REPONDÉRATIONS DANS LA NOUVELLE ENQUÊTE EMPLOI EN CONTINU.

Philippe FEVRIER (*), Pauline GIVORD (**)

(*) Insee, Unité Méthodes Statistiques
(**) Insee, Département Emploi et Revenus d'activité

Introduction

L'enquête Emploi de l'Insee est passée en 2002 d'une collecte annuelle (menée généralement au mois de mars de chaque année) à une collecte en continu. Désormais, l'enquête emploi est réalisée chaque trimestre dans environ 54 000 logements, et la collecte est uniformément répartie sur toutes les semaines du trimestre. Chaque logement est interrogé pendant six trimestres consécutifs, contre trois années successives dans l'enquête emploi annuelle. Ce nouveau mode de collecte devrait permettre un meilleur suivi conjoncturel de l'emploi et du chômage, ainsi que des trajectoires individuelles. Le questionnaire a également été modifié en profondeur dans un souci d'harmonisation européenne.

Le passage à cette nouvelle enquête est l'occasion de remettre à plat le système de calcul des pondérations de l'enquête, destiné à corriger des variations d'échantillonnage et des biais introduits par la non réponse.

1. La nouvelle Enquête Emploi en continu

1.1 Un échantillon aréolaire

Comme ses prédécesseurs, le nouvel échantillon de l'enquête emploi est aréolaire : tous les logements appartenant à une aire géographique délimitée sont interrogés. Les échantillons aréolaires ont de nombreux atouts : ils garantissent en principe un meilleur taux de réponse, ils réduisent les risques d'oubli de logements par l'enquêteur, ils permettent le suivi des logements neufs à un coût réduit et ils diminuent le nombre de déplacements des enquêteurs. Tous ces avantages ont leur revers : les personnes qui vivent dans des logements contigus ont souvent des caractéristiques proches et sont moins représentatives de la population que le même nombre de personnes tirées au hasard complet. Cet effet dit « effet de grappe » détériore la précision. Les aires doivent donc être de taille réduite (et si possible hétérogènes...).

Dans le précédent échantillon emploi, les aires comprenaient 20 logements en moyenne en zone urbaine et 40 logements en moyenne en zone rurale. Pour réduire l'effet de grappe, le nouvel échantillon emploi ne comporte que des aires de 20 logements en moyenne, que ce soit en zone urbaine ou en zone rurale.

Le tirage de l'échantillon a été effectué à partir des données du recensement de 1999. Il représente de manière relativement homogène l'ensemble des régions métropolitaines et des tranches d'unités urbaines (communes rurales, communes des unités urbaines de moins de 10 000 habitants, de 10 000 à 50 000 habitants, de 50 000 à 200 000 habitants, de plus de 200 000 habitants). On trouvera dans Christine (2000) une description détaillée de l'échantillon et de sa construction. Cet échantillon permettra de réaliser l'enquête emploi jusqu'à la fin du deuxième trimestre 2010. Un échantillon de réserve a également été tiré, au cas où le nouveau recensement ne permettrait pas le tirage d'un autre échantillon à temps.

La représentation des logements neufs sera assurée par un échantillonnage direct sur le terrain : les enquêteurs interrogeront tout ou partie des logements neufs construits dans l'aire d'enquête entre la date du recensement et la date de l'enquête. Auparavant, dans l'enquête emploi annuelle, une partie des logements neufs étaient échantillonnés à partir de la BSLN.

1.2 Des interrogations plus fréquentes et nombreuses

En régime permanent, environ 54 000 logements «ordinaires », c'est-à-dire hors collectivité (foyer, maison de retraites…) sont interrogées chaque trimestre. Le taux de sondage moyen est de l'ordre de 1/600^{e}, contre 1/300^{e} dans l'enquête annuelle. D'un trimestre à l'autre, un sixième de l'échantillon est renouvelé. Chaque logement est ainsi interrogé six fois de suite (quels que soient ses occupants), puis quitte l'échantillon.

Le champ de l'enquête est constitué des personnes vivant dans des logements ordinaires. Les ménages sont interrogés dans leur *résidence principale*, c'est-à-dire celle où ils vivent habituellement. En outre, les personnes résidant en logement collectif, mais ayant un lien avec un ménage ordinaire (enfant étudiant vivant en foyer universitaire, parent âgé vivant en maison de retraite) sont également interrogées dans le logement du ménage auxquelles elles sont rattachées. Ces personnes qui ne font pas partie du champ de l'enquête sont néanmoins rattachées au ménage conformément au règlement n° 577/98 du conseil européen sur les enquêtes Force de Travail.

La première et la sixième dernière interviews d'un logement sont réalisées en face à face ; les interviews intermédiaires le sont en général par téléphone.

A l'intérieur d'un trimestre, la collecte est étalée de manière uniforme sur chacune des 13 semaines qui le composent. Pour ce faire, l'ensemble des logements d'une même aire est affecté à l'une de ces 13 semaines, dite semaine de référence. Les enquêtés sont interrogés sur leur situation (notamment vis-à-vis du marché du travail) au cours de cette semaine de référence. La collecte pour une aire débute juste après la semaine de référence correspondante, et s'étale sur une période de deux semaines et deux jours.

La taille des échantillons interdit d'effectuer des estimations mensuelles (et a fortiori hebdomadaires…) avec un degré de précision acceptable. Les principales estimations se feront donc sur une base trimestrielle, ce qui permettra d'assurer un meilleur suivi conjoncturel de l'emploi et du chômage que ne le permettait par définition l'enquête annuelle. Les résultats ne seront en revanche pas directement comparables avec ceux de l'enquête annuelle, qui fournissaient une mesure ponctuelle au mois de mars de chaque année.

1.3 Un questionnaire modifié

Le questionnaire s'articule en 16 modules, qui reprennent l'ensemble des thèmes traités par l'enquête annuelle : activité professionnelle, salaires, employeur, durée du travail, recherche d'emploi, formation,... Certains d'entre eux ont été approfondis, afin de mieux répondre à la demande sociale sur ces sujets : le module durée du travail et le module revenu, avec la prise en compte des revenus annuels des indépendants et la perception d'allocations, sont deux exemples d'améliorations importantes du questionnement.

La modification majeure du questionnaire par rapport à celui de l'enquête emploi annuelle concerne la classification entre actif occupé, chômeur et inactif. L'ensemble de ces questions a été rassemblé en début du nouveau questionnaire, dans un module dit « module BIT ». Son architecture générale tente de coller au plus près des critères du Bureau International du Travail et suit les recommandations européennes du règlement n° 1897/2000, Eurostat : les premières questions portent donc sur l'exercice d'un travail «effectif » au cours de la semaine de référence, puis sur la recherche éventuelle d'un emploi sur la période récente, et enfin sur la disponibilité à exercer cet emploi.

La nouvelle enquête Emploi devrait donc permettre une mesure plus directe et donc plus précise de l'activité telle que la définit le BIT. Il est probable qu'elle devrait également permettre une meilleure comptabilisation des emplois atypiques. En revanche, elle ne permettra plus d'effectuer une classification de l'activité « au sens du recensement » : il n'est en effet plus demandé aux enquêtés de se définir eux-mêmes comme chômeurs, actifs, étudiants,… comme c'est le cas actuellement dans l'enquête annuelle (question Fi qui ouvre l'entretien).

1.4 Un taux d'impossibles à joindre plus élevé que pour l'enquête annuelle

L'enquête emploi en continu a débuté avant que ne se termine la série des enquêtes annuelles, en juillet 2001. Elle ne viendra remplacer définitivement l'enquête annuelle qu'à partir du 1er janvier 2003. Des informations sont donc d'ores et déjà disponibles sur la qualité du nouveau mode de collecte.

Le taux d'enquêtes acceptées observé sur les premiers trimestres de collecte est plus faible que dans l'enquête annuelle (tableau 2) : il est de 78% au troisième trimestre 2001, puis semble s'établir autour de 83 % sur les trois trimestres de collecte suivants, alors qu'il était de 86,6% sur l'échantillon aréolaire[1] de l'enquête annuelle emploi de mars 2002. Cette faiblesse du taux d'enquêtes acceptées s'explique à la fois par un taux de refus légèrement plus important (4,5% en moyenne, au lieu de 4% à l'enquête annuelle de mars 2002), mais surtout du fait d'un taux de ménages impossibles à joindre (IAJ) ou absents de leur domicile pour une longue durée (ALD) beaucoup plus élevé : 12 % (17% au T3 2001) contre 9 % dans le cadre de l'enquête annuelle de mars 2002.

Tableau 1 : Nombre de logements par trimestre et répartition selon leur statut d'occupation

	T3 2001	T4 2001	T1 2002	T2 2002
Nombres de logements	54 018	53 708	53 639	53772
Dont :				
- Résidences principales	43 247	43 251	43 456	43 736
- Résidences Occasionnelles	635	670	648	653
- Résidence Secondaires	4 827	4 909	4 891	4 925
- Logement Vacant	4 578	4 356	4 121	4 004
- Autres hors champs	731	522	523	454

Tableau 2 : Résultat de l'interview pour les résidences principales (en %)

	Enquête Emploi annuelle (échantillon aréolaire) Mars 2002	Enquête Emploi en continu			
		T3 2001	T4 2001	T1 2002	T2 2002
acceptés	86,6 %	78,5 %	83,4 %	83,3 %	83,6 %
IAJ, ALD…	9,3 %	17,2 %	12,0 %	13,2 %	11,5 %
Refus	4,1 %	4,4 %	4,6 %	4,7 %	4,9 %

1 C'est à dire hors « échantillon spécial », ce terme désignant l'échantillonnage spécifique qui était réalisé pour tenir compte de la construction neuve à partir de la BSLN, procédure qui n'a pas été conservé pour l'enquête en continu.

Ces comparaisons sont évidemment à prendre avec prudence. La structure des échantillons des premiers trimestres de collecte est particulière : le nouvel échantillon décrit ci-dessus, issu du recensement de 1999, ne sert sur le terrain que progressivement depuis le troisième trimestre 2001, par sixième. Une part importante de l'échantillon (5/6 au troisième trimestre 2001, 1/3 au deuxième trimestre 2002) des premiers trimestres de collecte est donc composée de logements qui faisaient initialement partie du précédent échantillon Emploi. Cette partie de l'échantillon a donc été déjà interrogée au moins trois fois dans le cadre de l'enquête emploi annuelle, et parfois plus : certains logements faisaient également parti du « dispositif léger » mis en place dès 1998 afin de permettre un test grandeur nature d'une enquête en continu. Il est probable que cette sollicitation lourde et répétée des mêmes logements, et donc souvent des mêmes individus, a un effet négatif sur le taux de réponse. Au-delà de ces effets de mise en route, plusieurs raisons laissent penser que le taux d'enquête de l'enquête en continu restera plus faible que celui de l'enquête annuelle :

En premier lieu, la période au cours de laquelle peut s'effectuer la collecte est plus courte dans l'enquête en continu (deux semaines et deux jours) que dans l'enquête annuelle (jusqu'à cinq semaines). Les enquêteurs peuvent rencontrer plus de difficultés à rencontrer tous les ménages d'une aire dans les temps impartis, augmentant sensiblement le nombre de ménages « impossibles à joindre ». De plus, cet effet est amplifié par le fait que la collecte en continu couvre des périodes au cours desquelles le taux d'absence des ménages est élevé (vacances scolaires…). Ce problème ne se posait pas avec l'enquête annuelle : le mois de mars avait précisément été choisi comme période de collecte parce qu'il s'agit du mois le moins sujet à ces perturbations saisonnières et d'un mois habituellement sans vacances scolaires. Enfin, le nombre et la fréquence des réinterrogations dans l'enquête en continu, tous deux significativement plus élevés que dans l'enquête annuelle, détériorent très probablement le taux de réponse[2]. On peut craindre de ce fait que le biais de rotation ne soit amplifié : les chômeurs répondent moins en réinterrogation que les autres personnes interrogées. Ce biais s'observe déjà dans l'enquête Emploi annuelle : le taux de chômage estimé par vague d'interrogation est systématiquement décroissant avec le rang d'interrogation. Ce résultat semble exister également dans les premières données issues de l'enquête en continu : le taux de chômage estimé sur les personnes interrogées pour la première fois est plus élevé que celui des vagues supérieures. Le faible recul ne permet pas cependant d'établir la validité de ce résultat.

2. Repondérations

Comme dans toute enquête, l'estimation d'une grandeur à partir d'un simple échantillon nécessite de disposer de poids qui permettent de pondérer chaque individu ou ménage de l'échantillon. En particulier, pour calculer l'estimateur d'Horvitz-Thompson en limitant les biais introduits par le phénomène de non-réponse, il est nécessaire d'estimer la probabilité que chaque individu soit in fine répondant. Cette probabilité est le produit de deux termes : la probabilité d'appartenir à l'échantillon interrogé et la probabilité de répondre conditionnellement au fait d'être interrogé. La première probabilité est bien connue car elle découle du plan de sondage, lequel est « maîtrisé ». C'est l'inverse de cette probabilité d'appartenir à l'échantillon qui fournit les poids initiaux. La deuxième probabilité en revanche est inconnue et il faut chercher un moyen de l'estimer. Cette estimation servira à transformer les poids initiaux pour prendre en compte la non-réponse. Il existe deux méthodes pour transformer les poids initiaux. La première consiste à estimer directement la probabilité de non-réponse dans les données et à corriger les poids initiaux à l'aide de cette estimation. Il est également possible de corriger la non-réponse à l'aide d'un calage et c'est cette deuxième méthode que nous présentons dans ce papier. Une autre transformation des poids est nécessaire pour limiter les fluctuations d'échantillonnage. Cette étape est, elle aussi, réalisée au moyen d'un calage (Deville et Sarndal, 1992).

[2] 6 interrogations, contre 3 dans l'enquête annuelle, et tous les 3 mois, contre tous les ans dans l'enquête annuelle.

2.1 Méthode générale de repondération par calage

Deux transformations sur les poids initiaux sont nécessaires pour estimer les poids finaux. Il est tout d'abord nécessaire d'estimer la probabilité que chaque individu soit in fine répondant. Cette estimation permettra de transformer les poids initiaux pour prendre en compte la non-réponse. Il existe deux méthodes pour corriger les poids. La première consiste à estimer directement la probabilité de non-réponse dans les données et à corriger les poids initiaux à l'aide de cette estimation. Il est également possible de corriger la non-réponse à l'aide d'un calage. C'est cette deuxième méthode que nous présentons dans ce papier.

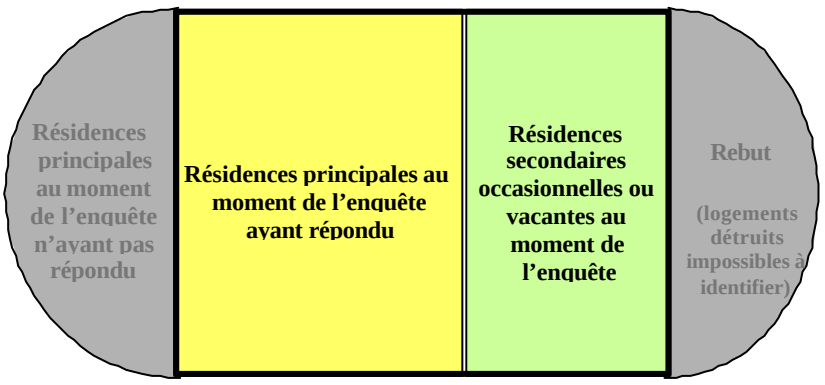
La façon de modéliser et d'appréhender un calage dans l'enquête emploi n'est pas chose aisée. La présentation la plus simple consiste à distinguer dans un premier temps le calage sur les variables du Recensement de la Population (RP) du calage sur les variables individus.

Commençons par la présentation de la logique sous-jacente au calage sur les variables RP. Nous appellerons échantillon initial, l'échantillon théorique composé de l'exhaustivité des logements ordinaires qui appartiennent aux aires à enquêter (logements recensés) et des logements ordinaires que l'on ajoute grâce au ratissage (logements neufs ou oubliés au RP).

Cet échantillon initial peut, après l'enquête, se décomposer en quatre parties comme indiqué sur le graphique suivant :

- la première partie se compose des logements qui sont résidence principale au moment de l'enquête mais qui n'ont pas répondu à l'enquête ;
- la deuxième partie se compose des logements qui sont résidence principale au moment de l'enquête et qui ont répondu à l'enquête ;
- la troisième partie se compose des logements qui sont résidence secondaire, occasionnelle ou vacante (SOV) au moment de l'enquête ;
- la quatrième et dernière partie forme le « rebut » de l'échantillon, elle est composée de logements détruits, impossibles à identifier, de logements ayant perdu leur usage d'habitation, de logements ayant été créés par erreur,...

Graphique 1 : Répartition de l'échantillon initial selon le statut d'occupation à l'enquête



Nous appellerons échantillon pour calage l'échantillon issu de l'échantillon initial (où nous avons retiré les parties grisées) et composé de la deuxième partie et de la troisième partie de celui-ci, c'est à dire des logements qui sont résidence principale au moment de l'enquête et qui ont répondu à l'enquête et des logements qui sont résidence secondaire, occasionnelle ou vacante au moment de l'enquête. C'est cet échantillon qui sera utilisé dans les calages, d'où son appellation. Il est nécessaire, pour obtenir in fine des estimations qui correspondent aux logements qui sont résidence principale au moment de l'enquête (voir plus bas), de garder à ce niveau les logements secondaires, occasionnels ou vacants.

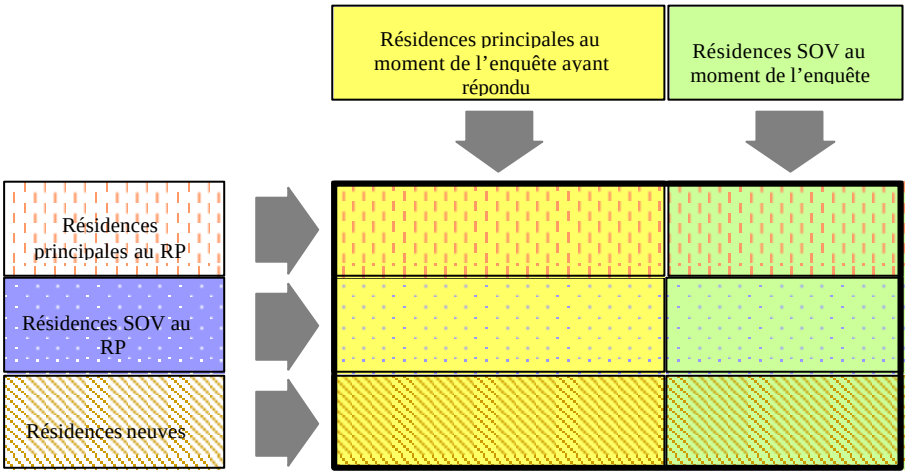
Pour bien comprendre comment cet échantillon peut être utilisé pour réaliser le calage que l'on désire, il est utile de remarquer qu'on peut lui-même le décomposer en plusieurs sous-échantillons, comme indiqué dans le graphique suivant :

- le premier sous-échantillon correspond aux logements qui étaient résidence principale au moment du Recensement de la Population (RP) ;

- le deuxième aux logements qui étaient résidence secondaire, occasionnelle ou vacante au moment du RP ;

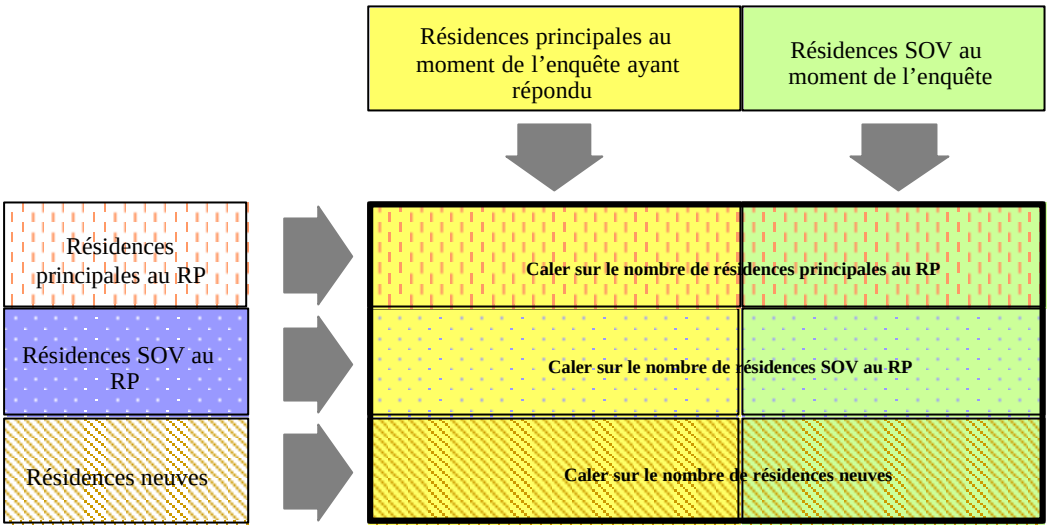
- le troisième aux logements neufs ;

Graphique 2 : Répartition de l'échantillon pour calage selon le statut d'occupation au RP 99



Nous obtenons finalement une décomposition en six sous-parties qui repose sur le croisement de la catégorie du logement à deux dates : au RP et au moment de l'enquête. Le calage que l'on aimerait réaliser est le suivant : les totaux obtenus sur les logements de l'échantillon pour calage qui étaient résidence principale au moment du RP et qui sont restées résidence principale au moment de l'enquête devraient être calés sur des totaux de la population globale de logements qui sont à la fois résidence principale aux deux dates ; et similairement pour les cinq autres sous-parties. Malheureusement, nous ne disposons pas de tels totaux et il faut donc regrouper les sous-parties deux à deux. Ainsi par exemple, les totaux sur l'ensemble des logements de l'échantillon pour calage qui étaient résidences principales au RP doivent être calés sur les totaux de la population totale des résidences principales au RP, qui sont connus. Il en est de même pour les autres sous-parties, comme l'indique le graphique suivant :

Graphique 3 : Calage sur les variables RP



Mathématiquement, le calage peut être écrit de la façon suivante pour les calages sur les variables concernant les logements qui sont résidence principale au RP :

$$\sum_{k \in E \cap RP} w_k x_k(t_{RP}) = \sum_{k \in U \cap RP} x_k(t_{RP})$$

E représente les logements de l'échantillon pour calage, RP représente les logements qui sont résidence principale au RP, U représente la population totale des logements ; tandis que $x_k(t_{RP})$ représente les caractéristiques d'un logement k à la date du RP. Il est important de remarquer que ce sont bien les caractéristiques à la date du RP qui sont utilisées dans le calage. On cherche ainsi à redresser l'échantillon de manière à estimer parfaitement bien la structure de la population au RP. Un calage analogue peut être écrit pour les logements qui sont résidence SOV au RP. Enfin, le calage sur le nombre de logement neuf s'écrit simplement :

$$\sum_{k \in E \cap N} w_k 1 = \sum_{k \in U \cap N} 1$$

où N représente l'ensemble des logements neufs.

Si la méthode de calage exposée ne pose pas de problème théorique, il n'en est pas de même d'un point de vue pratique. Tout d'abord, la variable qui permet de différencier dans l'échantillon initial les logements qui sont résidences principales au moment de l'enquête, des logements qui sont résidences SOV (secondaires, occasionnelles, vacantes) doit être de bonne qualité. Ensuite, pour diviser l'échantillon pour calage, il est nécessaire de connaître la catégorie du logement au RP. Or certains logements de l'échantillon qui étaient présents au RP n'ont pas pu être appariés (en particulier en raison du nombre important d'identifiants RP mal codés) [3]. Nous avons donc du recourir à des imputations pour traiter le cas de ces logements, puisqu'ils doivent participer au calage mais que les variables sont inconnues. Les valeurs imputées correspondent simplement aux valeurs moyennes calculées sur les logements pour lesquels l'information était disponible. Enfin, il faut traiter le cas des logements oubliés au RP. Ces logements ne doivent participer à aucun calage et les différentes

[3] Ces logements sont les logements qui sont notés au moment de l'enquête comme appartenant au RP (construction antérieure au RP, non oublié au RP) mais qui ne sont pas retrouvés lors de l'appariement global avec tous les logements recensés.

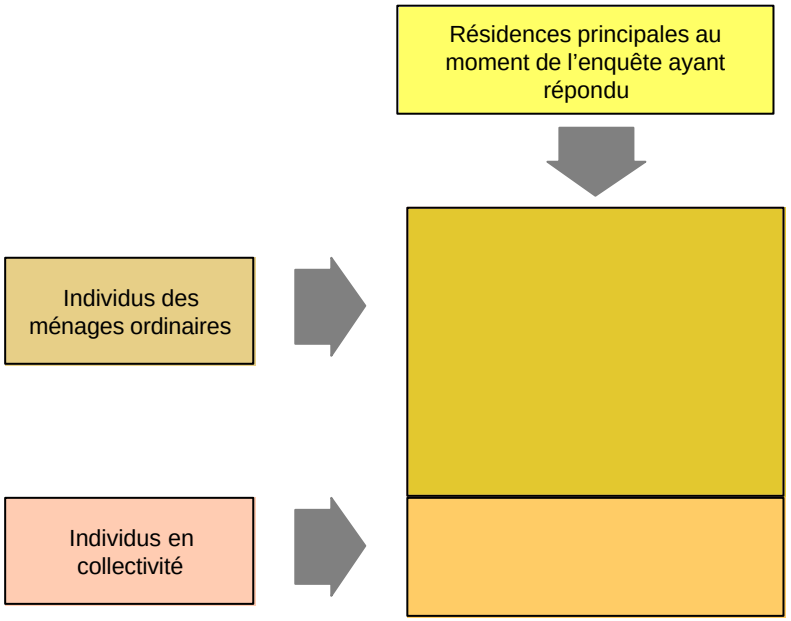
variables intervenant dans les calages sont donc mises à zéro pour ces logements. Le poids de ces logements n'est donc pas modifié par la procédure de calage, et ces logements restent avec leur poids initiaux.

Une fois le calage réalisé, on restreint l'échantillon pour calage repondéré aux seuls logements qui sont résidence principale au moment de l'enquête et qui ont répondu. Nous obtenons ainsi l'échantillon final de répondants. Les estimations calculées sur cet échantillon final correspondent ainsi à des estimations sur des résidences principales au moment de l'enquête, ce qui est cohérent avec le champ de l'enquête.

Le calage concernant les variables individu est en quelque sorte plus simple, puisqu'il ne concerne que les logements qui sont résidence principale au moment de l'enquête et qui ont répondu. Les valeurs de ces variables sont connues sur les logements en question et sont mises à zéro pour tous les autres logements.

Une difficulté survient néanmoins dans la comparaison entre le champ de l'enquête et ce que l'on cherche à mesurer. Le champ de l'enquête emploi en continu est en effet celui des ménages résidant en logements ordinaires dont sont théoriquement exclus les individus vivant en logements collectifs. C'est dans ce sens qu'est effectué l'échantillonnage. Cependant, conformément au règlement européen n° 577/98 du conseil européen sur l'organisation des enquêtes Force de Travail, à défaut d'autres enquêtes portant directement sur les ménages vivant en collectivité, « les personnes qui, dans ces ménages vivant en collectivité, ont gardé un lien avec un ménage privé sont prises en compte dans le cadre de ce dernier ». Ceci pose donc la question de savoir comment intégrer ces individus pour le calcul des pondérations et des estimations.

Graphique 4 : Répartition des répondants selon leur statut



Deux méthodes sont envisageables. La première possibilité pour intégrer les individus hors ménages ordinaires consiste à considérer que les individus « récupérés » sont représentatifs de l'ensemble de la population vivant en collectivité. Cette hypothèse ne semble pas justifiable a priori puisqu'il n'y a pas d'information précise sur ces individus, mais cette méthode a le mérite de poser clairement les hypothèses sur laquelle elle repose[4]. L'idée consiste alors à garder toute la population interrogée et à utiliser dans le calage une pyramide des âges qui correspond à l'ensemble total de la population.

La deuxième méthode envisageable, et qui est celle retenue dans l'enquête emploi annuelle, est la suivante. Les individus vivant en collectivité sont enlevés du fichier pour le calcul des pondérations. La pyramide des âges utilisée alors est la pyramide qui correspond à la population vivant en ménages ordinaires. Une fois les pondérations terminées, on réintègre dans le fichier les individus vivant en collectivité que l'on avait trouvés, en leur attribuant le poids du logement. Cette méthode n'a pas réellement de fondement théorique. Le calage réalisé correspond bien à la logique « ménages ordinaires », et la réintégration des individus vivant en collectivité se justifie par le but à atteindre i.e un taux de chômage sur la population totale. Eurostat dit d'ailleurs à ce sujet qu'il faut faire ce qu'on peut pour approcher au mieux ce but. Néanmoins, on ne sait pas bien dans cette méthode ce que représentent réellement les individus vivant en collectivité que l'on a ainsi ajoutés au fichier. D'autre part, la pyramide des âges qu'il faut utiliser ici est sans doute de moins bonne qualité que la pyramide des âges totale puisque le département de la démographie ne dispose que des informations au RP pour estimer le nombre de personnes vivant en collectivité ainsi que leur âge.

Pour réaliser ces deux types de calage différents, il nous a fallu créer deux types de fichier. La première version des fichiers qui correspond à un calage sur la pyramide totale sera notée version 1, tandis que la deuxième version, qui correspond à la méthode de l'enquête annuelle et qui utilise un calage sur la pyramide des ménages ordinaires sera notée version 2.

2.2 Plusieurs types de calage

Pour repondérer chaque version, plusieurs méthodes de calage sont envisageables. Ce sont ces méthodes que nous exposons maintenant. Nous ne retenons ici que les variables RP utilisées dans la repondération de l'enquête annuelle, à savoir la tranche d'unité urbaine, le type de logement et le nombre de pièces.

Par ailleurs, il est possible de repondérer soit le fichier total, soit chaque sous-échantillon de manière indépendante les uns des autres. Nous travaillerons ici au niveau de chaque sous-échantillon. Le premier avantage de cette méthode repose dans le fait que chaque sous-échantillon ne présente pas forcément la même structure de non-réponse. En effet, lorsqu'un ménage a été contacté, il n'y a aucune raison que sa probabilité de réponse soit identique à celle d'un ménage qui n'a jamais été contacté. Il est donc nécessaire, lorsqu'on travaille sur le fichier total, d'introduire des variables supplémentaires pour prendre en compte cet effet : variable du rang d'interrogation, croisement des variables avec ce rang d'interrogation,… En repondérant chaque sous-échantillon de manière indépendant, on évite ces problèmes. De plus, cette deuxième méthode à l'avantage de fournir des estimations sans biais au niveau de chaque sous-échantillon, ce qui peut être nécessaire lors du calcul d'estimateurs complexes (voir Bosredon et Février 2002).

Finalement, les différentes méthodes de calage envisagées sont les suivantes :

- Le premier calage considéré, noté calage 1 par la suite, correspond à une méthode approchée de la méthode de l'enquête annuelle. Dans une première étape, on redresse l'échantillon par calage sur des variables RP (tranche d'unité urbaine, type de logement et nombre de pièces). On redresse alors dans une deuxième étape l'échantillon obtenu précédemment par calage sur une pyramide des âges.

[4] La seule information disponible est celle du RP, mais elle ne nous permet pas de tester cette hypothèse.

L'avantage de la méthode réside dans la dissociation des deux étapes de calage qui n'ont pas le même but. Le premier permet en effet de corriger de la non-réponse, tandis que le deuxième est un « vrai » calage qui a pour but de limiter les fluctuations d'échantillonnage.

L'inconvénient majeur de cette méthode réside cependant dans le fait que le deuxième calage détruit le premier. On risque en particulier d'obtenir de mauvaises estimations du nombre de logements. Des études ont noté qu'il y avait des divergences entre l'enquête emploi annuelle et l'enquête logement pour l'estimation de cette variable. Ces études ont par ailleurs souligné le fait que les poids obtenus après le premier calage (extrl) permettaient d'estimer correctement le nombre de résidences principales, ce qui n'était plus le cas après le deuxième calage avec le nouveau jeu de pondérations (extri). Notons enfin que la méthode considérée ici diffère de celle employée pour l'enquête annuelle dans la mesure où ici, nous redressons le premier échantillon directement sur les totaux du RP, tandis que pour l'enquête annuelle, l'échantillon est redressé sur les totaux de l'échantillon théorique.

- Le deuxième calage, calage 2, reprend les variables du calage 1 (tranche d'unité urbaine, type de logement, nombre de pièces et âges des individus) mais le calage est réalisé cette fois en une unique étape.

  Les avantages et inconvénients sont inverses de ceux du calage 1: de meilleures estimations puisque le premier calage n'est pas détruit, au prix d'un peu opacité puisqu'il n'est plus possible de dissocier la variation de poids entre non-réponse et fluctuations d'échantillonnage.

- Le troisième calage, calage 3, reprend la méthode précédente en y ajoutant la variable logement neuf.

  L'avantage est bien sûr de mieux estimer la construction neuve depuis le RP en l'utilisant comme variable de calage. Cette information est fournie par le Ministère de l'Equipement (source SITADEL) et est utilisée par la division Logement dans la repondération de son enquête. Cette variable peut toutefois présenter des défauts de qualité et il faut en tenir compte pour décider d'intégrer ou non cette dernière. L'enquête Emploi annuelle en particulier a arrêté de caler sur cette source depuis 1995 alors qu'elle le faisait auparavant, mais le Ministère de l'Equipement a depuis fait des efforts pour augmenter la qualité de cette source.

- Le quatrième calage, calage 4, part du calage 3 et ajoute les variables nombre de chômeurs à la date t-1, nombre de personnes de nationalité française à la date t-1, nombre de personnes à temps partiel à la date t-1. L'introduction de ces variables à la date t-1 ne peut évidemment se faire que sur les sous-échantillons de 2 à 6 pour lesquels on dispose de ces variables[5].

  L'introduction de ces variables est motivée par le fait qu'elles sont elles aussi explicatives de la non-réponse. De plus, il est possible d'espérer que l'introduction de ces variables permettra de diminuer le biais de rotation mis en évidence sur l'enquête annuelle[6].

[5] L'écriture mathématique d'un tel calage diffère des expressions données précédemment, et il est important de bien comprendre ce qui est fait. Le calage s'écrit de la manière suivante :

$$\sum_{k \in E_{t-1} \cap E_t} w_k x_k (t-1) = \sum_{k \in E_{t-1}} w_k (t-1) x_k (t-1)$$

où E_t et E_{t-1} représente les échantillons de répondants aux dates t et t-1. On cherche donc à redresser les répondants à la date t, qui avaient répondu à la date t-1, pour estimer parfaitement des totaux à la date t-1.

[6] Le biais de rotation a été mis en évidence dans l'enquête annuelle. L'estimation du nombre de chômeurs dans le premier sous-échantillon (tiers entrant) était toujours supérieur à l'estimation obtenue dans le deuxième sous-échantillon (tiers médian), qui lui même était toujours supérieur à l'estimation obtenue avec le troisième et dernier sous-échantillon (tiers sortant). Une explication possible consiste à penser que ce sont les chômeurs qui souvent ne répondent plus lors des réinterrogations. Le calage sur les variables t-1 permet de corriger ce type de non-réponse.

Le tableau suivant résume simplement les différents types de version et de calage que nous allons donc étudiés ainsi que leur appellation dans la suite du papier :

Tableau 3 : résumé des méthodes de calage utilisé

	Version 1 : Pyramide de la population totale	*Version 2 : Pyramide de la population des ménages ordinaires*
Calage en deux étapes (tranche d'unité urbaine, type de logement, nombre de pièces)	V1C1	V2C1
Calage en une étape (tranche d'unité urbaine, type de logement, nombre de pièces)	V1C2	V2C2
Calage en une étape (tranche d'unité urbaine, type de logement, nombre de pièces, nombre de logements neufs)	V1C3	V2C3
Calage en une étape avec des variables t-1 (tranche d'unité urbaine, type de logement, nombre de pièces, nombre de logements neufs, nombre de chômeurs à la date t-1, nombre de personnes à temps partiel à la date t-1, nombre de personnes de nationalité française à la date t-1)	V1C4	V2C4

2.3 Résultats des différentes méthodes de calage

Nous avons estimé ces 8 types de calage différents et ceci pour six trimestres de l'enquête emploi en continu : le 3ème et le 4ème trimestre 2001, puis le 1er , le 2ème, le 3ème et le 4ème trimestre 2002. Rappelons également que chacun des six sous-échantillons de chaque enquête est traité de manière indépendante par rapport aux autres. Les estimations données dans cette partie pour une année correspondent donc pour la plupart à la moyenne des six estimations calculées sur chaque sous-échantillon.

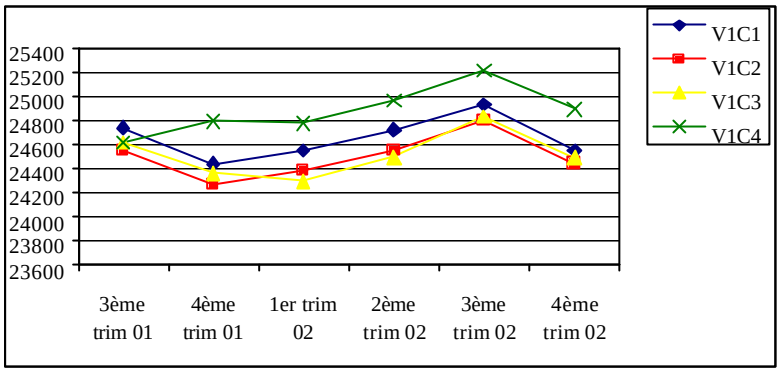
Les estimations obtenues selon les différentes méthodes de calage pour le nombre de résidences principales (i.e le nombre de ménage) et pour la répartition de l'activité au sens du BIT sont comparées aux résultats obtenus à partir de la dernière enquête Emploi annuelle de mars 2002. Le tableau et les graphiques ci-dessous regroupent les principaux résultats.

Tableau 5 : Résultats pour le 1er trimestre 2002

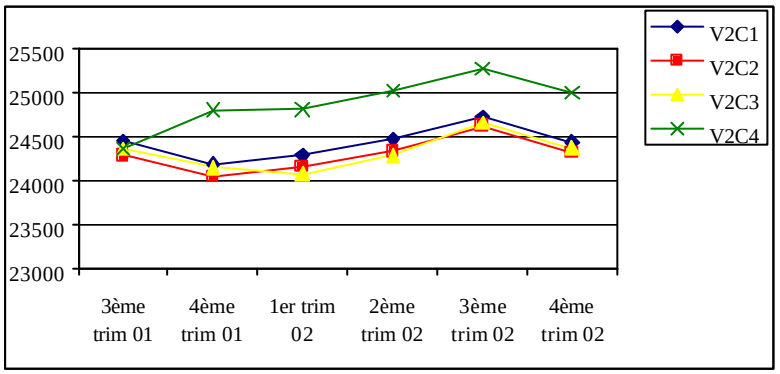
		Nombre de chômeurs BIT (en milliers)	Nombre d'actifs occupés BIT (en milliers)	Taux de chômage BIT (en %)	Nombre de résidences principales au 1er janvier 2002 (en milliers)
Référence		2 406	23 942	9,13	24 500
Version 1	Calage 1	2 450	24 550	9,07	25 378
	Calage 2	2 494	24 387	9,28	24 195
	Calage 3	2 448	24 295	9,15	24 271
	Calage 4	2 457	24 777	9,02	24 295
Version 2	Calage 1	2 429	24 289	9,09	24 862
	Calage 2	2 463	24 155	9,25	23 877
	Calage 3	2 417	24 071	9,12	23 970
	Calage 4	2 477	24 816	9,08	24 269

Ici la référence pour le nombre de chômeurs BIT est donnée par l'enquête emploi annuelle au point de mars 2002, à savoir 2 341 000 chômeurs, auxquels nous avons rajouté 65 000 chômeurs (estimé à partir des évolutions du nombre de Demandeurs d'Emploi en Fin de Mois - DEFM- inscrits à l'ANPE) pour tenir compte des fluctuations saisonnières au sein du 1er trimestre 2002. La référence pour le nombre d'actifs occupés BIT est simplement le nombre d'actifs occupés estimé dans l'enquête emploi annuelle (on suppose ici qu'il n'y a pas de variation saisonnière importante).

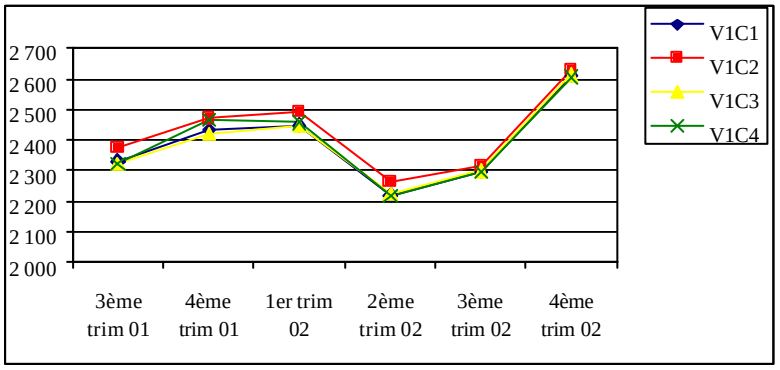
Graphique 5a : Estimations du nombre d'actifs occupés au sens du BIT dans la version 1



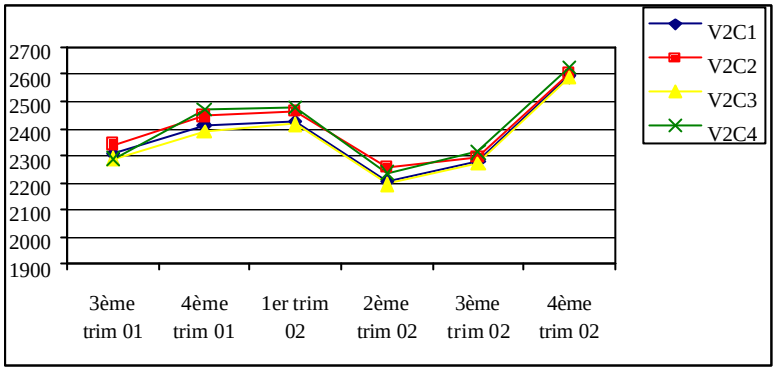
Graphique 5b : estimations du nombre d'actifs occupés au sens du BIT dans la version 2



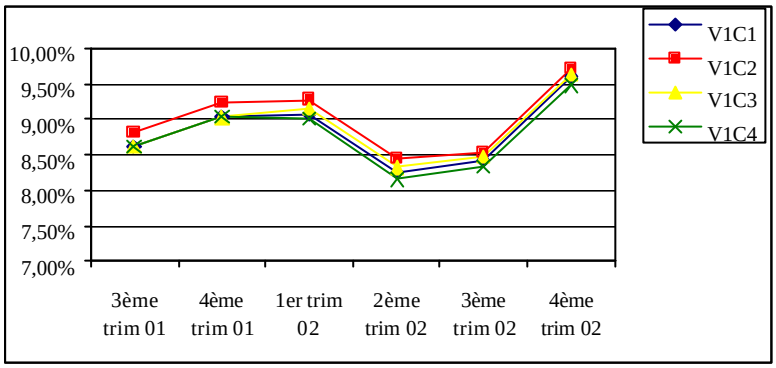
Graphique 6a : estimations du nombre de chômeurs au sens du BIT dans la version 1



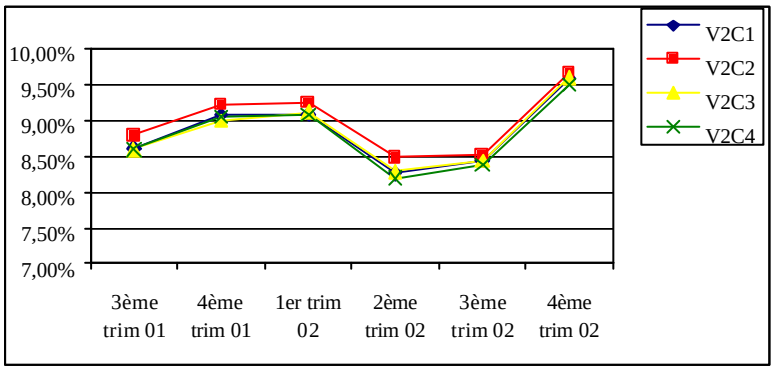
Graphique 6b : estimations du nombre de chômeurs au sens du BIT dans la version 2



Graphique 7a : Estimations du taux de chômage au sens du BIT dans la version 1



Graphique 7b :Estimations du taux de chômage au sens du BIT dans la version 2



Plusieurs observations peuvent être déduites de ces résultats :

La première d'entre elles consiste simplement à remarquer que l'estimation du nombre de chômeurs BIT dans l'enquête emploi en continu au 1er trimestre 2002 est cohérente, quelle que soit la méthode, avec l'estimation du nombre de chômeurs obtenue dans l'enquête annuelle, même si elle le surestime légèrement de manière assez systématique.
De plus, quelle que soit la méthode, le nombre d'actifs occupés au sens du BIT estimé au 1er trimestre 2002 est également supérieur au nombre d'actifs occupés estimés dans l'enquête annuelle. Cette observation peut se justifier par le fait qu'un des objectifs de l'enquête emploi en continu consistait

justement à mieux appréhender le nombre d'actifs occupés et en particulier ceux qui travaillent peu d'heures.

Au total, les taux de chômage BIT sont tous très proches les uns des autres, puisqu'ils varient tous autour de 9,1 %, ce qui est conforme à l'estimation donnée par l'enquête emploi annuelle. Il semble donc qu'il sera possible d'éviter une rupture brutale de cette série.

Le deuxième point repose sur la comparaison entre la version 1 et la version 2 :

- les deux méthodes prédisent des taux de chômage BIT similaires ;
- la version 1 prédit de meilleures estimations pour le nombre de résidences principales ;
- la version 1 a tendance à surestimer le nombre de chômeurs et d'actifs occupés par rapport à la version 2. Ce résultat est en accord avec les estimations du nombre d'actifs occupés à l'aide de sources administratives, basées sur la population totale (comme la version 1), qui surestiment le nombre d'actifs occupés estimés dans l'enquête annuelle, basée sur la population de ménages ordinaires (comme la version 2). De plus, le nombre de chômeurs dans les collectivités est d'environ 40 000[7] et seule une partie d'entre eux sont pris en compte dans la version 2, alors que la version 1 les englobe tous en théorie. Les résultats obtenus semblent donc raisonnables de ce point vue et il semble que la version 1 réussisse mieux à estimer le chômage dans ces populations;
- l'évolution des différentes variables sur la période considérée sont les mêmes dans les deux versions.

Le troisième point concerne les comparaisons entre les différentes méthodes de calage.

- Les quatre méthodes donnent des résultats presque identiques à la fois en niveau et en évolution pour l'estimation du nombre de chômeurs et du taux de chômage. De plus, l'évolution du nombre de chômeurs est cohérente avec l'évolution observée dans les DEFM ;
- Les trois premières méthodes donnent également des résultats très similaires sur le nombre d'actifs occupés, tandis que le calage 4 surestime celui-ci par rapport aux autres méthodes. L'évolution est en revanche la même dans les quatre méthodes[8] ;
- il n'en est pas de même pour l'estimation du nombre de résidences principales où on retrouve le fait que, comme dans l'enquête emploi annuelle, le calage 1 ne donne pas des résultats proches de ceux de l'enquête logement, ce qui peut être gênant. Le calage 3, qui semble meilleur que le calage 2 où le nombre de logements neufs est sous-estimé, et le calage 5 donnent les meilleurs résultats avec des estimations plus proches de celles de l'enquête logement ;
- le calage 3 a encore l'avantage de fournir de meilleures estimations de l'évolution du nombre de logements sur la période considérée (graphe non présenté), en particulier lorsqu'on se restreint aux sous-échantillons correspondant au RP99. Les autres calages donnent des évolutions plus marquées qui sont plus difficilement interprétables ;
- enfin, les quatre calages donnent une évolution du nombre d'actifs occupés qui semblent indiquer des mouvements saisonniers assez importants. L'évolution sur l'année 2002, estimée par la différence entre le 4ème trimestre 2002 et le 4ème trimestre 2001 varie entre 100 000 et 150 000 suivant la méthode de calage utilisée. Le calage 3, qui prédit une hausse de 100 000 actifs sur l'année 2002, fournit les résultats les plus proches de ceux publiés dans la note de conjoncture (95 000).

[7] Ce chiffre est obtenu à partir du RP. Si le concept de chômage au RP n'est pas le même que le concept BIT, ce chiffre n'en est pas moins une indication de l'ordre de grandeur des différences raisonnables qu'on peut attendre entre les deux méthodes.
[8] L'évolution du premier point pour le calage 4 est difficilement interprétable puisque le 3ème trimestre 2001 n'a pas pu être estimé par cette méthode (il n'y a pas d'information passée pour la première vague d'enquête) et a donc été fixé au niveau donné par le calage 3.

Conclusion

Les estimations obtenues avec l'enquête emploi en continu sont cohérentes avec les résultats obtenus avec l'enquête emploi annuelle. La comparaison des différentes versions et calage semble privilégier un calage en une étape sur un nombre restreint de variables RP (tranche d'unité urbaine, type de logement, nombre de pièces et âges des individus) et le nombre de logements neufs, en utilisant la pyramide des âges totale (version 1 avec le calage 3). C'est en effet cette méthode qui donne les résultats les plus proches des points de référence et qui assure une évolution temporelle raisonnable.

Bibliographie

[1] Bosredon J., et Fevrier P. (2002), « Estimations dans l'enquête Emploi en continu », in Actes des Journées de Méthodologie Statistique, Insee Méthodes n°100, pp 393-414.

[2] Caron, N, et Ravalet P. (2002), « estimation dans les enquêtes répétées : application à l'enquête emploi en continu », in Actes des Journées de Méthodologie Statistique, Insee Méthodes n°100, pp 327-392.

[3] Christine M. (2002), « La construction de l'échantillon de la future enquête emploi en continu à partir du recensement de 1999 », in Actes des Journées de Méthodologie Statistique, Insee Méthodes n°100, pp 175-229.

[4] Deville J.C., et Särndal C.E. (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, 87, 11, pp. 376-382.