PONDÉRATIONS LONGITUDINALES DANS L'ENQUÊTE EMPLOI DE L'INSEE

Pascal Ardilly

Insee, Département des méthodes statistiques, 165 Bd Garibaldi 69003 Lyon, France
pascal.ardilly@insee.fr

Résumé. L'enquête trimestrielle sur l'Emploi permet de former un échantillon dit 'longitudinal' d'individus enquêtés à l'occasion de deux trimestres distincts et de le pondérer de manière à estimer sans biais les transitions d'activité BIT (actif occupé / chômeur / inactif) entre ces deux trimestres. La difficulté provient de la non-réponse due à l'érosion, que l'on peut soupçonner d'être non-ignorable. Une technique de calage permet de former les poids longitudinaux. Il est aussi possible d'utiliser une estimation explicite des probabilités de réponse. Le calage a une double vertu car outre le traitement de la non-réponse, il permet d'assurer la cohérence des statistiques issues de l'échantillon longitudinal avec celles qui proviennent des échantillons transversaux. Différents scénarios relatifs à des jeux distincts de variables de calage sont testés. On constate que la pondération longitudinale a tendance à augmenter significativement les effectifs d'individus dont l'activité BIT change au cours de la période étudiée.

Mots-clés. Poids longitudinaux, calage, correction de la non-réponse, activité BIT

1 Contexte et objectifs

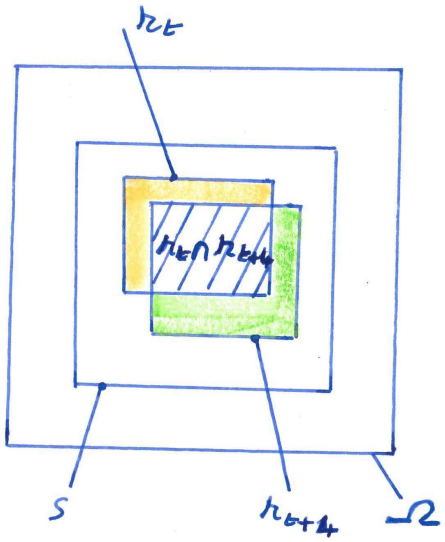
Parmi les statistiques importantes produites à partir de l'enquête Emploi trimestrielle en France (EEC), figurent les transitions entre les trois états d'activité au sens du Bureau international du travail (BIT) : actif occupé, chômeur, inactif. Il s'agit d'estimer, sur une période qui peut être le trimestre ou l'année, une matrice de transition entre états de dimension 3 x 3. L'estimation des flux de transition repose sur un échantillon de personnes physiques dit 'longitudinal'. Cet échantillon regroupe les individus qui fournissent les informations relatives à leur activité BIT à la fois aux dates de début et de fin de période. Pour l'obtenir, on forme l'intersection de l'échantillon transversal (trimestriel) de la date de début de période avec l'échantillon transversal de la date de fin de période, en retenant seulement les individus qui répondent aux deux dates. La construction du poids longitudinal va répondre à deux types d'exigence :

- d'une part, les exigences *techniques*, qui conduisent à rechercher un poids sans biais (en pratique peu biaisé) et produisant une faible variance (si possible). La réduction du biais renvoie aux techniques de correction de la non-réponse et la réduction de variance s'obtient par la mise en œuvre d'un redressement. Pour ce qui est de la pondération transversale, la pratique actuelle de l'Insee consiste à traiter les deux objectifs simultanément au travers d'un calage dit « en une étape » - pratique dont il est naturel de tenir compte pour définir la stratégie de pondération longitudinale.

- d'autre part les exigences de *communication*. En la circonstance, elles sont essentielles car les échantillons longitudinaux seront diffusés auprès d'utilisateurs divers qui bien évidemment pourront les exploiter pour produire des statistiques en niveau relatives respectivement aux dates de début et de fin de période (malgré toutes les précautions et mises en garde qu'on voudra bien formuler…). Il est impossible de pondérer l'échantillon longitudinal de façon à retrouver toutes les statistiques transversales des dates de début et de fin de période. En revanche, on peut pondérer cet échantillon de manière à retrouver certaines statistiques transversales. Le choix de ces statistiques est par nature conventionnel.

2 Eléments théoriques

Ignorons pour l'instant la question de la définition de la population d'inférence. Pour simplifier, plaçons-nous dans le cas d'une évolution annuelle (l'évolution trimestrielle relève exactement de la même théorie). L'enquête Emploi est une enquête à échantillon rotatif, dont un sixième est renouvelé chaque trimestre. De fait, chaque trimestre, l'échantillon transversal est constitué de six sous-échantillons panélisés. Dans un sous-échantillon donné, un échantillon de logements s donnant lieu à la date t_0 à un échantillon d'individus $\tilde{s}_0$ a été tiré (selon un plan complexe). Le trimestre t, cet échantillon de logements - qui est fixe dans le temps - donne lieu au sous-échantillon d'individus répondants r_t. Le trimestre $t+4$, donc un an plus tard, l'échantillon d'individus répondants dans ces mêmes logements est r_{t+4}. Puisqu'une année s'est écoulée, l'intersection entre r_t et r_{t+4} est assez grande, mais il n'y a pas d'inclusion loin de là, et la configuration que l'on rencontre est du type :



L'intersection $r_{t+4}^{t} = r_t \cap r_{t+4}$ est l'échantillon d'individus longitudinal (zone hachurée). Pour un trimestre t fixé, l'enquête permet de mobiliser à cette fin deux sous-échantillons en tout et pour tout : le sous-échantillon entrant et celui qui est interrogé pour la seconde fois (du fait de la rotation trimestrielle, les quatre autres sont sortis avant le trimestre $t+4$). La non-identité entre r_t et r_{t+4} résulte de phénomènes divers. Primo, il faut tenir compte de tous les changements de ménage (déménagements) - le sous-échantillon étant un vrai panel de logements mais pas de ménages. Secundo, il y a au cours du temps des modifications de

périmètre des ménages (une personne part, une personne arrive…). Tertio, on subit le phénomène de non-réponse proprement dite du ménage, de manière évolutive selon le rang d'interrogation, en particulier l'attrition qui est due souvent à la lassitude. Enfin, bien que ce soit assez marginal, il peut y avoir de la non-réponse individuelle au sein d'un ménage considéré comme répondant. Il y a donc *in fine* de nombreuses raisons pour que r_{t+4} diffère significativement de r_t.

Les logements i de s ont tous un poids de sondage noté $w_i^{(1)}$. Ce poids est aussi le poids de sondage transversal des individus résidant dans les logements tirés, quelle que soit la date considérée. Du fait de la non-réponse, l'individu i de r_t a un poids (approche transversale « classique ») noté $w_i^{(2)}$ et l'individu i de r_{t+4} a un poids $w_i^{(3)}$. Ces poids sont aussi les poids des ménages auxquels appartiennent les individus, ce qui résulte du fait que dans l'approche transversale tous les individus d'un même ménage ont même poids (par construction).

A la date t, l'estimation (sans biais) $\hat{Y}^t$ du total Y^t de n'importe quelle variable individuelle Y_i^t est donc

$$\hat{Y}^t = \sum_{i \in r_t} w_i^{(2)} \cdot Y_i^t$$

L'objectif recherché ici est de pouvoir utiliser l'échantillon r_{t+4}^t pour produire des estimations d'évolutions entre t et $t+4$ à partir d'une variable individuelle $\Delta_i^{t/t+4}$. Puisqu'on s'intéresse aux matrices de transition de manière privilégiée, par exemple $\Delta_i^{t/t+4}$ vaudra 1 pour tous les individus i qui passent au cours de la période de l'état de chômeur à l'état d'actif occupé et 0 pour tous les autres. On peut ainsi estimer le nombre total d'individus qui passent d'un état à l'autre au cours d'un intervalle de temps glissant d'une année. Cela suppose qu'il existe un poids « longitudinal » individuel w_i^L adéquat permettant de former l'estimateur (sans biais)

$$\sum_{i \in r_{t+4}^t} w_i^L \cdot \Delta_i^{t/t+4}$$

Disposer des estimations des effectifs pour toutes les cases de la matrice de transitions (neuf cases en la circonstance) conduit naturellement à disposer des marges estimées, donc des effectifs par modalité d'activité, estimés en toute généralité aux trimestres respectifs t et $t+4$. Or ces effectifs sont tous estimés par ailleurs en utilisant les échantillons trimestriels, dits transversaux. Les estimations transversales sont notées Z_t^* de façon générale pour toute variable individuelle indicatrice d'activité Z_i^t relative au trimestre t et elles forment des effectifs de référence : successivement, on obtient le nombre total de chômeurs, le nombre total d'actifs occupés et le nombre total d'inactifs. Pour les raisons techniques et stratégiques exposées dans la partie 1, moyennant certaines adaptations préalables et incontournables qui seront précisées plus loin, **on s'impose dans tous les cas de figure, pour chacun des trois états d'activité et pour chacune des deux dates concernées, une égalité entre les estimations obtenues via l'échantillon longitudinal et celles obtenues via l'échantillon transversal**, soit neuf égalités au total, du type

$$\sum_{i \in r_{t+4}^{t}} w_i^L \cdot Z_i^t = Z_t^* \quad \text{et} \quad \sum_{i \in r_{t+4}^{t}} w_i^L \cdot Z_i^{t+4} = Z_{t+4}^*$$

Ces neuf égalités font partie des conventions et elles ne sont jamais remises en cause.

Cela étant, on peut s'attendre à ce qu'un utilisateur de l'échantillon longitudinal procède dans le même esprit pour produire des estimations en évolution avec d'autres variables de collecte de nature qualitative, soit directement liées à l'activité (variables définissant un parcours antérieur en matière d'activité par exemple), soit plus classiquement appartenant au registre sociodémographique (sexe, âge, nationalité, diplôme,…). Dans un second temps, rien n'empêche cet utilisateur d'en tirer des effectifs trimestriels estimés en niveau. Sur le plan stratégique, comme il y a beaucoup de variables de collecte, on ne peut pas garantir la cohérence entre estimations longitudinales et estimations transversales pour chacune d'entre elles. Néanmoins pour un « certain nombre » de variables de collecte au niveau individu X_i^t et X_i^{t+4}, on peut assurer les égalités de type

$$\sum_{i \in r_{t+4}^{t}} w_i^L \cdot X_i^t = X_t^* \quad \text{et / ou} \quad \sum_{i \in r_{t+4}^{t}} w_i^L \cdot X_i^{t+4} = X_{t+4}^*$$

Par exemple X_i^t sera une indicatrice de sexe (à la date de début de période) si on veut assurer la cohérence longitudinale et transversale (à la date de fin de période) en estimant le nombre d'hommes et le nombre de femmes. Cet objectif relève purement de la communication et il s'atteint sur le plan opératoire par un calage. Bien évidemment, s'ajoute la préoccupation technique, qui conduit à retenir parmi les variables de calage celles qui sont explicatives des transitions d'activité. Et comme il y a également une correction de non-réponse à effectuer si on pratique un calage en une seule étape, il faut en sus que X_i^t soit explicative du comportement de réponse. La différence entre les Z_i^t et les X_i^t tient au caractère plus ou moins contraint de la cohérence recherchée, les variables X_i^t constituant un ensemble à géométrie variable. En conclusion, on place dans la liste des variables X_i^t toutes celles pour lesquelles on recherche une cohérence d'affichage transversal / longitudinal ainsi que toutes les variables repérées et disponibles expliquant à la fois les transitions d'activité BIT et le comportement de réponse individuel.

Au final, il s'agit donc, ni plus ni moins, de résoudre un système d'équations de calage dans lesquelles on décide d'inclure systématiquement les structures d'activité à la date de départ et à la date d'arrivée, et auxquelles on convient d'ajouter telle ou telle variable, soit pour représenter une situation à la date de début de période, soit pour représenter une situation à la date de fin de période - voire aux deux dates simultanément.

Cette méthodologie a pour conséquence forte d'imposer en amont certaines adaptations au niveau des effectifs sont lesquels on cale. En effet, la somme des poids calés représente toujours l'estimation de la taille totale de la population d'inférence. Même si celle-ci n'est pas définie de façon claire, dans tous les cas de figure il n'y a qu'une seule taille de population envisageable ! Pour cette raison, la somme des effectifs estimés relatifs aux différentes modalités de n'importe quelle variable qualitative est une constante systématiquement égale à la somme des poids calés. Les populations ayant la mauvaise habitude d'évoluer avec le temps, cela explique qu'il est impossible d'assurer conjointement un calage sur des effectifs à

la date de départ et sur des effectifs à la date d'arrivée sans avoir préalablement effectué une opération de mise en cohérence des marges. La mise en cohérence est laissée libre et on peut faire ce que l'on veut dès lors qu'on aboutit à une unique taille de population. L'approche retenue dans les travaux effectués à l'Insee en 2015 a consisté à **conserver les rapports entre les effectifs relatifs aux différentes modalités et à imposer à la taille totale de la population à la date de fin de période d'être égale à la taille totale de la population à la date de début de période**.

Cela se fait très simplement par des « règles de trois ». Si la taille de population estimée à la date de début de période t avec l'échantillon transversal considéré à t est $\hat{N}_t$, si l'échantillon transversal à $t+4$ estime à la valeur $\hat{N}_{t+4}^k$ le nombre total d'individus vérifiant la modalité k d'une variable qualitative *lambda* à K modalités, on transforme chaque marge ainsi :

$$\tilde{\tilde{N}}_{t+4}^k = \frac{\hat{N}_{t+4}^k}{\sum_{k=1}^{K} \hat{N}_{t+4}^k} \cdot \hat{N}_t$$

Une possibilité alternative serait de maintenir $K-1$ des effectifs (au choix) à leur valeur initiale $\hat{N}_{t+4}^k$ et d'adapter la valeur de l'ultime effectif $\hat{N}_{t+4}^K$ selon $\hat{N}_{t+4}^K = \hat{N}_t - \sum_{k=1}^{K-1} \hat{N}_{t+4}^k$.

L'avantage de la première méthode est de conserver tous les ratios, puisque pour tout couple de modalités (k,l) on a

$$\frac{\tilde{\tilde{N}}_{t+4}^k}{\tilde{\tilde{N}}_{t+4}^l} = \frac{\hat{N}_{t+4}^k}{\hat{N}_{t+4}^l}$$

C'est ainsi qu'on doit modifier le nombre d'actifs occupés, de chômeurs et d'inactifs sur lesquels on se cale à la date de fin de période $t+4$, en contrepartie de quoi on peut assurer que le taux de chômage, le taux d'activité et le taux d'inactivité que permettra d'obtenir l'exploitation de l'échantillon longitudinal à la date $t+4$ seront parfaitement en cohérence avec la statistique officielle relative à la situation $t+4$, c'est-à-dire ce qui vient de l'échantillon transversal à $t+4$. La seconde méthode permettrait (par exemple) d'estimer parfaitement le nombre d'actifs occupés et le nombre de chômeurs à $t+4$ (donc le taux de chômage) mais en revanche les taux impliquant les inactifs ne répondraient à aucune logique. De manière symétrique, on peut aussi imaginer de caler sur la population à la date d'arrivée, soit $\hat{N}_{t+4}$ au lieu de $\hat{N}_t$.

Le calage fait intervenir trois dimensions : l'échantillon, les variables de calage (fichier de collecte d'une part, marges de l'autre, mais en la circonstance les marges sont très particulières puisqu'elles sont issues du fichier de collecte, au moins pour une grande partie d'entre elles) et les poids initiaux. Le reste suit. En la circonstance, la première composante est complètement définie, s'agissant de l'échantillon d'individus physiques r_{t+4}^t, dit 'longitudinal'. La seconde dimension offre beaucoup plus de liberté puisqu'on a vu que l'on peut adapter à sa convenance (plus ou moins, car il faut que l'équation de calage ait une solution mathématique…) la liste des variables de calage de type X_i^t. Évidemment, on peut à ce niveau inclure et combiner des variables au choix issues du fichier de collecte presque à l'infini. Il y a une variante qui consiste à inclure dans cette liste des variables de la base de

sondage, lesquelles pourraient être en tout ou partie des variables actuellement mobilisées pour le calage des échantillons transversaux. La troisième composante est la plus subtile et permet d'engager des scénarios (plus ou moins) distincts : il s'agit du poids initial (fourni en entrée à Calmar) dont le poids calé w_i^L cherche à être le plus proche possible. Diverses possibilités semblent trouver une justification, mais les deux scénarios suivants ont été privilégiés, en rappelant que la théorie se conçoit sous-échantillon par sous-échantillon :

** Scénario 1 de pondération*

Considérant la propriété d'inclusion $r_{t+4}^t \subset r_t$, on peut partir du poids transversal individuel $w_i^{(2)}$, lequel intègre la correction pour non-réponse à la date de départ. On reste dans l'esprit du calage en une étape puisque r_{t+4}^t est considéré comme un échantillon tiré dans la population complète des individus présents à t : dans ces conditions, ce calage est sensé corriger seulement la non-réponse qui se traduit par la perte d'informations individuelles entre t et $t+4$, c'est-à-dire la non-réponse à $t+4$ sachant qu'on a répondu à t. Bien que significative, cette non-réponse reste modérée à échéance d'une année, donc la repondération associée ne devrait pas être violente 'en moyenne ' (même s'il faut s'attendre à des évolutions fortes de certains poids dès lors que certaines marges traduisent des sous-populations de petite taille). Par ailleurs, cette optique est bien compatible avec le calage sur une batterie de variables X_i^t liées à la collecte, certaines étant bien explicatives à la fois du comportement de réponse et des variables liées à l'activité.

** Scénario 2 de pondération*

On abandonne l'optique de calage en une étape et on adopte le traitement - plus traditionnel - de correction des poids au travers d'une probabilité de réponse estimée. On part de $w_i^{(2)}$ (comme dans le scénario 1) mais on estime la probabilité de réponse $P(i \in r_{t+4} | i \in r_t)$ par un modèle adapté. Le poids en entrée de Calmar est alors

$$\frac{w_i^{(2)}}{\hat{P}(i \in r_{t+4} | i \in r_t)}$$

En pondération transversale, il est souhaitable que tous les calages s'effectuent au niveau ménage, pour deux raisons : primo il y a des variables d'intérêt définies au niveau ménage (le ménage est une unité statistique intéressante quand on se place à une date fixée), secundo la non-réponse 'ponctuelle' (ou faut-il dire 'instantanée' ?) est un phénomène qui se situe presque exclusivement au niveau ménage (grâce au système de 'proxy'), et enfin c'est la seule façon de garantir que tous les individus d'un ménage donné aient même poids en fin d'opération. En approche longitudinale au contraire, toutes les variables d'intérêt sont fondamentalement des variables définies au niveau individuel car le ménage est une entité instable dans le temps (aussi courte que soit la période couverte). Par ailleurs, la non-réponse comprend une composante individuelle indéniable qui ne peut pas se transférer au niveau du ménage (cf. ci-dessous). Aussi, le principe consistant à imposer le même poids pour tous les membres du ménage paraît beaucoup plus difficile à assurer et même à justifier. C'est pourquoi les calages n'ont rien imposé au niveau ménage - d'ailleurs le niveau 'ménage' a été totalement ignoré dans toute la procédure. Aussi, *in fine*, **chaque individu physique a un poids qui lui est propre**.

Pour traiter la non-réponse des individus, on peut s'interroger sur les parts explicatives qui reviennent respectivement à l'individu et au ménage dans son ensemble. En effet, à cause de l'effet « proxy », le statut répondant / non répondant d'un individu est dû en partie à la composition du ménage. Mais il y a en sus et à l'évidence une forte composante de nature spécifiquement 'individuelle', qui se traduit en particulier par des modifications du périmètre du ménage (l'individu peut quitter le ménage, dont la composition évolue avec le temps). Il importerait de savoir si l'individu disparu reste dans le champ de l'enquête. Cette dernière condition semble hélas invérifiable[1] au moment de la collecte, par conséquent il est impossible de mettre en place une pondération individuelle rigoureuse (noter qu'on a exactement le même problème si le ménage dans son intégralité quitte le logement !). Ce problème n'existe pas dans la problématique transversale, où on a coutume de dire que la non-réponse individuelle « instantanée » est négligeable : effectivement, grâce au « proxy », il est très rare si le ménage répond qu'un individu donné de ce ménage soit spécifiquement non-répondant.

Quelle que soit la méthode utilisée (calage en une étape ou estimation explicite de probabilités de réponse), le choix des variables explicatives des probabilités de réponse individuelles est essentiel parce que les phénomènes motivant les disparitions du panel des individus ont toutes les raisons d'être corrélés à l'activité : les individus qui quittent le logement ont des transitions spécifiques, c'est-à-dire qu'on est *a priori* en contexte de non-réponse non-ignorable.

Si on accepte de produire une pondération au niveau individu (et non ménage), il est naturel de prendre en compte des variables explicatives relatives au ménage, dont on ne peut pas nier le caractère important pour expliquer la non-réponse des individus : dans cet esprit, ont été retenues une variable de type de ménage et deux variables géographiques (la tranche d'unité urbaine et l'appartenance à une Zone Urbaine Sensible (ZUS) - la tranche d'unité urbaine est bien connue pour être un fort déterminant de la propension du ménage à répondre). La variable 'type de ménage' propose des modalités qui caractérisent assez bien la taille totale du ménage - dont on sait qu'elle constitue une variable très explicative de la probabilité qu'a l'enquêteur d'obtenir les réponses de l'ensemble des membres du ménage.

Pour ce qui concerne le traitement de la non-réponse, le scénario 2 est probablement préférable sur le plan théorique à cause des limites bien connues attachées au calage en une étape mais il s'appuie sur une technique de pondération qui n'est pas celle de l'enquête transversale, ce qui peut être un obstacle en terme de communication. Lorsque le traitement des non-réponses entre t et $t+4$ est intégralement[2] assuré par un calage (scénario 1), les exigences en matière d'information sont différentes de celles qui prévalent lorsqu'on estime en amont des probabilités de réponse : avec le calage, il n'est pas utile de disposer de l'information explicative individuelle pour les non-répondants, en revanche elle doit être connue de manière agrégée sur toute la population (la marge doit naturellement être calculée sur la population entière du champ). En la circonstance, dans le match qui pourrait opposer les

[1] En tout cas on n'a pas systématiquement cette information. On sait par exemple bien repérer les cas de non-réponse individuelle pour cause de refus ou parce que le proxy ne sait pas ou ne veut pas répondre. En revanche si un individu a quitté le ménage, on n'a plus d'information le concernant, on ne sait pas s'il réside toujours en France et en ménage ordinaire, s'il est toujours vivant,…

[2] Au sens où il n'y a aucune intervention spécifique du statisticien pour procéder à l'estimation des probabilités de réponse. Dans le calage « standard » en une étape, la prise en compte de la non-réponse est transparente (bien qu'elle relève d'hypothèses fortes, mais implicites !).

scénarios 1 et 2, il faut voir que l'échantillon transversal qui sert de base à la formation de l'échantillon longitudinal offre une information extrêmement riche, aussi bien au niveau des répondants de l'échantillon longitudinal (qui en est un sous-échantillon) qu'au niveau du calcul des marges : cela fait que dans ces circonstances très particulières, pour ce qui concerne le traitement de la non-réponse et en terme d'information à mobiliser, la technique de calage perd à la fois son avantage et son inconvénient par rapport à la technique de traitement en deux étapes !

Compte tenu de ces éléments, le scénario 1 apparaît *a priori* comme un bon compromis entre simplicité, communication et efficacité. Il a l'avantage d'être cohérent avec la pratique de pondération actuellement utilisée à l'Insee (calage en une étape) pour les échantillons transversaux. De plus, argument supplémentaire, les applications numériques conduisent à des matrices de transitions identiques ou quasi identiques à celles du scénario 2 - donc autant privilégier la méthode la plus simple à appliquer pour aboutir aux mêmes résultats !

Reste *in fine* la **question de la population d'inférence**. Traditionnellement, on précise la population sur laquelle on définit les paramètres. Dans le cas présent, il n'y a hélas pas d'éclairage satisfaisant à apporter à cette question. En effet, les calages prétendent produire une inférence sur deux populations à la fois (population Ω_t des individus du trimestre t et population Ω_{t+4} des individus du trimestre $t+4$ - laquelle ne peut pas être la même que la première…). L'opération de calage sur la population d'arrivée est donc de nature « cosmétique », ce qui n'enlève rien à sa respectabilité, mais de ce fait il vaut mieux ne pas trop s'attarder sur la nature de l'inférence. En la circonstance, la moins mauvaise position consiste à dire qu'à échéance d'une année les populations de personnes physiques ne changent encore pas trop. En revanche, si la pondération doit couvrir des périodes plus larges (plusieurs années), le problème peut devenir crucial.
Au demeurant, les normalisations préalables des effectifs par modalité des variables qualitatives définies dans la population d'arrivée traduisent une forme de confusion quant à la définition de la population d'inférence et d'une certaine façon il devient plus acceptable d'utiliser la technique moins naturelle et moins robuste de calage en une étape.

Les calages effectués pour pondérer l'échantillon longitudinal ont porté sur son intégralité et non pas sous-échantillon par sous-échantillon (comme cela se fait pour la pondération transversale). Utiliser plusieurs calages a semblé introduire un raffinement qui ne se justifiait pas, d'autant plus qu'il n'y a pas de distinction entre les sous-échantillons quant à la disponibilité de l'information sur l'activité utilisée pour produire les estimations longitudinales essentielles[3].

3 Les résultats de l'estimation sur période annuelle, du T1 2014 au T1 2015

Les travaux menés récemment à l'Insee et utilisant des pondérations longitudinales ont produit des estimations de transitions sur deux périodes annuelles et sur deux périodes trimestrielles. Dans tous les cas, ce sont les mêmes phénomènes qui apparaissent, mais ils sont plus marqués sur la période annuelle. On présente ici le cas d'une de l'estimation annuelle couvrant la période du T1 2014 au T1 2015. Le champ retenu est formé par l'ensemble des

[3] La justification essentielle du calage par sous-échantillon, dans l'approche transversale tient au fait que certaines variables ne sont collectées que pour les sous-échantillons entrant, voire entrant et sortant . Mais il n'y a aucune distinction de cette nature en approche longitudinale.

individus ayant entre 15 et 74 ans, l'âge étant celui déclaré à la date de l'enquête. Le calage de l'échantillon longitudinal peut s'effectuer selon quatre scénarios :

a) Dans l'esprit du scénario de pondération N°1 les variables de calage sont les structures d'activité (variable ACTEU) classiques considérées aux dates respectives de début et de fin de période (après la normalisation requise sur les estimations transversales de la date de fin de période) ;

b) Aux variables précédentes, on ajoute deux variables spécifiques construites à partir de données d'activité antérieures déclarées (déclaration 'spontanée') à la date de début de période : les variables TRANSIT et PARCOUR, qui caractérisent d'une certaine façon le passé de l'individu en matière de statut d'activité sur une année complète précédant la date de début de période ;

c) Aux variables précédentes, on ajoute encore un ensemble de variables sociodémographiques : le sexe, l'âge (en tranches), le diplôme, la nationalité, la catégorie sociale remaniée, le type de ménage, la tranche d'unité urbaine, et l'appartenance à une ZUS - plus deux variables complexes *ad hoc* liées à l'activité et construites à partir de la nature de l'employeur principal, du type de contrat et de l'ancienneté au chômage. Trois de ces variables sont exploitées à la fois aux dates de début et de fin de période ;

d) On applique cette fois le scénario de pondération N°2, en procédant à une estimation préalable des probabilités de réponse, par une régression logistique avec sélection *stepwise* des régresseurs. On distingue une version non pondérée et une version pondérée de la régression. Les variables initialement mobilisées pour expliquer ces probabilités et les variables de calage sont celles du traitement c) mais uniquement pour ce qui concerne la date de début de période.

Pour construire la variable TRANSIT, on considère les déclarations spontanées d'activité mensuelle en se plaçant à la date de début de période et en retenant l'activité déclarée[4] le mois d'enquête (SP00) et l'activité déclarée exactement une année auparavant (SP12). Puis on forme les transitions sur la base de ces deux situations, exactement de la même façon que pour les variables ACTEU qui définissent la matrice de transition, soit :

Si SP12 = 1 ou 2 et SP00 = 1 ou 2,	alors transit = 1 ; (activité / activité)
Si SP12 = 1 ou 2 et SP00 = 4,	alors transit = 2 ; (activité / chômage)
Si SP12 = 1 ou 2 et SP00 différent de 1, 2, 4	alors transit = 3 ; (activité / inactivité)
Si SP12 = 4 et SP00 = 1 ou 2,	alors transit = 4 ; (chômage / activité)
Si SP12 = 4 et SP00 = 4,	alors transit = 5 ; (chômage / chômage)
Si SP12 = 4 et SP00 différent de 1, 2, 4	alors transit = 6 ; (chômage / inactivité)
Si SP12 différent de 1, 2, 4 et SP00 = 1 ou 2,	alors transit = 7 ; (inactivité / activité)
Si SP12 différent de 1, 2, 4 et SP00 = 4,	alors transit = 8 ; (inactivité / chômage)
Si SP12 différent de 1, 2, 4 et SP00 différent de 1, 2, 4,	alors transit = 9 ; (inactivité / inactivité)

[4] Modalités de SPxx : 1 ou 2 = actif occupé ; 4 = chômeur ; autres = différents états d'inactivité.

La variable PARCOUR prétend caractériser d'une manière différente - et *a priori* plus fine que TRANSIT - le parcours de l'individu sur une période de 12 mois consécutifs. On mobilise alors les 12 variables d'activité mensuelle SPxx déclarées sur les 12 mois précédant le mois de collecte, plus l'information du mois de collecte lui-même (date de début de période). Puis on réduit l'information en retenant 3 états d'activité pour chacun des 12 mois: actif occupé déclaré, chômeur déclaré, inactif déclaré. Partant de là, on forme les parcours sur les 12 mois en distinguant 10 modalités :

L'individu se déclare actif occupé chaque mois : *parcour* = 1
L'individu se déclare chômeur chaque mois : *parcour* = 2
L'individu se déclare inactif chaque mois : *parcour* = 3
L'individu change au moins une fois de situation en oscillant entre les états d'actif occupé et de chômeur (seulement), mais sur les 12 mois il ne déclare qu'une transition : *parcour* = 4
L'individu change au moins une fois de situation en oscillant entre les états d'actif occupé et de chômeur (seulement), mais sur les 12 mois il déclare 2 transitions ou plus : *parcour* = 5
L'individu change au moins une fois de situation en oscillant entre les états d'actif occupé et d'inactif (seulement), mais sur les 12 mois il ne déclare qu'une transition : *parcour* = 6
L'individu change au moins une fois de situation en oscillant entre les états d'actif occupé et d'inactif (seulement), mais sur les 12 mois il déclare 2 transitions ou plus : *parcour* = 7
L'individu change au moins une fois de situation en oscillant entre les états de chômeur et d'inactif (seulement), mais sur les 12 mois il ne déclare qu'une transition : *parcour* = 8
L'individu change au moins une fois de situation en oscillant entre les états de chômeur et d'inactif (seulement), mais sur les 12 mois il déclare 2 transitions ou plus : *parcour* = 9
L'individu, sur les 12 mois, connaît au moins une fois chacun des trois états : ce sont donc les parcours complexes qui forment le complément des modalités précédentes : *parcour* = 10

Pour des raisons liées au protocole de collecte de l'enquête, le statut d'occupation du logement - qui distingue les propriétaires des locataires - ne fait pas partie des variables impliquées dans les calages, ce qui est regrettable car cette information est fortement corrélée à la mobilité.

Au T1 2014 et avant tout filtre sur l'âge, la fraction de l'échantillon transversal utilisé comme base de sélection de l'échantillon longitudinal (donc r_t) comprend 32 672 individus répondants. Parmi ces individus, 5 463 disparaissent dans l'année qui suit (donc $r_t - r_{t+4}^t$), pour toutes les raisons déjà signalées, ce qui fait que l'échantillon longitudinal (donc r_{t+4}^t) comprend exactement 27 209 individus, tous âges confondus. Pour former le champ retenu, le filtre sur l'âge fait passer l'échantillon longitudinal 'opérationnel' à 23 807 personnes.

- *Bilan : comparaison des transitions estimées selon six méthodes*

On donne, dans le tableau synthétique suivant, les neuf proportions caractérisant la matrice des transitions, d'abord dans la version (biaisée) de l'échantillon longitudinal pondéré par le poids transversal (colonne 1 de référence), ensuite lorsque cet échantillon est pondéré par les poids longitudinaux, variant selon le traitement appliqué.

Ventilation de la population d'inférence entre les neuf types de transitions, exprimée en %
Evolution T1 2014 - T1 2015

Transition	Référence (biaisée)	Méthode a)	Méthode b)	Méthode c)	Méthode d) non pondérée	Méthode d) pondérée
AO → AO	51.61	51.74	51.19	51.00	51.00	51.00
AO → C	1.66	1.88	2.05	2.05	2.05	2.05
AO → I	2.94	2.74	3.30	3.31	3.31	3.31
C → AO	2.03	2.19	2.45	2.50	2.52	2.52
C → C	2.47	3.00	2.84	2.67	2.66	2.66
C → I	1.26	1.26	1.17	1.28	1.27	1.27
I → AO	2.01	2.09	2.38	2.51	2.50	2.50
I → C	1.43	1.67	1.67	1.84	1.85	1.85
I → I	34.59	33.43	32.95	32.84	32.84	32.84

Ci-dessous, le même tableau en distinguant seulement deux situations individuelles : soit il y a stabilité du statut d'activité à échéance d'une année (AO → AO ou C → C ou I → I), soit il y a un « changement » de statut (l'une des six autres situations). Autrement dit, la première ligne ('*stabilité*') somme les trois proportions associées à la diagonale de la matrice de transitions, la seconde ligne ('*changement*') somme les six autres proportions.

Ventilation de la population totale selon le degré de stabilité de l'activité, exprimée en %
Evolution T1 2014 - T1 2015

Transition	Référence (biaisée)	Méthode a)	Méthode b)	Méthode c)	Méthode d) non pondérée	Méthode d) pondérée
Stabilité	88.67	88.17	86.98	86.51	86.50	86.50
Changement	11.33	11.83	13.02	13.49	13.50	13.50

La conclusion principale est que l'ajout de marges de calage tend à augmenter significativement la proportion d'individus dont l'activité BIT change au cours de la période considérée (cases 'hors diagonale' de la matrice de transition). L'échantillon longitudinal a donc bien spontanément tendance à être constitué de trop d'individus « stables », essentiellement (pour ne pas dire exclusivement…) parce que la non-réponse agit en ce sens. On peut gagner ainsi plus de 2 points de pourcentage par rapport à la méthode naïve de référence, grâce essentiellement à la transition inactif-inactif. Le « décrochage » le plus spectaculaire intervient lors de l'application de la méthode b), surtout si on considère qu'il s'agit d'introduire seulement deux marges supplémentaires (versus 13 marges supplémentaires quand on passe de b à c). En contrepartie, la dispersion des poids augmente avec le nombre de marges, si bien que l'on prend plus de risques avec la méthode c) en cas d'exploitation de l'échantillon longitudinal sur des sous-populations fines.

Bibliographie

[1] Biausque, A. , Juillard, M. et Lebrère, A. (2012), *Utilisation de l'enquête Emploi en panel-Non-réponse et calage,* Actes des Journées de Méthodologie Statistique de l'Insee.

[2] Clarke, P. et Chambers, R. (1998), *Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage,* Techniques d'enquête, Vol 24, N°2, pp. 133-140.

[3] Jauneau, Y. et Nouël de Buzonnière, C. (2011), *Proposition de pondération longitudinale pour utiliser l'enquête Emploi en panel annuel,* Document de travail Insee N° F1107.