LES PROBLÈMES DE CALAGE DANS LES ENQUÊTES ENTREPRISES

Nathalie CARON

Insee, Direction des statistiques d'entreprises

Introduction

Lorsqu'on réalise une enquête auprès des ménages ou des entreprises, il est rare de ne disposer d'aucune information (extérieure à l'enquête) sur les unités appartenant au champ de l'enquête. Cette information, appelée information auxiliaire, peut soit être connue à un niveau global, sous forme d'un total par exemple, pour la totalité du champ de l'enquête ou sur une partie, soit être disponible au niveau individuel pour tous les individus de la base de sondage. Il est important d'en tenir compte pour améliorer la précision des statistiques obtenues à partir du fichier de diffusion.

Selon le degré de finesse et les délais de mise à disposition de l'information, celle-ci est mobilisée au niveau de la conception du plan de sondage ou au niveau de l'expression finale de l'estimateur après réalisation de l'enquête. Dans ce dernier cas, c'est ce que l'on appelle le redressement de l'enquête. L'objectif du redressement consiste à « construire » une forme d'estimateur permettant de faire coïncider le total d'une variable estimé à partir de l'échantillon avec celui provenant d'une autre source considérée comme certaine. Ce nouvel estimateur s'obtient en modifiant les pondérations initiales des unités sélectionnées dans l'échantillon, qui sont fonction du plan de sondage réalisé. Les nouveaux poids sont aussi proches que possible des poids initiaux, au sens d'une distance choisie au préalable, tout en respectant un ensemble de contraintes appelées équations de calage : la somme pondérée des valeurs de chaque variable auxiliaire obtenue sur les individus de l'échantillon correspond au total connu de cette variable auxiliaire. Il existe plusieurs méthodes de redressement. Les plus simples sont l'estimateur post-stratifié, l'estimateur par le ratio et l'estimateur par régression. Toutes ces méthodes peuvent être considérées comme des cas particuliers d'une famille de méthodes plus générale : les méthodes de calage. De façon générale, ces méthodes peuvent utiliser des informations auxiliaires provenant aussi bien de variables numériques que qualitatives et peuvent être appliquées avec la macro SAS CALMAR, développée à l'Insee.

Largement utilisées pour les enquêtes réalisées auprès de ménages, les méthodes de calage ne le sont qu'occasionnellement pour celles réalisées auprès des entreprises, en particulier à la direction des statistiques d'entreprises de l'Insee. A première vue, cette situation semble paradoxale. En réalité, elle s'explique par des difficultés supplémentaires propres à la méthodologie des enquêtes « entreprises »

qui apparaissent lors de la mise au point du calage. En effet, la réalisation d'un calage ne se résume pas à savoir utiliser la macro CALMAR. Le statisticien en charge de cette opération doit faire un certain nombre de choix : choix des marges utilisées, choix de la méthode de calage, etc. Pour les enquêtes entreprises, le choix d'une bonne source d'information auxiliaire est plus difficile que pour les enquêtes ménages, compte tenu des nombreux liens existant entre les sources relatives aux entreprises. Or, utiliser comme données auxiliaires des variables mises à jour par les résultats d'une enquête par sondage que l'on cherche à caler est source de biais, et est par conséquent à déconseiller.

Dans cet article, nous présentons dans une première partie les principaux problèmes auxquels un statisticien se trouve confronté pour effectuer un calage, puis les problèmes spécifiques supplémentaires liés aux données entreprises. Ces problèmes méthodologiques et les choix réalisés seront illustrés dans le cas de la mise en œuvre des techniques de calage sur la principale enquête structurelle réalisée auprès des entreprises : l'enquête annuelle d'entreprise (EAE). Plus précisément, nous avons utilisé les données de l'EAE-commerce réalisée en 2000 sur l'exercice 1999 auprès de 60 000 entreprises. Ajoutons que les résultats obtenus ne l'ont été qu'à titre d'expérimentation méthodologique et ne constituent pas à ce titre des éléments de référence pour le secteur du commerce. Pour évaluer l'apport du calage en termes de meilleure appréhension économique du secteur, l'analyse d'un expert du domaine est indispensable. Les aspects mathématiques de la théorie du calage ne sont pas rappelés dans ce papier (le lecteur dispose d'un résumé en annexe et pour plus de détails nous l'invitons à se reporter aux références bibliographiques [4], [5] et [6]).

1. Les problèmes classiques

Réaliser un calage ne se résume pas à savoir utiliser la macro SAS CALMAR. Cela nécessite que le statisticien en charge de cette opération fasse des choix méthodologiques. Il doit en particulier décider de la source auxiliaire, des variables de calage et de leur nombre ainsi que de la méthode de calage utilisée.

1.1 Choix de la source auxiliaire

Pour faire un calage, il est tout d'abord indispensable de disposer d'information auxiliaire. Par conséquent, la première difficulté consiste à déterminer une « bonne » source auxiliaire et à l'utiliser correctement.

Les résultats théoriques sur le calage et sur le redressement des enquêtes en général sont établis en considérant que la source auxiliaire ainsi que les marges utilisées sont exactes. Or, il est bien rare d'avoir une source auxiliaire « certaine » à notre disposition. On se contente, en général, d'utiliser des marges nettement plus précises que celles obtenues directement à partir de l'enquête que l'on cherche à redresser. Il est par conséquent inutile de caler les résultats d'une enquête sur ceux d'une autre enquête conduisant à une précision inférieure ou égale ou même juste meilleure. En pratique, on recommande l'utilisation d'une enquête comme source auxiliaire lorsque la taille de son échantillon est approximativement 10 fois plus grande que celle de l'enquête à caler. A titre d'exemple, à l'Insee, c'est en général l'enquête Emploi (réalisée auprès de 80000 ménages environ) qui sert de source de calage pour les enquêtes ménages (dont la taille moyenne de l'échantillon est de 10000). La seule enquête ménages qui n'utilise pas cette source auxiliaire est l'enquête Logement réalisée auprès de 40000 ménages.

1.2 Choix des variables de calage

Une fois la source auxiliaire déterminée, il faut choisir les variables de calage. La théorie nous indique que le gain en termes d'amélioration de la précision sera d'autant plus important que la variable d'intérêt est corrélée aux variables auxiliaires utilisées pour le calage. Cet élément théorique permet de guider le statisticien dans le choix des variables de calage. Celui-ci doit ensuite se demander jusqu'à quel niveau de finesse il souhaite réaliser son calage. Par exemple, si pour une enquête ménage, on retient comme information auxiliaire la répartition des ménages selon leur taille issue de l'enquête Emploi, il faut se fixer le nombre de modalités de la variable « taille ». Ce choix détermine le nombre de contraintes de calage à respecter ce qui influe directement sur la procédure de calage. En effet, lorsque le nombre de contraintes est important, il est en général plus difficile de réaliser un calage.

Le statisticien doit ensuite veiller à assurer une cohérence des variables de calage entre le fichier de l'enquête et la source externe. Autrement dit, ce doit être les mêmes variables, c'est-à-dire les mêmes concepts recueillis à une même date, qui sont utilisées à la fois pour constituer les marges et dans le fichier d'enquête « à caler ». Si la source auxiliaire est exhaustive, cette condition est automatiquement vérifiée à condition d'utiliser pour les individus de l'échantillon la valeur de la variable disponible dans la source auxiliaire. Prenons un exemple simple concernant une enquête « ménages » réalisée en 2000. On supposera que l'information auxiliaire est issue du recensement de la population de 1999, par exemple le nombre total d'individus en 1999. Sous cette hypothèse, il faudra utiliser pour le calage comme variable dans le fichier d'enquête le nombre d'individus par ménage connu au moment du recensement de 1999, et surtout pas le nombre d'individus par ménage recueilli au moment de la réalisation de l'enquête. De façon plus générale, pour constituer le total connu de la source auxiliaire, on ne doit pas utiliser une variable mise à jour dans la base de sondage par les résultats de l'enquête que l'on cherche à redresser par calage (ni par ceux d'une autre enquête dont l'échantillon lui est coordonné). En effet, pour une enquête donnée, on ne peut modifier la base de sondage que pour les unités sélectionnées par sondage dans l'échantillon. Or, ces unités représentent un certain nombre d'unités de leur secteur. Par conséquent, toute modification constatée sur une unité enquêtée ayant une certaine pondération devrait se traduire, au niveau de la base de sondage, par autant de modifications que la pondération associée à cette unité, ce qui est impossible. L'utilisation de variables auxiliaires mises à jour de cette manière conduit à des estimations biaisées.

Le choix des marges et des variables de calage peut parfois conduire à utiliser comme fichier d'entrée du logiciel CALMAR un fichier différent de celui des répondants de l'enquête que l'on cherche à caler. Prenons un exemple pour illustrer ces propos. Pour l'enquête logement réalisée par l'Insee en 2001, la source auxiliaire utilisée est le recensement de la population de 1999. Or, entre 1999 et 2001, la population de logements a évolué : disparition de logements, création de logements neufs, changement de catégorie (un logement principal en 1999 peut être inoccupé en 2001). Afin de respecter le principe de cohérence évoqué ci-dessus, il faut alors redonner à tous les logements sélectionnés dans l'échantillon les valeurs des variables recueillies au recensement de 99. Ainsi, par exemple, si on souhaite caler l'enquête logement sur le nombre total de logements principaux en 1999, le fichier en entrée du logiciel CALMAR contient l'ensemble des logements de l'échantillon qui étaient déclarés principaux en 1999. Par définition, ce fichier est différent de celui d'exploitation de l'enquête qui contient l'ensemble des logements répondants déclarés principaux en 2001.

Quel que soit le soin apporté lors du choix des marges, un seul calage ne constitue pas une solution optimale pour toutes les études envisagées à partir d'un fichier. En effet, dans le cadre d'une enquête entreprises, une procédure de calage sur le nombre d'entreprises par secteur efficace au niveau national peut perdre beaucoup de son intérêt lorsqu'on se restreint aux données d'une région. Dans ce cas, il est préférable d'utiliser la répartition sectorielle des entreprises au niveau régional qui est rarement identique à celle au niveau national. Pour chaque étude, l'idéal est de rechercher s'il existe une information auxiliaire adéquate afin de l'utiliser dans la procédure de calage. Ce type d'information

n'est pas toujours disponible à un niveau fin. Ajoutons que réaliser différents calages pour une étude nationale et pour des études régionales risque de poser des problèmes de réconciliation de résultats.

1.3 Choix de la méthode de calage

Après avoir déterminé les marges, il est nécessaire de choisir la méthode de calage. En effet, la macro SAS CALMAR[1] permet d'utiliser quatre méthodes de calage : la méthode linéaire, la méthode du raking-ratio, la méthode logit et la méthode linéaire tronquée.

Le choix de la méthode peut être considéré comme un faux problème. En effet, d'un point de vue théorique, toutes les méthodes sont asymptotiquement équivalentes au sens où elles conduisent à des estimateurs ayant la même précision. Cependant, du point de vue de l'utilisateur, le choix de la méthode n'est pas sans conséquence. En effet, la méthode linéaire peut dans certains cas conduire à des pondérations négatives, ce qui n'est pas forcément très simple à expliquer aux utilisateurs du fichier. Le statisticien doit décider du choix de la méthode selon son objectif : minimiser la variance des pondérations finales, minimiser l'étendue de la distribution des pondérations finales ou encore obtenir une « bonne » allure générale de cette distribution.

2. Les problèmes spécifiques aux enquêtes entreprises

Pour appliquer les méthodes de calage à une enquête « entreprises » réalisée à l'Insee, le statisticien en charge de cette opération se trouve confronté à des problèmes supplémentaires.

2.1 Un choix de variables auxiliaires plus délicat

Pour les enquêtes «entreprises », le choix de l'information auxiliaire est plus délicat que pour les enquêtes « ménages » car de nombreux liens existent entre les diverses sources relatives aux entreprises. En particulier, l'enquête annuelle d'entreprises (EAE), principale source d'information sur l'activité économique des entreprises, notamment le chiffre d'affaires et sa répartition en branches, la valeur ajoutée, l'investissement et l'emploi non salarié, a des relations privilégiées avec le répertoire Sirene. En effet, ses résultats servent de référence pour le code de l'activité principale exercée (APE) et permettent sa mise à jour de façon quasi-automatique dans le répertoire Sirene. Ces liens rendent particulièrement difficile l'application de méthodes de calage.

En effet, comme nous l'avons déjà signalé ci-dessus, utiliser comme données auxiliaires des variables mises à jour par les résultats d'une enquête par sondage que l'on cherche à caler est source de biais et est par conséquent à déconseiller. Prenons un exemple pour illustrer ce problème. Considérons une base de sondage contenant 1000 entreprises identifiées comme ayant la même activité principale exercée (qui sera notée A par la suite). On suppose que l'on sélectionne dans cette base de sondage 100 entreprises par un sondage aléatoire simple. Lors de la réalisation de l'enquête, on constate que sur ces 100 entreprises, 50 ont toujours la même activité A alors que les 50 autres ont changé d'activité, celle-ci sera notée B. La mise à jour de l'activité dans la base de sondage à partir des résultats de l'enquête conduit donc à avoir 50 entreprises dont l'activité est B et 950 dont l'activité reste A. Alors que la répartition des entreprises selon l'activité est correctement estimée à partir de l'échantillon, celle obtenue dans la base de sondage actualisée est certainement « fausse » même si elle est « meilleure »

[1] La seconde version de CALMAR propose une méthode de calage supplémentaire (Voir [3]).

que celle obtenue à partir de la base initiale. Caler les résultats de cette enquête sur la répartition des entreprises selon l'activité connue dans la base de sondage actualisée n'a alors aucun sens. Dans l'idéal, pour une enquête «entreprises» donnée, il faut rechercher une variable auxiliaire disponible dans une source qui n'est pas mise à jour par les échantillons coordonnés à cette enquête, y compris l'échantillon de l'enquête lui-même s'il s'agit d'un panel. Pour les enquêtes « ménages », ce problème ne se pose pas car il n'y a aucun lien entre l'enquête Emploi et les autres enquêtes « ménages ».

2.2 Le traitement de la partie exhaustive de l'enquête

Le plan de sondage des enquêtes «entreprises» est en général un plan de sondage stratifié où les strates correspondent à un croisement de l'activité principale de l'entreprise (APE) avec la taille de l'entreprise exprimée en tranches d'effectifs salariés. Dans chaque strate, on réalise un sondage aléatoire simple. Le taux de sondage est très variable selon les strates. En principe, les grandes entreprises en termes d'effectifs salariés sont automatiquement enquêtées : elles appartiennent à la partie exhaustive. Pour les enquêtes ménages, les plans de sondage sont en général plus complexes (stratifiés à plusieurs degrés) et il n'existe aucune partie exhaustive.

Les unités de la partie exhaustive sont à examiner avec soin. En effet, si on effectue un calage sur un fichier global y compris la partie exhaustive, la procédure de calage réalisée conduit à **modifier à la hausse ou à la baisse tous les poids des entreprises y compris ceux de la partie exhaustive.** Après calage, le poids final de certaines de ces entreprises peut être inférieur à 1. Ceci peut paraître surprenant car dans ce cas, la modification des poids de ces entreprises ne peut être interprétée ni comme une correction des fluctuations d'échantillonnage ni comme une correction de la non-réponse. Pour éviter ce problème, une solution consisterait à « forcer » dans la procédure de calage la pondération finale de ces entreprises à 1. Autrement dit, cela revient à leur donner une fonction de calage différente de celle des autres entreprises. Cependant, cette possibilité n'existe pas directement dans la version du logiciel CALMAR disponible à ce jour. En pratique, une solution consiste à ne réaliser le calage que sur la partie échantillonnée. Une autre façon de procéder consisterait à créer dans le fichier autant de variables indicatrices que d'entreprises de la partie exhaustive et à réaliser le calage en y ajoutant les marges correspondantes à ces variables.

3. Application à l'EAE-Commerce

A titre d'expérimentation méthodologique, nous avons testé les méthodes de calage sur la principale enquête structurelle réalisée auprès des entreprises : l'enquête annuelle d'entreprise (EAE). Plus précisément, nous avons utilisé les données de l'EAE-commerce sur l'exercice 1999. Dans cette partie, nous présentons les choix que nous avons retenus pour le calage de cette enquête.

3.1 Présentation des enquêtes annuelles d'entreprise

Les EAE sont réalisées dans six grands secteurs : l'industrie manufacturière, les industries agricoles et alimentaires, le commerce, les services, la construction et les transports. Les secteurs du commerce et des services sont enquêtés par l'Insee, les autres par les services statistiques des ministères. Ces enquêtes qui concernent annuellement un échantillon de 240 000 entreprises permettent d'obtenir des données structurelles de référence sur les différents secteurs d'activité. Le champ et la sélection des unités concernées par ces enquêtes sont spécifiques à chaque secteur. Dans l'industrie et les industries agroalimentaires, l'enquête est réalisée sur le champ des entreprises de 20 salariés et plus et est exhaustive. Dans les autres secteurs, elle concerne toutes les entreprises et est exhaustive au-dessus d'un seuil déterminé en fonction du nombre de salariés et réalisée par sondage en dessous de ce seuil.

En ce qui concerne l'EAE - secteur commerce, la taille de l'échantillon est d'environ 65 000 pour une population de 650 000. Les grandes entreprises en terme d'effectifs salariés ou de chiffre d'affaires reçoivent systématiquement un questionnaire chaque année ; elles appartiennent à la partie exhaustive. Pour les autres entreprises (c'est-à-dire les petites), un sondage est effectué. Les entreprises sont alors sélectionnées par un plan de sondage stratifié où les strates sont définies par le croisement de l'activité exprimée en classe NAF et de l'effectif exprimé en tranche. De plus, afin de privilégier les estimations en évolution, l'échantillon de l'EAE-Commerce est renouvelé au quart chaque année. Par conséquent, une petite entreprise qui rentre dans l'échantillon y figure au moins pendant quatre ans à moins qu'elle n'ait cessé son activité.

3.2 Les variables de calage retenues

Nous avons examiné plusieurs variables de calage possibles en nous attachant plus particulièrement à leur procédure de mise à jour afin de comprendre si leur utilisation conduirait ou non à des estimateurs biaisés. La liste exhaustive des variables étudiées est décrite dans le document [2]. Cette étude nous a permis de retenir les marges suivantes :

- le nombre d'entreprises actives au 31/12/1999 selon l'activité principale exercée (APE), ce qui conduit dans notre cas à 116 totaux.
- les effectifs totaux en termes de salariés au 31/12/1999 selon l'APE en classe NAF, ce qui conduit aussi à 116 totaux.

Ces marges sont constituées à partir de trois variables disponibles au niveau de l'entreprise dans le répertoire Sirene : l'APE, le caractère actif ou non de l'entreprise et le nombre de salariés. Les deux premières variables doivent être utilisées avec précaution. En effet, la variable APE est principalement actualisée par les résultats des enquêtes annuelles d'entreprise. Quant à la variable « active/inactive », sa mise à jour ne s'effectue pas non plus de manière indépendante du traitement des EAE. En particulier, elle peut être déclenchée par les résultats d'enquêtes administratives dites «enquêtes d'amélioration du répertoire » sur les non-répondantes à l'EAE. Ces enquêtes concernent les entreprises sur lesquelles les gestionnaires des EAE n'ont aucun retour ; parmi celles-ci, ces derniers ne savent pas distinguer les entreprises ayant réellement disparu de celles toujours actives (les « vraies » non-répondantes à l'enquête). En revanche, les effectifs totaux des salariés au 31/12 proviennent des déclarations annuelles des données sociales (DADS), d'origine différente ; leur utilisation ne pose aucun problème particulier. Afin de réduire au maximum le biais, les valeurs des variables ont été prises dans l'état où elles étaient avant la prise en compte des résultats de l'enquête administrative sur les non-répondantes à l'EAE à caler et avant l'éventuelle modification de l'APE.

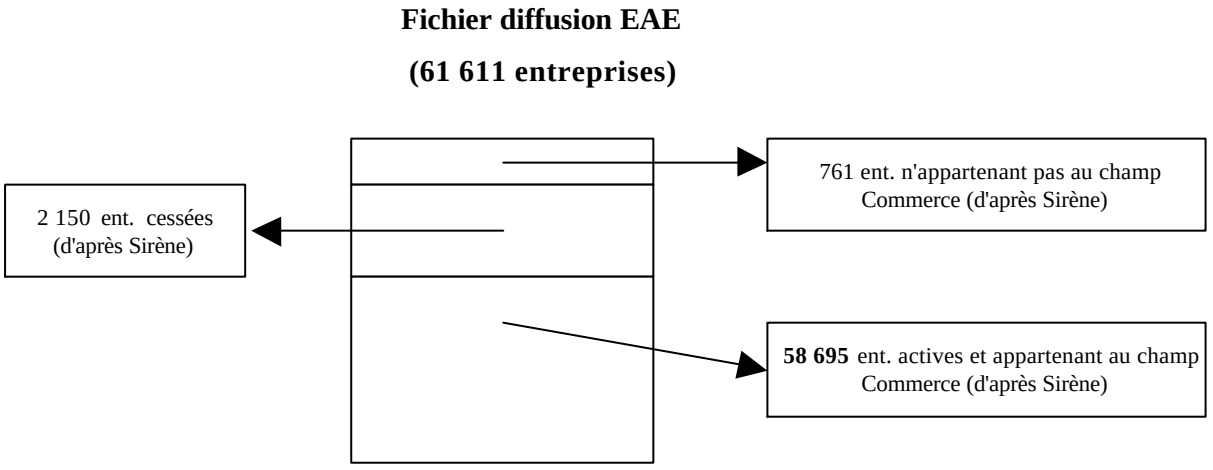
3.3 Constitution du fichier de données qui sera « calé »

L'expérience sur les EAE a été menée sur les données recueillies sur l'exercice 99 pour l'EAE-commerce. Le fichier de diffusion à partir duquel les principaux résultats de l'enquête sont établis contient les données de 61611 entreprises. Afin de respecter le principe de cohérence des variables entre les marges et le fichier d'enquête, les variables de la source auxiliaire (le répertoire Sirene dans notre cas) sont « incorporées » au fichier de données et sont utilisées pour le calage. Par conséquent, la variable «active/inactive » et la variable APE utilisées lors de la procédure de calage sont **celles issues du répertoire Sirene et non celles renseignées lors de la réalisation de l'enquête.**

L'information auxiliaire retenue conduit à utiliser, pour le calage, un fichier limité aux entreprises de l'échantillon qui, d'après le répertoire Sirene, sont actives et appartiennent au secteur du commerce au 31/12/99 (par la suite, nous les appellerons les «actives champ commerce »). Ce fichier n'a aucune

raison de coïncider avec le fichier de diffusion de l'enquête. En effet, le fichier diffusion contient trois types d'entreprises :

- les entreprises qui d'après les renseignements Sirène sont cessées au 31/12/99 (2150 entreprises sont dans ce cas). En effet, une entreprise peut être déclarée cessée au 31/12/99 dans le fichier Sirène mais avoir sur l'année concernée plus de 6 mois d'exercice et par conséquent faire partie du champ de l'enquête EAE et contribuer à ses résultats.
- les entreprises qui d'après les renseignements Sirène ne font pas (ou plus exactement ne font pas encore) partie du champ Commerce (761 entreprises sont dans ce cas). Rappelons que ce sont précisément les résultats des enquêtes EAE qui permettent de mettre à jour le code APE dans Sirène.
- les entreprises déclarées « actives champ commerce » au sens Sirène (58695 entreprises sont dans ce cas).



Sur ces 3 catégories d'entreprises, nous n'avons retenu pour le calage que les entreprises appartenant à la dernière catégorie c'est-à-dire celles déclarées « actives champ commerce ». De plus, afin de ne pas introduire de biais, il est nécessaire de prendre aussi en compte les entreprises déclarées hors-champ par le service enquêteur de l'EAE-Commerce (soit pour cause de cessation, soit pour changement de secteur) qui sont actives dans le champ Commerce au sens Sirène et qui sont par conséquent comptabilisées dans les marges. Sur les 4882 entreprises déclarées hors-champ, 3811 sont dans ce cas.

Le calage est donc réalisé sur 62506 entreprises (dont 58695 entreprises actives qui appartiennent au fichier de diffusion et 3811 entreprises déclarées hors-champ par le service enquêteur). A noter que pour l'exploitation de l'enquête, il faudra reconstituer le fichier de diffusion qui comprend 58695 entreprises qui ont participé au calage et 2916 entreprises qui n'ont pas participé à la procédure de calage. Ces 2916 entreprises conserveront leur poids initial pour les exploitations dans le fichier final.

3.4 Choix retenus pour le calage

La solution que nous avons retenue est de réaliser le calage à partir du fichier des entreprises présentes dans l'échantillon de l'EAE-Commerce et qui sont « actives champ commerce » **au sens de Sirène**. Avec les données dont nous disposons, il était aussi possible d'envisager un calage sur chacune des catégories d'entreprises (les actives au sens de Sirène, les cessées au sens de Sirène et celles n'appartenant pas au champ Commerce). Cependant, pour la catégorie des entreprises n'appartenant pas au champ commerce et celle des entreprises cessées (au sens de Sirène), le calage ne peut pas être réalisé à un niveau aussi fin que pour les entreprises actives (au sens de Sirène). On aurait pu envisager de caler les entreprises « cessées » sur le nombre d'entreprises de cette catégorie d'après Sirène et les entreprises « en dehors du champ commerce » au niveau global (c'est-à-dire sans distinction des secteurs) sur le nombre d'entreprises de cette catégorie et sur le nombre de salariés au 31/12/99 obtenus d'après Sirène. Nous avons privilégié la solution qui consiste à ne réaliser un calage que sur les « actives champ commerce » car elle conduit à un nombre de contraintes de calage plus faible que la seconde.

En ce qui concerne le choix de la méthode de calage, notre point de vue consiste à privilégier au maximum l'information provenant du plan de sondage. Dans cette optique, le calage doit modifier le moins possible les pondérations d'origine, en général directement issues du plan de sondage utilisé. Nous avons donc privilégié l'utilisation de la méthode logit : son avantage est de pouvoir définir a priori une borne inférieure et une borne supérieure pour les rapports du poids après calage sur celui avant calage. Dans notre expérimentation, les valeurs minimale et maximale du rapport du poids après calage sur celui avant calage ont été fixées respectivement à 0,6 et 4,4. Cette procédure de calage conduit à une hausse globale des poids de près de 5 %, soit une pondération finale moyenne des unités du fichier de 10,1.

3.5 Première analyse des résultats

Comme nous l'avons signalé dès l'introduction, les résultats obtenus ne l'ont été qu'à titre d'expérience méthodologique. A ce titre, nous avons abordé l'analyse des résultats sous l'angle statistique en nous intéressant en particulier au gain de variance obtenu. Une validation plus complète des résultats obtenus, et plus précisément de la pertinence des agrégats économiques estimés, ne peut être réalisée que par des experts connaissant bien le domaine du commerce.

La théorie indique que le calage permet d'augmenter la précision des résultats, c'est-à-dire de diminuer la variance d'échantillonnage. Il n'est pas toujours facile de vérifier cet argument théorique en calculant des estimations de gain de précision. Parfois, les conclusions obtenues peuvent même aller à l'encontre de la théorie. Ceci s'explique par le fait que l'estimation de variance admet elle aussi une variance et une estimation de variance. L'estimation de variance peut alors conduire à une valeur plus élevée après calage qu'avant. C'est en particulier le cas lorsque le nombre d'unités dans le domaine est faible car l'estimateur de variance est alors très instable. À partir de nos données, nous avons comparé les estimations de variance obtenues avant et après calage pour le chiffre d'affaires total estimé dans 6 domaines de diffusion (exprimés en classe NAF). Pour 3 d'entre eux, la procédure de calage permet de gagner de 0,1 à 2,7 points sur la précision relative estimée des estimateurs. Cependant, pour les 3 autres domaines, l'estimation de la variance obtenue après calage est plus grande que celle obtenue avant.

Une approche complémentaire permettant d'évaluer l'apport du calage consiste à examiner les résultats économiques obtenus à partir de l'EAE avec les nouveaux «poids ». D'un point de vue descriptif, nous avons étudié les modifications obtenues avec les nouveaux poids sur l'estimation du total du chiffre d'affaires et celle du nombre d'entreprises par secteur.

- Pour l'ensemble du commerce de détail et de l'automobile, le chiffre d'affaires total estimé à partir des poids issus du calage diffère de moins de 1 % de celui obtenu en prenant les anciens poids. Cependant, ce chiffre cache de grandes inégalités selon les secteurs. Le calage conduit à des variations pouvant atteindre plus de 20 % dans plusieurs secteurs, en particulier dans des secteurs où le nombre estimé d'entreprises a été considérablement modifié suite à la procédure de calage.
- En ce qui concerne le nombre estimé d'entreprises dans le commerce de détail, la procédure de calage conduit à une hausse du nombre estimé d'entreprises de 3%. Cet écart est très différent selon les secteurs étudiés. Une hausse (resp. baisse) du nombre d'entreprises due à la procédure de calage va en général de pair avec une hausse (resp. baisse) du chiffre d'affaires estimé mais ce n'est pas systématiquement le cas.

4. Conclusion

Dans ce papier, nous avons listé les principales difficultés méthodologiques auxquelles se heurte le statisticien qui souhaite mettre en œuvre une procédure de calage pour son fichier d'enquête. Cette liste est loin d'être exhaustive ; en particulier, nous n'avons pas abordé les problèmes concernant directement la manipulation de la macro SAS CALMAR : le nombre maximum de variables de calage autorisé par la macro, la constitution correcte de la table des marges, le problème de colinéarité approchée des variables de calage, la numérotation des modalités des variables qualitatives…

La mise en œuvre des méthodes de calage sur l'EAE nous a permis d'identifier des problèmes méthodologiques propres aux enquêtes entreprises. Certains d'entre eux sont directement reliés au fait que l'enquête considérée est l'EAE. En effet, cette enquête sert en particulier à mettre à jour le code d'Activité Principale Exercée des entreprises (APE) dans Sirene. De plus, le renouvellement au quart de l'échantillon EAE chaque année qui conduit à conserver au sein de l'échantillon une petite entreprise 4 ans de suite une fois qu'elle est sélectionnée ajoute des problèmes méthodologiques supplémentaires. En toute rigueur, le fichier auxiliaire utilisé ne devrait pas tenir compte des mises à jour relatives à une entreprise tant qu'elle n'est pas sortie de l'échantillon. Dans notre étude, nous avons supposé pour simplifier que les informations recueillies les années antérieures par l'enquête l'auraient été de toutes les façons par une autre source (cette hypothèse est sans doute inexacte pour les petites entreprises en particulier celles qui n'ont aucun salarié). Les difficultés évoquées seraient de moins grande ampleur pour un calage concernant une enquête thématique (à condition de s'assurer que la source auxiliaire utilisée n'a pas été mise à jour par les résultats d'une enquête qui lui serait coordonnée).

Il serait inexact d'en déduire à la lecture de ce papier que la réalisation d'un calage à l'Insee est beaucoup plus simple pour une enquête sociale que pour une enquête réalisée auprès des entreprises. En effet, il est plus facile de respecter le principe de cohérence évoqué dans la partie IV pour les enquêtes entreprises car les sources auxiliaires disponibles sont en général exhaustives. Par conséquent, il suffit d'« injecter » pour chaque individu du fichier d'enquête la valeur disponible dans la source auxiliaire et de l'utiliser pour le calage. En revanche, pour les enquêtes sociales, la source externe n'est en général pas exhaustive sauf dans des cas très particuliers (lorsque par exemple l'enquête sociale a lieu la même année que le recensement de la population). La mise en œuvre du calage nécessite donc d'utiliser dans le fichier d'enquête et dans la source externe des informations qui n'ont pas été recueillies par le même protocole d'enquête. Cette procédure risque alors d'introduire un biais difficilement mesurable supérieur au gain de variance espéré.

Bibliographie

[1] Ardilly P., *Techniques de sondages*, éditions Technip, 1994.

[2] Caron N., "Application des méthodes de calage à l'EAE-commerce", *Document de travail* n°0201 série Méthodologie Statistique, Insee, 2002.

[3] Deville J.-C., Le Guennec J. et Sautory O., "L'emploi de CALMAR II : calage généralisé et application à la non-réponse", *Présentation aux JMS de décembre 2002*.

[4] Deville J.-C. et Särndal C., "Calibration estimators in Survey Sampling", *JASA*, vol. 87, n° 418, p. 376-382, 1992.

[5] Deville J.-C., Särndal C. et Sautory O., "Generalized raking procedures in Survey Sampling", *JASA*, vol. 88, n° 423, p. 1013-1020, 1993.

[6] Deville J.-C., Särndal C. et Sautory O., "Generalized raking procedures in Survey Sampling", *JASA*, vol. 88, n° 423, p. 1013-1020, 1993.

ANNEXE : aspects mathématiques du calage sur marges

Soit une population $U = \{1 \dots k \dots N\}$ de N individus, dans laquelle on a sélectionné un échantillon *s* de taille n avec les probabilités d'inclusion $\pi_k = P(k \in s)$. Dans le cas d'un sondage aléatoire simple, les probabilités d'inclusion sont constantes et égales à n/N. On notera ***Y*** la variable d'intérêt dont on veut estimer le total Y sur la population. Par définition, $Y = \sum_{k \in U} y_k$. L'estimateur de Horvitz-Thompson est défini par $\hat{Y}_\pi = \sum_{k \in s} \frac{1}{\pi_k} y_k$.

On notera, par la suite, d_k l'inverse de la probabilité d'inclusion de l'individu k. Il correspond au poids de sondage de l'individu k et permet d'extrapoler les données obtenues sur l'échantillon à l'ensemble de la population concernée.

On suppose que l'on dispose de J variables auxiliaires notées $X_1 \dots X_j \dots X_J$ connues sur s et dont on connaît **les totaux sur la population** $X_j = \sum_{k \in U} x_{jk}$. L'objectif des méthodes de calage est de déterminer un nouvel estimateur de Y de la forme $\hat{Y}_w = \sum_{k \in s} w_k y_k$ où les nouveaux poids w_k sont « proches » des poids initiaux d_k, au sens d'une distance qu'il faut définir, et vérifient les **équations de calage** définies par : $\forall j = 1 \dots J \quad \sum_{k \in s} w_k x_{jk} = X_j$.

La résolution théorique passe par le choix d'une « fonction de distance » G, d'argument $x = w_k / d_k$, positive et convexe (et de classe C^2), telle que $G(1) = G'(1) = 0$. En posant $x_k = (x_{1k} \dots x_{Jk})'$ et $X = (X_1 \dots X_J)'$, le programme de minimisation s'écrit :

$$\min_{w_k} \sum_{k \in s} d_k G(w_k / d_k) \text{ avec } \sum_{k \in s} w_k x_k = X$$

Celui-ci se résout en introduisant un vecteur de multiplicateurs de Lagrange λ qui est déterminé par la résolution (par la méthode itérative de Newton) du système non linéaire des **équations de calage** : $\sum_{k \in s} d_k F(x_k' \lambda) x_k = X$ où F est la fonction réciproque de la dérivée de G.

Dans la macro SAS CALMAR, quatre « fonctions de distance » – autrement dit, quatre méthodes – sont disponibles : la méthode linéaire, la méthode «raking ratio », la méthode logit et la méthode « linéaire tronquée ».