid14* = 6 x *extri14* en vague 1, *extrid14* = 0 sinon ;
- *extridf14* est le poids correspondant au sous-échantillon des individus des vagues 1 et 6. On a *extridf14* = 3 x *extri14* en vagues 1 et 6, *extrid14* = 0 sinon.
- *coeffq* : poids individuel trimestriel transmis à Eurostat. Ce poids correspond à la variable *extri*,

ainsi que le poids individuel annuel *coeffy* transmis à Eurostat. C'est ce poids dont on cherche ici à expliciter la méthode de calcul.

Dans la suite de ce document, on omettra le millésime des poids. On parlera donc des poids *extri*, *extrilog*…

Par la suite, on appellera :

- totaux trimestriels les résultats de l'EEC compilés à partir des poids trimestriels *extri* d'une / de base(s) trimestrielle(s) *z* ;
- moyennes annuelles des totaux trimestriels les **moyennes arithmétiques** des quatre totaux trimestriels. Les poids correspondants sont donc égaux à *extri*/4 ;
- totaux annuels spontanés du sous-échantillon annuel les totaux calculés à partir du sous-échantillon annuel et des poids *extrid*/4 (si seule la vague 1 est retenue pour le sous-échantillon annuel) ou *extridf*/4 (quand les deux vagues 1 et 6 sont retenues) ;
- calage des totaux du sous-échantillon annuel sur les moyennes annuelles des totaux trimestriels, ou plus simplement calage annuel, l'opération qui consiste à calculer des poids pour le sous-échantillon annuel permettant de « retrouver » les moyennes annuelles des totaux trimestriels pour un ensemble déterminé de variables. Ces poids sont appelés poids annuels calés.
- totaux annuels calés les totaux calculés à partir du sous-échantillon annuel et des poids annuels calés.

Les pondérations au lancement du calage annuel sont celles issues de l'EEC trimestrielle (*extrid*/4 ou *extridf*/4), plutôt que les poids d'échantillonnage. Cela implique qu'**on ne tient pas compte des logements hors champ** (résidences secondaires et logements vacants) **lors du calage annuel**.

[2] Ces ménages ne sont pas réinterrogés en vague intermédiaire (vagues 2 à 5 de l'enquête), mais sont intégrés aux micro-données du trimestre courant ; leur statut d'activité est supposé stable au cours du temps.

[3] i.e. le millésime du recensement duquel sont tirées les marges utilisées pour caler les poids de l'EEC. Ce millésime est omis dans le nom des poids de l'EEC lorsque ceux-ci deviennent définitifs et ne font plus l'objet de révisions.

Le calage est effectué par la macro SAS Calmar dont on trouvera une présentation dans [1].

2. Un calage à partir d'une base de logements

Dans le cadre du calage annuel, le règlement européen impose que tous les individus d'un même logement aient le même poids annuel calé. Cela oblige à caler un échantillon de logements, et non un échantillon d'individus, et ce alors que l'unité répondante de l'EEC est l'individu.

Dans un tel calage, une difficulté provient de l'existence de logements où certains individus interrogés répondent à l'enquête et d'autres non. Puisqu'il s'agit d'une non-réponse totale d'une partie des individus du logement, cette non-réponse sera appelée par la suite « non-réponse totale partielle ».

Dans un fichier d'individus, on peut tenir compte de cette non-réponse totale partielle en distinguant le poids d'un individu du poids du logement dont il est issu. Par exemple, si un ménage est composé d'un couple mixte dont seule la femme a répondu à l'enquête, le poids de la personne répondante, corrigeant de la non-réponse totale partielle mais avant calage, sera le double de celui du logement.

Dans un fichier de logements, le poids d'une observation est nécessairement un poids « logement ». Pour le calage annuel, on partira du poids *extrilog* issu du calage trimestriel, multiplié par un coefficient tenant compte de l'abandon de certaines vagues dans le sous-échantillon annuel et du passage de trimestriel en annuel.

Dès lors, pour tenir compte de la non-réponse totale partielle au sein d'un logement, et permettre le calage sur des marges de population, c'est sur les variables individuelles qu'on « intervient ». Dans le cas du couple mixte évoqué ci-dessus, on considérera qu'ont répondu à l'enquête deux personnes ayant toutes les caractéristiques (âge, sexe, activité…) de la personne ayant effectivement répondu.

3. Une typologie des variables de calage

Le règlement impose le calage sur le croisement des variables sexe, âge et activité, soit 2 x 5 x 3 égale 30 marges distinctes.

Toutefois, on peut souhaiter imposer d'autres marges de calage. Celles-ci peuvent être :

- soit des marges de population, soit des marges de logements (*cf*. a)
- soit des marges annuelles, soit des marges trimestrielles (*cf*. b).
- soit des marges déjà utilisées pour le calage trimestriel, soit des résultats de l'EEC (*cf.* c).

a. Marges de population et marges de logements

Certaines marges sont des marges de population, comme par exemple la population de chacune des régions de France métropolitaine, d'autres des marges de « logements », comme le nombre total de résidences principales en France métropolitaine.

Le calage annuel doit permettre de caler simultanément sur des marges de population et des marges de logements, à l'instar du calage trimestriel de l'EEC.

b. Marges annuelles et marges trimestrielles

Le règlement européen impose « *la cohérence entre les totaux annuels des sous-échantillons et les moyennes annuelles des échantillons complets* », qu'on appellera cohérence annuelle.

On peut souhaiter être plus exigeant pour certaines variables, et garantir la cohérence entre le total trimestriel du sous-échantillon et le total trimestriel de l'échantillon complet, qu'on appellera cohérence trimestrielle.

On remarque que, si la cohérence trimestrielle est respectée, alors la cohérence annuelle l'est aussi, puisque la moyenne annuelle des totaux trimestriels retenue est la moyenne arithmétique.

c. Marges utilisées en trimestriel et résultats de l'EEC

Par ailleurs, il faut distinguer :

- les marges de calage qui sont utilisées comme marges des enquêtes trimestrielles ;
- les résultats des enquêtes trimestrielles, qui peuvent être utilisés comme marges de calage du sous-échantillon annuel.

Les marges déjà utilisées lors du calage trimestriel

Les marges utilisées pour le calage de l'EEC trimestrielle sont les suivantes :

- effectifs par sexe, tranche d'âge et région administrative, souvent de façon croisée, provenant du recensement de la population actualisé à partir des données de l'état civil
- effectifs par sexe x tranche d'âge quinquennale, provenant du recensement de la population actualisé à partir des données de l'état civil
- marges issues du fichier de la taxe d'habitation (TH) : décile de revenu du ménage, nombre de pièces du logement, statut d'occupation du logement, type de logement
- nombre de logements construits après 2000, marge issue des comptes du logement
- marges des variables NCH et NACTOP. Il s'agit du nombre de chômeurs (NCH) et du nombre d'actifs occupés (NACTOP). La prise en compte de l'enquête auprès des non-répondants (ENR) dans la mécanique de l'EEC trimestrielle fait de ces variables des variables de calage[4].

Chaque trimestre, chacune des six vagues d'enquête est calée séparément. Les marges issues du bilan démographique sont identiques pour toutes les vagues, contrairement aux marges issues du fichier TH, qui sont fonction du millésime du fichier.

De ce fait, les totaux annuels spontanés du sous-échantillon annuel sont déjà calés pour les variables issues du bilan démographique. Pour conserver le calage, il faut quand même les intégrer aux marges du calage annuel.

Les résultats issus de l'EEC

On peut souhaiter garantir la cohérence annuelle, voire la cohérence trimestrielle, pour certaines variables, et notamment des variables « importantes » de l'EEC.

Eurostat impose déjà la cohérence annuelle sur le croisement des données âge, sexe et activité (***cf.*** **Introduction**). Il est possible de la proposer pour d'autres variables. Ont été également retenues les variables « diplôme le plus élevé » en onze modalités (variable DIP11) ainsi que la catégorie socioprofessionnelle en neuf modalités (variable CSTOTR). Eurostat utilise ces deux variables DIP11 et CSTOTR pour évaluer des indicateurs principaux diffusés en trimestriel et en annuel. Inclure ces variables parmi les variables de calage permet d'assurer la cohérence des résultats.

4. Les options de calage

Les programmes de calage permettent de choisir entre :

- trois méthodes de constitution et de calage du sous-échantillon annuel (macro-variable *meth*[5]) ;
- deux options de calcul des marges des variables de calage TH (macro-variable *optionTH*) ;
- cohérence trimestrielle et cohérence annuelle pour chaque variable de calage (macro-variables *varannu* et *vartrim*).

Les trois méthodes de constitution et de calage du sous-échantillon annuel sont les suivantes :

- *meth*=1 : le sous-échantillon annuel n'est composé que des vagues 1 de chaque trimestre ;

[4] Dit de façon très simplifiée, l'EEC est calée sur le nombre de chômeurs et le nombre d'actifs occupés qu'on obtient à partir des résultats spontanés de l'ensemble des deux enquêtes EEC et ENR calés uniquement sur les marges externes à l'enquête (celles provenant de la TH, du recensement et de l'état civil, des comptes du logement).

[5] On fait référence aux noms des macro-variables dans les programmes de calage.

- *meth*=2 : le sous-échantillon annuel est composé des vagues 1 et 6 et on cale séparément sur chaque vague ;
- *meth*=3 : le sous-échantillon annuel est composé des vagues 1 et 6 et le calage se fait globalement sur ces deux vagues.

Les deux options de calcul des marges des variables de calage issues de la TH sont les suivantes :

- *optionTH*=1 : les marges de calage issues de la TH sont calculées à partir des seules vagues du sous-échantillon annuel ;
- *optionTH*=2 : les marges de calage issues de la TH sont calculées à partir de toutes les vagues.

Dès lors qu'on a fixé la liste des variables de calage, et pour chacune sa fréquence de cohérence, annuelle ou trimestrielle, on a le choix entre 2 « options TH » x 3 méthodes = 6 méthodes de calage, sans compter les options propres à la macro Calmar.

5. Synthèse sur les variables de calage

Le tableau 1 ci-dessous synthétise les choix effectués pour chaque variable de calage.

Tableau 1 : synthèse sur les variables de calage annuel

	A	B	C	D	E	F	G	H	I
1	**Description de la variable**	**Nom**	**Nature**	**Nombre de modalités**	**Sources des marges**	**Marge utilisée lors du calage de l'EEC trimestrielle ?**	**Les totaux annuels spontanés sont-ils précalés ?**	**La marge trimestrielle dépend-elle de la vague ?**	**Variable de calage annuelle ?** ***Si oui, cohérence trimestrielle ou annuelle ?***
2	Effectif métropolitain par tranche d'âge et par sexe	*aqAAsS*	Individu	(4 x) 2 x 16	Recensement	Oui	Oui	Non	Oui. *Trimestrielle*
3	Effectif par région métropolitaine, éventuellement croisé par tranche d'âge et/ou par sexe	*iRRRAAS*	Individu	(4 x) 94	Recensement	Oui	Oui		
4	Taux d'étudiants de 19 à 24 ans par sexe	*etud19S*	Individu	(4 x) 2	EEC 2012	Oui	Oui		
5	Nombre de chômeurs	*nch (1 à 6)*	Individu	6	EEC trimestrielle et enquête non-répondants (ENR)	Oui	Oui	Une valeur par vague	Non, car elle est redondante avec la variable *sexe* x *age* x *acteu*
6	Nombre d'actifs occupés	*nactop (1 à 6)*	Individu	6		Oui	Oui		
7	Nombre de résidences principales	*resp*	Logement	(4 x) 1	Comptes du logement	Oui	Oui	Non	Oui. *Trimestrielle*
8	Nombre de logements	*log*	Logement	1	Comptes du logement	Oui	Oui	Non	Non, car elle est redondante avec la variable *resp* puisqu'on supprime les logements hors champ pour le calage annuel.
9	Décile de revenu du ménage	*nbdXXth*	Logement	(4 x) 10	TH	Oui	Oui : *optionTH* =1 Non : *optionTH* =2	Oui, via le millésime TH	Oui. *Trimestrielle.* Les marges utilisées ne sont pas les marges TH trimestrielles car les logements hors champ ne sont pas pris en compte pour le calage annuel. Les marges sont issues de la base *z* . Ce sont soit celles de la / des vague(s) du sous-échantillon annuel (***optionTH* =1**), soit celles de l'échantillon complet (***optionTH* =2**).
10	Type de logement (maison, appartement)	*nbMA/APth*	Logement	(4 x) 2	TH	Oui			
11	Nombre de pièces du logement	*nbpiNth*	Logement	(4 x) 6	TH	Oui			
12	Statut d'occupation du logement	*nbPROP/LOCth*	Logement	(4 x) 2	TH	Oui			
13	Résidence principale / secondaire	*nbRESP/SECth*	Logement	(4 x) 2	TH	Oui			
14	Zonage en aires urbaines	*nbzauXth*	Logement	(4 x) 2	TH	Oui			
15	Statut HLM du logement	*nbhlmth*	Logement	(4 x) 1	TH	Oui			
16	Nombre de logements construits depuis 2000	*nbneufth*	Logement	(4 x) 1	Comptes du logement	Oui		Non	
17	Sexe x âge x activité	*sexe x age x acteu*	Individu	(4 x) 2x5x3	EEC trimestrielle	Non	Non	sans objet	Oui. *Trimestrielle*
18	Diplôme	*dip11*	Individu	11		Non	Non	sans objet	Oui. *Annuelle*
19	Catégorie socioprofessionnelle	*cstotr*	Individu	9		Non	Non	sans objet	Oui. *Annuelle*

II. Différents scénarios de calage

Plusieurs calages respectant les indications du **tableau 1** ci-dessus ont été effectués. Chacun de ces calages est appelé scénario.

Les scénarios sont présentés au **paragraphe II.1,** leur qualité statistique et leur effet sur les totaux de certaines variables de l'enquête sont ensuite abordés aux **paragraphes II.2 et II.3** respectivement.

1. Description des scénarios

Différents scénarios de calage ont été testés :

- un scénario de base, dit « scénario 1 » ;
- huit autres scénarios, chacun d'eux ayant fait l'objet d'une unique modification, appelée alternative, par rapport au scénario 1. Ces « scénarios alternatifs » sont numérotés de 2 à 8 ;
- deux « scénarios bis », dérivés des scénarios alternatifs 3 et 4 : scénarios 3b et 4b.

Les alternatives testées sont les suivantes :

- *optionTH*[6] = 1 (scénario 1) ou *optionTH* = 2 (scénario 2) ;
- la modalité de la variable CSTOTR (catégorie socioprofessionnelle) qui n'est pas calée est la modalité 5 « Ouvriers » (scénario 1) ou la modalité 6 « Employés » (scénario 3). Dans le scénario 3b, toutes les modalités sont utilisées pour le calage. L'intérêt de cette alternative est lié à la présence de valeurs manquantes pour la variable CSTOTR ;
- la modalité de la variable DIP11 (diplôme le plus élevé obtenu) qui n'est pas calée est la modalité 71 « Sans diplôme » (scénario 1) ou la modalité 50 « CAP, BEP ou équivalents » (scénario 4). Dans le scénario 4b, toutes les modalités sont utilisées pour le calage. L'intérêt de cette alternative est lié à la présence de valeurs manquantes pour la variable DIP11 ;
- la méthode de calage est le « *raking ratio* », méthode 2 de Calmar, sauf pour le scénario 5 : méthode « *logit* », méthode 3 de Calmar, avec pour intervalle « admissible » de rapports de poids [1/3 ; 3] ;
- la version de Calmar utilisée est la première version de Calmar, « Calmar 1 », sauf pour le scénario 6 : Calmar 2 ;
- la « méthode » utilisée, au sens indiqué au **paragraphe I.5**, est la méthode 1 (scénarios 1 à 6), 2 (scénario 7) ou 3 (scénario 8).

2. Examen des scénarios

Le **tableau 2** ci-dessous récapitule les options de chaque scénario (en jaune, apparaît l'option alternative par rapport au scénario 1) et fournit quelques statistiques pour chaque scénario :

- nombre d'itérations de Calmar
- intervalle des rapports de poids
- écart-type des rapports de poids
- poids minimal, maximal et centile 99.

Dans tous les scénarios, la distribution des rapports de poids ne s'écarte pas fortement d'une loi normale, y compris pour le scénario 5 dont la méthode de calage est une méthode *Logit*, ainsi que pour les scénarios de la **partie III**. Dans la suite de ce document, on n'évoquera plus la distribution des rapports de poids au titre des critères de qualité des calages.

Tableau 2 : options et statistiques sur les scénarios

	scen1	scen2	scen3	scen4	scen4b	scen5	scen6	scen7	scen8
Option TH	1	2	1	1	1	1	1	1	1
Modalité de CSTOTR non calée (en plus de la modalité 0)	Ouvriers	Ouvriers	Employés	Ouvriers	Ouvriers	Ouvriers	Ouvriers	Ouvriers	Ouvriers
Modalité de DIP11 non calée	Sans diplôme	Sans diplôme	Sans diplôme	CAP/BEP	Aucune	Sans diplôme	Sans diplôme	Sans diplôme	Sans diplôme
Type de calage	2 (Raking ratio)	2	2	2	2	3 [1/3 ; 3]	2	2	2
Calmar	1	1	1	1	1	1	2	1	1
Méthode	1	1	1	1	1	1	1	2	3
Nombre d'itérations	4	7	4	4	23	6	4	4 et 6	4
Rapport de poids min	0,21	0,20	0,19	0,21	0,05	0,35	0,21	0,21 et 0,096	0,54
Rapport de poids max	3,15	4,73	3,05	3,15	21,6	2,88	3,15	3,15 et 3,25	1,78
Ecart-type des rapports de poids	0,114	0,147	0,116	0,113	0,168	0,114	0,114	0,114 et 0,098	0,046
Poids minimal	53	53	55	53	43	53	53	53 et 90	28
Centile 99 de poids	2 451	2 656	2 452	2 450	2 449	2 453	2 451	2 451 et 2 460	1 214
Poids maximal	28 072	28 470	28 105	28 017	27 992	27 958	28 072	28 072 et 33 208	17 285

On déduit de ce tableau les observations suivantes sur les scénarios :

[6] Le paramètre *optionTH* est présenté au **paragraphe I.5**.

- le **scénario 4b** est manifestement une version dégradée du **scénario 4** : Calmar a besoin de davantage d'itérations pour converger, l'intervalle des rapports de poids est beaucoup plus large, l'écart-type des rapports de poids plus élevé.

  Pour expliquer ce phénomène, il faut avoir à l'esprit que toutes les personnes de 15 ans et plus sont interrogées sur leurs diplômes mais que certaines personnes répondantes à l'enquête n'ont pas répondu à la question : leur variable DIP11 est égale à « blanc ». Le nombre de personnes de 15 ans et plus est un total sur lequel on se cale implicitement, avec la variable aqAAsS. Dès lors, si chacune des modalités (autres que « blanc ») de la variable DIP11 est aussi une variable de calage, le poids total de la modalité « blanc » est déterminé, et égal à la différence entre le nombre de personnes de 15 ans et plus et la somme des marges correspondant aux modalités de la variable DIP11 autres que « blanc ». La modalité « blanc » de la variable DIP11 est donc alors une **modalité de calage implicite**.

  On observe que le nombre d'individus dont la variable DIP11 vaut « blanc » augmente fortement avec le numéro de vague. Dans ce cas, le calage de la vague 1 sur la moyenne des 6 vagues d'enquête oblige à fortement dilater les poids des individus dont la variable DIP11 vaut « blanc ». Le calage est plus difficile, ce qui se traduit par les statistiques de calage médiocres. Quand on supprime une modalité autre que « blanc » du calage, tout se passe comme si on fusionnait la modalité « blanc » avec la modalité non utilisée dans le calage (CAP/BEP pour le scénario 4, d'après le tableau 2), ou comme si on imputait à CAP/BEP les valeurs manquantes, aussi bien dans le sous-échantillon de la vague 6 que pour le calcul des marges.

  Par ailleurs, avec la méthode 3, qui intègre la vague 6 dans le sous-échantillon annuel, l'écart entre la valeur issue du sous-échantillon et la marge calculée à partir des six vagues d'enquête pour la modalité « blanc » est nettement réduite, si bien que le calage sur toutes les modalités de la variable DIP11 ne pose alors pas de problème.

- dans le même ordre d'idée, un **scénario 3b** (non représenté dans le **tableau 2**) a été testé. Toutes les modalités de la variable CSTOTR autres que la modalité 0, qui représente les valeurs manquantes, étaient utilisées pour le calage. Ce calage ne converge pas.

  Le phénomène est le même que pour le scénario 4b. Dès lors qu'on cale sur toutes les modalités de la variable CSTOTR sauf une, alors on cale implicitement sur la modalité manquante. À nouveau, la modalité 0 est de plus en plus fréquente quand le numéro de vague augmente. Cela ne pose à nouveau pas de problème avec la méthode 3, qui utilise les vagues 1 et 6, mais fait échouer le calage avec la méthode 1.

- C'est pourquoi il est nécessaire de supprimer du calage une autre modalité que la modalité 0 de CSTOTR ou que la modalité « blanc » de DIP11. Le choix doit se porter sur des modalités fréquentes, moins susceptibles de poser des problèmes de calage, ce qui est cohérent avec l'interprétation en terme d'imputation implicite mentionnée précédemment. Les modalités choisies ici sont "Employés" pour CSTOTR (**scénario 3**) et "Sans diplôme" pour DIP11 (**scénario 4**).

  Toutefois, au vu des statistiques de calage, il semble que les scénarios 3 et 4 soient très proches du scénario 1, le scénario 4 étant même légèrement meilleur.

- le **scénario 2** est une version dégradée du scénario 1. L'option TH 1, où les marges de calage des variables issues de la taxe d'habitation sont calculées à partir du sous-échantillon annuel, est donc préférable à l'option TH 2, où ces marges sont calculées à partir de l'échantillon complet.

  Cela se comprend bien. Les marges TH ne sont pas les mêmes pour toutes les vagues d'un même trimestre d'EEC. En effet, les marges TH d'une vague sont fonction du millésime du fichier TH dans lequel cette vague a été tirée. Ce fichier TH est actualisé chaque année pour la vague entrante au 4^{e} trimestre. Un trimestre donné, on a donc deux (pour chacun des trois premiers trimestres d'une année) voire trois marges différentes (pour le 4^{e} trimestre) selon les vagues. Avec l'option TH 2, le sous-échantillon annuel n'est pas calé au préalable. Cela complique le calage, d'où un écart-type du rapport de poids plus élevé. On observe qu'avec la méthode 3, les statistiques de cette option sont beaucoup plus proches de celles de l'option TH 1.

- le **scénario 6** donne des résultats identiques au scénario 1 : Calmar2 se comporte exactement comme Calmar.

- le **scénario 5** impose un intervalle de rapport de poids restreint par rapport au scénario 1. Toutefois, cette méthode semble n'avoir qu'une influence mineure sur le calage : ni l'écart-type des rapports de poids ni les poids élevés (centile 99 et maximum) ne sont réduits. Toutefois, la méthode de *raking ratio*, et en particulièrement l'influence de la fenêtre des rapports de poids, fera l'objet de la **partie IV**.

- les scénarios 7 et 8 intègrent les vagues 6 au sein du sous-échantillon annuel : le calage se fait sur les vagues 1 et les vagues 6 de façon séparée (**scénario 7**) ou non (**scénario 8**).

  Dans le scénario 7, le calage de la vague 1 correspond exactement au scénario 1. Le calage de la vague 6 semble selon les critères plus (écart-type du rapport de poids) ou moins (nombre d'itérations, poids minimum) facile que celui de la vague 1.

  Dans le scénario 8, au vu des statistiques sur les rapports de poids, le calage semble beaucoup plus facile que dans le scénario 7. Cela se comprend aisément : la taille double de l'échantillon facilite le calage. Toutefois, le centile 99, au préalable multiplié par deux pour le rendre comparable à ceux des autres scénarios, est à peine inférieur à ceux des autres scénarios. Quant au poids maximal et au nombre d'individus tronqués, ils sont même supérieurs. Il est difficile de dire s'il s'agit d'un phénomène contingent, spécifique à l'échantillon 2013, ou d'un phénomène structurel.

3. L'effet du calage sur quelques variables de l'enquête

Le **tableau 3** ci-dessous est relatif à l'effet du calage sur trois variables de l'enquête non utilisées pour le calage : les variables HALO (halo autour du chômage), SP00 (situation principale le mois d'enquête), CHPUB (nature de l'employeur de la profession principale).

Tableau 3 : effet des scénarios de calage sur quelques variables de l'enquête

	A	B	C	D	E	F	G	H	I	J	K	L
1			En différence par rapport aux poids trimestriels extri									
2		Poids extri	extrid	extridf	scen1	scen2	scen3	scen4	scen4b	scen5	scen7	scen8
3	Total pour HALO	62 278 232	0	0	0	0	0	0	0	0	0	0
4	1. : recherchent un emploi, mais ne sont pas disponibles	295 791	56 008	46 127	44 465	43 922	45 182	44 634	42 702	44 472	37 601	40 839
5	2. : disponibles pour prendre un emploi, mais n'en recherchent pas	325 035	58 601	40 006	49 355	48 300	49 747	49 235	49 367	49 442	34 094	36 795
6	3. : souhaitent un emploi, mais n'en recherchent pas et ne sont pas disponibles	668 230	193 016	107 803	175 315	172 841	175 386	175 525	167 603	175 390	97 594	101 232
7	9. : Sans objet (personnes en emploi, chômeurs ou inactifs hors Halo)	60 989 176	-307 625	-193 936	-269 135	-265 063	-270 315	-269 394	-259 672	-269 303	-169 288	-178 866
8	Total pour SP00	50 627 132	-8 345	3 374	-3 404	-3 404	-3 404	-3 404	-3 404	-3 404	8 327	7 880
9	1. Vous travaillez en tant que salarié (y compris apprentissage ou stage rémunéré)	23 317 385	25 082	16 079	30 433	30 606	33 718	30 837	34 762	30 549	24 205	23 689
10	2. Vous travaillez à votre compte ou en tant qu'aide familial ou conjoint collaborateur	2 238 983	-135 546	-74 210	-88 510	-93 585	-89 048	-88 590	-90 376	-88 562	-62 353	-59 723
11	3. Vous êtes en cours d'études, en stage non rémunéré	4 695 563	-14 708	7 638	26 309	21 413	28 017	26 767	34 387	26 536	9 996	12 653
12	4. Vous êtes au chômage (inscrit ou non à pôle Emploi)	3 636 017	176 570	61 955	63 896	69 178	61 498	63 990	60 927	64 276	26 030	35 300
13	5. Vous êtes retraité ou préretraité	13 224 900	-213 525	-116 617	-282 307	-275 423	-281 886	-281 789	-282 675	-282 728	-116 360	-94 705
14	6. Vous êtes en congé parental à temps plein	203 866	17 909	-9 540	6 512	9 397	6 506	6 618	5 780	6 521	-14 476	-9 973
15	7. Vous êtes homme (femme) au foyer	1 836 523	65 094	74 285	169 533	168 548	166 667	168 981	165 420	169 789	101 625	71 479
16	8. Vous êtes inactif pour cause d'invalidité	1 003 132	67 889	28 017	61 811	56 272	63 772	61 238	61 334	61 567	22 042	14 829
17	9. Vous êtes dans une autre situation	470 764	2 891	15 767	8 920	10 190	7 352	8 544	7 037	8 649	17 617	14 330
18	Total pour CHPUB	22 784 764	49 658	-8 737	53 298	56 227	54 113	53 419	53 713	53 325	2 102	-730
19	1. Entreprise privée ou association	15 661 392	106 998	32 384	144 445	157 123	64 599	144 256	134 480	144 577	39 088	40 940
20	2. Entreprise publique (EDF, La Poste, SNCF, etc.)	962 965	87 759	-7 947	89 814	84 121	98 527	90 044	93 037	89 706	-7 095	-8 745
21	3. État	2 532 923	-173 428	-41 299	-172 970	-174 230	-155 605	-173 326	-175 954	-172 886	-40 186	-42 947
22	4. Collectivités territoriales	1 761 066	72 995	37 976	57 290	56 261	80 589	57 673	53 471	57 128	40 521	39 156
23	5. Hôpitaux publics	952 677	-29 683	-28 978	-45 822	-43 533	-33 983	-45 578	-31 208	-45 824	-26 410	-26 136
24	6. Sécurité sociale	17 614	17 403	1 206	18 644	18 020	18 491	18 624	18 643	18 652	1 938	1 304
25	7. Particulier	896 127	-32 386	-2 080	-38 104	-41 536	-18 505	-38 274	-38 756	-38 027	-5 754	-4 302

Dans ce tableau :

- la colonne B correspond aux moyennes annuelles des totaux trimestriels (poids *extri*)

- les colonnes C et D correspondent aux totaux annuels spontanés du sous-échantillon annuel (*cf*. définition au **paragraphe I.1**), quand celui-ci est constitué de la seule vague 1 (poids *extrid*, colonne C) ou des vagues 1 et 6 (poids *extridf*, colonne D), en différence par rapport à la colonne B. Ces colonnes représentent l'effet spécifique de la vague 1 ou des vagues 1 et 6.

Pour les scénarios 1 à 5, les totaux de référence avant calage annuel sont calculés avec les poids *extrid*, et pour les scénarios 7 et 8, avec les poids *extridf*. Pour les variables d'enquête utilisées lors du calage annuel, les marges de calage correspondent aux moyennes annuelles des totaux trimestriels (poids *extri*).

Pour les autres variables de l'enquête, on s'attend à ce que le calage annuel permette de se rapprocher des moyennes annuelles des totaux trimestriels (colonne B), et donc à ce que les totaux résultant du calage (colonnes **E à L**), calculés en différence par rapport à la colonne B, soient dans l'intervalle compris entre leur valeur annuelle spontanée (colonne C pour les scénarios 0 à 5, colonne D pour les scénarios 7 et 8) et la moyenne annuelle des totaux trimestriels (0, par construction), intervalle appelé intervalle naturel. Idéalement, ces totaux en différence devraient se rapprocher de 0.

Avant calage, les totaux utilisant les vagues 1 et 6 (colonne D) sont plus proches des totaux sur l'échantillon complet que les totaux utilisant la seule vague 1 (colonne C), ce qui est naturel puisque le sous-échantillon est constitué de deux vagues et non plus d'une seule. On s'attend à ce que cette propriété soit conservée après calage annuel, c'est-à-dire que les totaux résultant des scénarios 7 et 8 soient plus proches de 0 que ceux résultant des scénarios 1 à 5.

Tous ces résultats attendus sont bien observés pour la variable HALO Les totaux résultant des scénarios 1 à 5, très proches les uns des autres, sont bien compris entre les totaux utilisant les poids avant calage *extrid* (colonne C) et ceux utilisant les poids *extri* (0, en différence). De même, les totaux résultant des scénarios 7 et 8 sont bien compris entre les résultats de la colonne D et 0. Toutefois, ces totaux sont nettement plus proches des totaux annuels spontanés du sous-échantillon annuel (colonnes C ou D) que des moyennes annuelles des totaux trimestriels (0). Le total pour la variable HALO correspond à la population totale, ce qui explique pourquoi les cellules **C3 à L3** du tableau 3 sont toutes nulles.

Pour la variable CHPUB, la proximité des résultats pour les scénarios 1 à 5 d'une part, 7 et 8 d'autre part, est conservée, hormis pour le scénario 3. Dans ce scénario 3, on cale implicitement la variable « employés + non déclaré » et explicitement la variable « ouvriers » ; dans les autres scénarios, on cale « ouvriers + non déclaré » et « employés ». Comme les « non déclaré » sont sous-représentés dans les sous-échantillons annuels, on augmente leur poids, mais donc aussi ceux de la CS qui leur est implicitement associée : d'où des poids plus forts pour les employés et plus faibles pour les ouvriers dans le scénario 3, si bien que 300 000 ouvriers sont « convertis » en employés. Cela a un effet sur la nature de l'employeur. En revanche, les scénarios fournissent des totaux qui sortent souvent de l'intervalle naturel, en étant souvent « en deçà » de leur valeur annuelle spontanée.

On obtient le même type de résultats pour la variable SP00 : les résultats sortent souvent de l'intervalle naturel, la plupart du temps « en-deçà » de leur valeur annuelle spontanée (modalités 1 ou 7), mais parfois « au-delà » des moyennes annuelles des totaux trimestriels (modalité 3). Le total pour la variable SP00 correspond à la population de 15 ans et plus. On pourrait s'attendre à ce que les cellules **C8 à L8** soient toutes nulles, puisqu'on cale sur les effectifs par tranche d'âge, mais ce n'est pas le cas. En effet, le calage est effectué par rapport à l'âge **en milieu de trimestre** tandis que ce sont les personnes de 15 ans et plus **la semaine de référence** qui sont interrogées pour l'EEC, et en particulier pour la question SP00.

En conclusion, le calage annuel semble avoir un effet assez modeste sur la convergence vers les moyennes annuelles des totaux trimestriels pour les variables autres que celles utilisées pour le calage.

III. Effet sur le calage du nombre de variables de calage

Dans cette partie, il s'agit de mesurer l'influence sur la qualité du calage du nombre de variables de calage.

Une variable de calage supplémentaire est introduite : la variable HALO2, qui correspond à la variable HALO après fusion des modalités 1 à 3, éventuellement croisée avec les variables sexe et tranche d'âge. Cette variable HALO2 est donc simplement l'indicatrice du fait d'appartenir au halo autour du chômage, sans distinction de ses trois catégories constitutives.

Plusieurs variables de contrôle supplémentaires sont examinées : les variables NAFG004N (code NAF de l'employeur en quatre postes), SOUSEMPL (situation de sous-emploi), STATUTR (statut), TPPRED (temps partiel ou non), plus une variable dite « pseudo-DEFM ». Un individu est considéré comme pseudo-DEFM s'il n'est pas un actif occupé (variable *Acteu* différente de 1) et s'il se déclare dans l'enquête comme demandeur d'emploi inscrit auprès de Pôle Emploi, un opérateur de placement ou une association d'insertion (variable *Officc* égale à 1).

1. Les nouveaux scénarios

Sept nouveaux scénarios, distincts de ceux de la **partie II.** et appelés N1 à N7, sont examinés. On passe d'un scénario au suivant en ajoutant une ou plusieurs contraintes de calage (cellules en jaune). Le scénario N2 correspond à peu près au scénario 1 du **paragraphe II.2** (*cf*. infra), le scénario N1 contient une variable de calage en moins, les scénarios N3 à N7 des variables de calage en plus.

Le tableau 4 ci-dessous présente ces nouveaux scénarios, et leurs statistiques.

Tableau 4 : options et statistiques sur les nouveaux scénarios

Scénario	Variable HALO2		Calage sur la variable CSTOTR ?	Nombre de contraintes de calage	Nombre d'itérations	Rapport de poids				Nombre d'individus "tronqués"
	Calage	Cohérence				Ecart-type	Min	C99	Max	
N1	Pas de calage	sans objet	Non	750	4	0,1016	0,193	1,288	3,15	69 517
N2	Pas de calage	sans objet	Oui	757	5	0,1038	0,202	1,296	3,15	70 837
N3	Halo2	Annuelle	Oui	758	4	0,1170	0,217	1,326	3,23	85 285
N4	Halo2 x Sexe	Annuelle	Oui	759	4	0,1171	0,218	1,326	3,23	85 295
N5	Halo2 x Sexe, Halo2 x Age	Annuelle	Oui	763	4	0,1184	0,214	1,334	3,22	88 100
N6	Halo2 x Sexe, Halo2 x Age	Trimestrielle	Oui	781	4	0,1211	0,218	1,347	3,20	84 491
N7	Halo2 x Sexe x Age	Trimestrielle	Oui	797	5	0,1240	0,193	1,353	3,20	77 645

Le scénario N2 ne diffère du scénario 1 du **paragraphe II.2** qu'en raison de corrections apportées à la base *z*. L'écart-type des rapports de poids s'en trouve réduit de 0,114 à 0,104.

De façon attendue, l'écart-type des rapports de poids augmente avec le nombre de contraintes de calage. L'augmentation la plus importante a lieu au moment de l'introduction de la variable HALO2 parmi les variables de calage (scénario N3), bien que cela ne crée qu'une contrainte supplémentaire. Cela s'explique par le fort biais de rotation pour cette variable : le nombre d'individus dans le halo varie structurellement avec la vague d'enquête. Malgré tout, les statistiques de calage du scénario N7 restent tout à fait acceptables.

2. Effet des nouveaux scénarios de calage sur les variables de contrôle

Le tableau 5 ci-dessous présente l'effet des nouveaux scénarios sur les variables de contrôle.

Pour chaque modalité des variables de contrôle :

- la colonne B correspond à la moyenne annuelle des totaux trimestriels (poids *extri*) ;
- les colonnes C et D correspondent à la « distance » entre le total annuel spontané du sous-échantillon annuel (*cf*. définition au **paragraphe I.1**) et la colonne B, mesurée en différence (colonne C) ou en écart relatif (colonne D) ;
- pour chacun des **scénarios N1 à N7**, les totaux après calage ont été transformés par une fonction affine, dont le résultat, appelé « position », apparaît aux colonnes E à K. Cette fonction affine est déterminée par les deux contraintes suivantes : si le total après calage est

égal au total avant calage, la position est 0 ; si ce total après calage correspond au chiffre de la colonne B, la position est 100 %. Ainsi, l'intervalle naturel (*cf.* **paragraphe II.2**), après transformation affine, est l'intervalle de position [0 ; 100 %]. Un total « en-deçà » de sa valeur annuelle spontanée correspond à une position négative ; un total « au-delà » de la moyenne annuelle des totaux trimestriels correspond à une position supérieure à 100 %.

Les cellules en rose sont celles qui sortent de l'intervalle [- 50 % ; 200 %] ; les cellules en jaune sont celles qui sortent de l'intervalle [0 % ; 120 %], tout en restant dans l'intervalle [- 50 % ; 200 %].

Tableau 5 : effet des nouveaux scénarios sur les variables de contrôle

	A	B	C	D	E	F	G	H	I	J	K
1		Poids extri	extrid-extri	extrid/extri-1	Position après calage selon le scénario						
2					N1	N2	N3	N4	N5	N6	N7
3	Total pour **CSTOTR** (Nombre de 15 ans et plus la semaine de référence)	50 623 728	-4 941	0,0%	100%	100%	100%	100%	100%	100%	100%
4	0. Non renseigné	49 149	-39 642	-80,7%	0%	0%	-1%	-1%	0%	-1%	-1%
5	1. Agriculteurs exploitants	523 375	-16 046	-3,1%	-29%	100%	100%	100%	100%	100%	100%
6	2. Artisans, commerçants, et chefs d'entreprise	1 680 747	-59 845	-3,6%	16%	100%	100%	100%	100%	100%	100%
7	3. Cadres et professions intellectuelles supérieures	4 688 327	-104 662	-2,2%	-51%	100%	100%	100%	100%	100%	100%
8	4. Professions intermédiaires	6 856 173	95 363	1,4%	65%	100%	100%	100%	100%	100%	100%
9	5. Employés	8 110 277	51 684	0,6%	-31%	100%	100%	100%	100%	100%	100%
10	6. Ouvriers	6 220 154	80 333	1,3%	-99%	51%	50%	50%	50%	50%	50%
11	7. Retraités	16 129 556	31 542	0,2%	99%	100%	100%	100%	100%	100%	100%
12	8. Autres personnes sans activité professionnelle	6 365 971	-43 669	-0,7%	116%	100%	100%	100%	100%	100%	100%
13	Total pour **HALO** (Nombre d'individus)	62 278 232	0	0,0%							
14	1. Recherchent un emploi, mais ne sont pas disponibles	295 791	56 008	18,9%	1%	0%	121%	121%	124%	122%	119%
15	2. Disponibles pour prendre un emploi, mais n'en recherchent pas	325 035	58 601	18,0%	3%	3%	129%	127%	122%	102%	101%
16	3. Souhaitent un emploi, mais n'en recherchent pas et ne sont pas disponibles	668 230	193 016	28,9%	0%	1%	85%	86%	86%	93%	94%
17	9. Sans objet (personnes en emploi, chômeurs ou inactifs hors halo)	60 989 176	-307 625	-0,5%	1%	1%	100%	100%	100%	100%	100%
18	**Pseudo-DEFM**	3 318 124	225 979	6,8%	39%	39%	85%	85%	83%	82%	83%
19	Total pour **SOUSEMPL**	1 678 893	217 975	13,0%	-4%	-1%	-1%	-1%	-2%	-1%	-1%
20	1. Temps partiel, souhait de travailler plus d'heures, disponible pour le faire et à la recherche d'un autre emploi	359 315	51 839	14,4%	-12%	-7%	-10%	-10%	-11%	-9%	-9%
21	2. Temps partiel, souhait de travailler plus d'heures, disponible et ne recherchant pas d'emploi	1 182 700	156 835	13,3%	-1%	1%	2%	2%	1%	2%	1%
22	3. Temps plein, ou temps partiel (autre que les deux cas ci-dessus), ayant involontairement moins travaillé que d'habitude (pour chômage partiel ou intempéries)	136 879	9 301	6,8%	-14%	-17%	-3%	-3%	-6%	-1%	5%
23	Total pour **TPPRED** (Nombre d'actifs occupés)	25 763 495	-65 899	-0,3%	100%	100%	100%	100%	100%	100%	100%
24	1. Temps complet	21 016 625	-269 117	-1,3%	21%	23%	27%	27%	26%	27%	26%
25	2. Temps partiel	4 746 870	203 218	4,3%	-5%	-2%	3%	3%	2%	3%	2%
26	Total pour **STATUTR** (Nombre d'actifs occupés)	25 763 495	-65 899	-0,3%	100%	100%	100%	100%	100%	100%	100%
27	1. Non salariés (indépendants, employeurs)	2 894 228	-86 166	-3,0%	-6%	86%	85%	85%	85%	85%	86%
28	2. Intérimaires	508 290	44 209	8,7%	-33%	-14%	-12%	-12%	-13%	-11%	-13%
29	3. Apprentis	406 319	-33 870	-8,3%	-12%	-14%	-22%	-22%	-23%	-20%	-18%
30	4. CDD	2 064 265	146 405	7,1%	-7%	-4%	-5%	-5%	-6%	-7%	-8%
31	5. CDI	19 889 121	-136 188	-0,7%	37%	-11%	-9%	-10%	-10%	-11%	-14%
32	Non renseigné (actif occupé)	1 272	-290	-22,8%	15%	31%	33%	33%	30%	31%	31%
33	Total pour **NAFG004N** (Nombre d'actifs occupés)	25 763 495	-65 899	-0,3%	100%	100%	100%	100%	100%	100%	100%
34	ES. Agriculture	781 720	-31 530	-4,0%	-4%	48%	48%	48%	48%	49%	48%
35	ET. Industrie	3 397 620	62 576	1,8%	-18%	9%	6%	6%	7%	8%	8%
36	EU. Construction	1 676 972	-15 155	-0,9%	116%	102%	107%	105%	102%	98%	93%
37	EV. Tertiaire	19 365 627	294 766	1,5%	-12%	-13%	-12%	-12%	-12%	-13%	-13%
38	00. Non renseigné	541 556	-376 557	-69,5%	1%	1%	1%	1%	1%	1%	1%
39	Total pour **SP00** (Nombre de 15 ans et plus)	50 623 336	-4 548	0,0%	109%	109%	109%	109%	109%	109%	109%
40	1. Vous travaillez en tant que salarié (y compris apprentissage ou stage rémunéré)	23 315 511	26 957	0,1%	-209%	6%	-16%	-16%	-17%	-21%	-17%
41	2. Vous travaillez à votre compte ou en tant qu'aide familial ou conjoint collaborateur	2 238 983	-135 546	-6,1%	-1%	41%	40%	40%	40%	40%	40%
42	3. Vous êtes en cours d'études, en stage non rémunéré	4 695 130	-14 275	-0,3%	194%	202%	384%	373%	426%	429%	436%
43	4. Vous êtes au chômage (inscrit ou non à pôle Emploi)	3 635 159	177 428	4,9%	49%	49%	118%	118%	115%	114%	113%
44	5. Vous êtes retraité ou préretraité	13 224 461	-213 087	-1,6%	0%	0%	18%	17%	15%	14%	15%
45	6. Vous êtes en congé parental à temps plein	203 866	17 909	8,8%	17%	17%	-14%	-15%	-24%	-20%	-20%
46	7. Vous êtes homme (femme) au foyer	1 836 523	65 094	3,5%	-39%	-38%	-67%	-71%	-64%	-64%	-56%
47	8. Vous êtes inactif pour cause d'invalidité	1 003 132	67 889	6,8%	23%	19%	-32%	-29%	-22%	-20%	-22%
48	9. Vous êtes dans une autre situation	470 573	3 082	0,7%	-50%	-26%	117%	108%	116%	112%	98%
49	Total pour **CHPUB**	22 784 764	49 658	0,2%	-137%	17%	16%	16%	16%	17%	19%
50	1. Entreprise privée ou association	15 661 392	106 998	0,7%	-56%	4%	6%	6%	6%	7%	8%
51	2. Entreprise publique (EDF, La Poste, SNCF, etc.)	962 965	87 759	9,1%	-7%	-6%	-8%	-8%	-7%	-7%	-7%
52	3. État	2 532 923	-173 428	-6,8%	2%	4%	6%	6%	6%	7%	6%
53	4. Collectivités territoriales	1 761 066	72 995	4,1%	-1%	12%	17%	16%	16%	17%	17%
54	5. Hôpitaux publics	952 677	-29 683	-3,1%	-43%	-43%	-35%	-35%	-34%	-34%	-33%
55	6. Sécurité sociale	17 614	17 403	98,8%	-5%	-8%	-9%	-9%	-8%	-9%	-9%
56	7. Particulier	896 127	-32 386	-3,6%	33%	7%	6%	7%	6%	8%	9%

Pour certaines variables de contrôle, les résultats sont conformes aux attentes : les résultats sont dans l'intervalle naturel et la position se rapproche de 100 % quand le nombre de contraintes de

calage augmente. C'est le cas notamment pour la variable « pseudo-DEFM » quand on introduit la variable HALO2 au sein des variables de calage. C'est aussi le cas pour la variable HALO : le renforcement du calage sur la variable HALO2 (scénarios N3 à N7) permet aux modalités 1 à 3 de se rapprocher de la position 100 %.

Mais pour la plupart des variables de contrôle, la position après calage est très éloignée de 100 %, et peut même sortir de l'intervalle naturel, sans que la hausse du nombre de contraintes de calage n'y change grand chose. C'est le cas pour les variables SOUSEMPL et TPPRED, ainsi que pour la variable CSTOTR dans le scénario N1 (cette variable devient une variable de calage dans les autres scénarios).

Pour certaines modalités, la position s'éloigne même de 100 % au fur et à mesure qu'on introduit des variables de calage. C'est le cas pour la modalité 3 de la variable SP00. Ce constat doit être modéré puisque l'intervalle naturel est étroit : le total spontané n'est qu'à 0,3 % de la moyenne annuelle des totaux trimestriels.

On peut chercher à comparer les scénarios de calage au vu de leur effet sur les positions des différentes variables de contrôle. On observe que le scénario 2 est meilleur que le scénario 1 au vu des variables NAFG004N, STATUTR, et à un degré moindre SP00. Cela incite à conserver la variable CSTOTR parmi les variables de calage.

Il est moins évident de classer les autres scénarios. L'introduction de la variable HALO2 parmi les variables de calage améliore la position de la variable « pseudo-DEFM », voire celle de TPPRED, mais détériore globalement celle de SP00. Enfin, les scénarios N3 à N7 semblent assez comparables, hormis, comme indiqué plus haut, pour la variable HALO.

En conclusion, l'écart-type des rapports de poids augmente avec le nombre de contraintes de calage. Malgré tout, dans l'exemple traité, les statistiques de calage restent à tout moment très correctes. Dès lors, on choisira plutôt parmi les scénarios en fonction des variables qu'on souhaite vraiment caler.

IV. Effet sur le calage de la fenêtre des rapports de poids dans la méthode *Logit*

Dans cette partie, on cherche à mesurer l'influence sur la qualité du calage de la fenêtre des rapports de poids dans la méthode *Logit*.

Le scénario N2 est repris de la **partie III**. Les différents scénarios du **tableau 6** ne se distinguent entre eux que par la méthode de calage utilisée : méthode du *Raking ratio* pour le scénario N2 contre méthode *Logit* pour les scénarios L1 à L12. Pour chacun des scénarios utilisant la méthode *Logit*, la borne *Low* de la fenêtre des rapports de poids a été choisie égale à l'inverse de la borne *Up*.

Tableau 6 : statistiques de calage selon la fenêtre des rapports de poids

Scénario	Méthode	Fenêtre des rapports de poids		Nombre d'itérations	Statistiques de rapport de poids			Remarque
		Low	*Up*		Écart-type	Min	Max	
N2	*Raking ratio*	sans objet		5	0,1038	0,2018	3,149	
L1	*Logit*	0,1000	10,000	5	0,1038	0,2274	3,114	
L2	*Logit*	0,1667	6,000	5	0,1038	0,2509	3,077	
L3	*Logit*	0,2500	4,000	5	0,1038	0,2903	3,007	
L4	*Logit*	0,3333	3,000	6	0,1037	0,3446	2,882	
L5	*Logit*	0,3448	2,900	7	0,1037	0,3534	2,857	
L6	*Logit*	0,3509	2,850	9	0,1037	0,3525	2,843	
L7	*Logit*	0,3521	2,840	10	0,1037	0,3521	2,840	
L8	*Logit*	0,3522	2,839	9	0,1037	0,3522	2,839	
L9	*Logit*	0,3571	2,800	8	0,1037	0,3571	2,800	Non convergence de la méthode avec Calmar
L10	*Logit*	0,3676	2,720	8	0,1036	0,3676	2,720	
L11	*Logit*	0,3690	2,710	8	0,1036	0,3690	2,710	
L12	*Logit*	0,3704	2,700	Non convergence de la méthode avec Calmar 2				

Du scénario N2 au scénario L12, la contrainte de calage imposée par la fenêtre des rapports de poids est de plus en plus forte. À partir du scénario L7, chaque borne de la fenêtre est atteinte pour une observation au moins. Pour les scénarios L9 à L11, Calmar ne converge plus alors que Calmar2 continue de converger. Cela s'explique par l'utilisation par Calmar2 de matrices inverses généralisées, quand Calmar ne connaît que les inverses "classiques". Pour le scénario L12, Calmar2 ne converge plus non plus

Étonnamment, le nombre d'itérations n'est pas monotone : il augmente entre N2 et L7 où il atteint son maximum, avant de diminuer.

Dans le scénario L1, la fenêtre des rapports de poids est suffisamment large pour contenir tous les rapports de poids de la méthode N2 non contrainte. On observe néanmoins un resserrement des rapports de poids par rapport à cette méthode N2, visible à la fois sur les extremums et sur l'écart-type des rapports de poids.

L'intervalle des rapports de poids décroît, au sens de l'inclusion d'ensemble, avec la fenêtre des rapports de poids, à part le poids minimum qui diminue entre les scénarios L5 et L7. L'écart-type décroît également, mais de façon légère. À aucun moment, on n'observe que la loi de distribution s'écarte d'une loi normale pour se rapprocher d'une loi "en U", c'est-à-dire une loi avec deux pics marqués, à chacune des deux bornes.

Le **tableau 7** ci-dessous s'intéresse à la distribution des rapports de poids calés entre le scénario N2 et le scénario L11.

Tableau 7 : distribution des rapports de poids entre le scénario N2 et le scénario L11

Plus petit	2e plus petit	1er centile	1er décile	9e décile	99e centile	2e plus grand	Plus grand
0,837	0,844	0,995	0,999	1,001	1,006	1,113	1,849

La quasi-totalité de ces rapports de poids est comprise entre 0,99 et 1,01. Par rapport au *Raking ratio*, la méthode *Logit* a donc, dans ce cas, peu d'influence sur les poids individuels, même quand la fenêtre est réduite au maximum, hormis pour un petit nombre d'observations sur lesquelles porte la contrainte.

Si on regardait l'impact de ces scénarios sur les variables de contrôle, comme dans le **tableau 5** plus haut, on verrait que, de façon cohérente avec l'observation précédente, la position après calage des variables dans tous les scénarios L1 à L11 est très voisine de leur position dans le scénario N2.

En conclusion, dans l'exemple ci-dessus, on observe que la méthode *Logit* a peu d'effet sur les variables de contrôle et sur la variance des rapports de poids. On préférera conserver la méthode *Raking ratio*, pour sa simplicité.

Bibliographie

[1] Sautory Olivier « La macro CALMAR Redressement d'un échantillon par calage sur marges », Insee, Document de travail n°F9310, 25 novembre 1993.