Utilisation de l'enquête Emploi en panel - Non réponse et calage

V.Biausque, M.Juillard, A.Lebrère

1 Introduction

Un intérêt majeur dans l'utilisation de l'enquête Emploi en panel est de pouvoir mesurer des transitions individuelles sur le marché de travail français. Les rapports du CNIS sur les inégalités sociales (Freyssinet J. et al., 2007) ou sur l'emploi, le chômage et la précarité (de Foucauld et al, 2008) ont insisté sur l'importance de telles statistiques permettant une meilleure description et donc une meilleure compréhension du marché du travail. En outre, Eurostat s'intéresse de plus en plus à des indicateurs portant sur les données longitudinales, comme par exemple le pourcentage des 15-64 ans ayant suivi au moins une formation au cours des douze derniers mois ; cette statistique ne pouvant être produite qu'en utilisant le caractère longitudinal et trimestriel de l'enquête Emploi.

Toutefois, le calcul de pondérations longitudinales pour l'enquête Emploi ne va pas de soi. Cette enquête est un panel rotatif de logements dont la méthodologie d'implémentation a été bâtie pour répondre à deux objectifs principaux : mesurer le taux de chômage en coupe et en évolution. Par conséquent, les panels d'individus construits à partir de cette enquête souffrent de plusieurs inconvénients. Tout d'abord, ils représentent une échelle de temps relativement courte : un individu est interrogé trimestriellement et au maximum six fois (au final, les panels les plus longs portent sur 15 mois). Ensuite et surtout, ces panels sont composés par nature d'individus n'ayant pas déménagé sur la période. Ce dernier phénomène est de nature à biaiser sérieusement les estimations naïves sur les transitions individuelles. Ce problème s'apparente à un effet d'attrition endogène dans les panels : la non-réponse longitudinale liée aux déménagements a des risques d'être corrélée aux transitions sur le marché du travail. On suspecte ainsi qu'une utilisation non corrigée des poids de l'enquête Emploi dans la mesure des flux sur le marché du travail surestime les diagonales des matrices de transitions (les personnes immobiles géographiquement pourraient être les personnes les moins mobiles sur le marché du travail).

Afin de construire des pondérations longitudinales à l'enquête Emploi permettant de corriger les problèmes évoqués, il faut dans un premier temps définir soigneusement la population d'inférence et repérer ces individus dans les données. Par exemple, il n'est pas toujours possible dans l'enquête Emploi de faire la différence entre des individus qui sortent du champ pendant la période d'étude et des personnes qui arrêtent de répondre. Des données du recensement viendront alors à l'appui pour comptabiliser le nombre de morts et le nombre de sorties de la métropole dans la période. Outre les déménagements, il faut ensuite savoir quels sont les principaux facteurs explicatifs de la non-réponse longitudinale. Grâce à cette étape permettant de choisir de « bonnes marges », un calage sera alors mis en œuvre afin de rendre le panel d'individus issu de l'enquête Emploi extrapolable à notre population d'inférence.

Une application à l'étude des flux sur le marché du travail 2009 sera alors présentée.

2 L'enquête Emploi en continu

La collecte de la nouvelle enquête Emploi de l'Insee a débuté le premier juillet 2001. La construction de son échantillon (voir Christine et Loonis) permet une collecte en continu sur tout le territoire métropolitain et permet ainsi un suivi trimestriel du marché du travail. L'Insee est par conséquent en mesure de produire chaque trimestre des indicateurs sur l'emploi, le chômage ou l'inactivité.

Des indicateurs en coupe ou en évolution sont ainsi construits. Par exemple, le nombre de chômeurs en France métropolitaine pour un trimestre t est le nombre moyen (sur les 12 ou 13 semaines du trimestre) de personnes âgées de 15 ans et plus, habitant en logement ordinaire et en situation de chômage BIT la semaine d'interrogation. L'évolution de taux de chômage entre les dates t et $t+1$ est simplement la différence entre les taux de chômage mesurés à ces périodes. Notons qu'il s'agit d'une évolution transversale étant donné que la population du champ de l'enquête Emploi a légèrement évolué entre ces deux dates (des personnes sont entrées, d'autres sont sorties du champ).

Le type d'échantillonnage (panel rotatif) en place pour l'enquête Emploi en continu a l'avantage de permettre des estimations performantes aussi bien en coupe qu'en évolution, bien que d'un point de vue théorique, ces objectifs soient contradictoires. En effet, si on note Y_t et Y_{t+1} les taux de chômages en t et $t+1$, et

$$\Delta = Y_t - Y_{t+1} \quad \text{et} \quad \overline{Y} = \frac{1}{2}(Y_t + Y_{t+1})$$

deux indicateurs fondamentaux de l'enquête Emploi. Alors

$$\mathbb{V}(\Delta) = \mathbb{V}(Y_t) + \mathbb{V}(Y_{t+1}) - 2 \cdot Cov(Y_t, Y_{t+1})$$

et

$$\mathbb{V}(\overline{Y}) = \frac{1}{4}(\mathbb{V}(Y_t) + \mathbb{V}(Y_{t+1}) + 2 \cdot Cov(Y_t, Y_{t+1})).$$

On voit bien dans ces deux relations que le rôle de $Cov(Y_t, Y_{t+1})$ est fondamental. En effet, la mesure de l'évolution est d'autant plus précise que $Cov(Y_t, Y_{t+1})$ est élevé, alors que la mesure d'une moyenne sur deux dates est d'autant plus précise que $Cov(Y_t, Y_{t+1})$ est faible. En d'autres termes, pour mesurer précisément une évolution, il est préférable d'utiliser un panel pur, alors que pour mesurer une moyenne annule à partir d'estimations trimestrielles, il est préférable d'utiliser des échantillons indépendants.

L'échantillonnage rotatif mis en place en 2001 par l'Insee consiste à tirer trimestriellement un échantillon panélisé de logements pour une durée limitée de 6 trimestres ; ainsi l'échantillon est renouvelé tous les trimestres par sixième. Dans chaque logement sélectionné, tous les individus de 15 ans et plus s'y rattachant sont interrogés. Les interrogations se font en face à face en première et sixième vague, alors qu'elles se font par téléphone lors des vagues intermédiaires (deuxième, troisième, quatrième et cinquième interrogation). Il est à souligner que les six sous-échantillons panel qui cohabitent chaque trimestre sont des panels de logements et non d'individus ; ceux qui déménagent d'un trimestre sur l'autre ne sont donc pas suivis. On pourra voir Figure 1 une représentation schématique de l'échantillonnage de l'EEC depuis le troisième trimestre 2001 ; les panels y sont numérotés de façon croissante en fonction de leur date d'entrée.

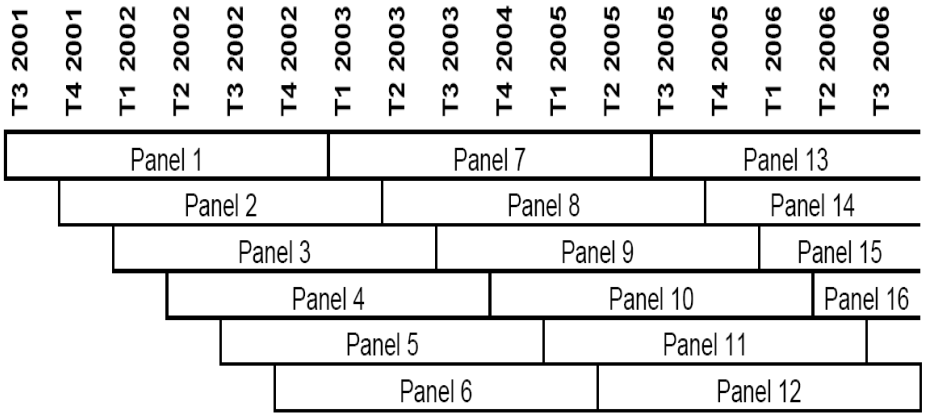

Figure 1 - Échantillonnage de l'enquête Emploi en continu

3 Population d'inférence

3.1 Approches transversale et longitudinale

Le choix de la population d'inférence est crucial pour les indicateurs que l'on souhaite produire. En effet, quand on mesure des évolutions, il est assez différent de raisonner à population constante ou à population évolutive. Reprenons l'exemple de l'évolution du taux de chômage entre deux dates t et $t+1$. Il existe deux approches qui sont en fait complémentaires : l'une transversale prenant en compte l'évolution de la structure de la population entre les deux dates, et l'autre longitudinale qui permet d'étudier l'évolution d'une même population. Imaginons qu'un boom démographique survienne entre les dates t et $t+1$ (un entrée massive d'individus dans le champ), la différence entre ces deux approches peut alors être notable. Dans ce cas, il se peut que l'on observe une hausse transversale du taux de chômage tandis que la proportion longitudinale stagne. Ce phénomène peut s'expliquer par le fait que les « entrants démographiques » de la période ont plus de chance d'être au chômage en $t+1$ que les autres.

3.2 Population d'inférence

Dans le suivi longitudinal des individus entre deux dates t et $t+1$, la population d'inférence est constituée des individus qui sont dans le champ de l'enquête à la fois en t et en $t+1$. Cette population est communément appelée celle des « présent-présent ». Afin de mieux cerner cette population, voici quelques exemples d'individus qui n'en font pas partie :

– les individus qui décèdent ou qui quittent la métropole entre t et $t+1$,

– les individus âgés de 14 ans en t et 15 ans en $t+1$,
– les individus qui ne résidaient pas en logement ordinaire en t, mais qui sont bien dans le champ en $t+1$,
– ...

Par la suite, on notera cette population Ω.

3.3 Marges sur la population d'inférence

3.3.1 Classement des individus

A partir des fichiers de l'enquête Emploi, il est aisé de calculer des marges sur des populations en t et en $t+1$. En revanche, calculer des marges sur notre population d'inférence, celle des « présent-présent », est plus complexe car il faut posséder des informations sur les individus aux deux dates. A titre d'exemple, les individus du sous échantillon sortant en t ne répondent plus en $t+1$!

Chaque fichier peut être divisé en trois catégories :
– les individus longitudinaux (individus du fichier dont on sait qu'ils font partie de la population d'inférence),
– les hors-champ (individus du fichier dont on sait qu'ils ne font pas partie de la population d'inférence),
– les incertains (individus du fichiers pour lesquels on ne peut pas trancher de façon certaine).

On peut encore diviser la catégorie des individus longitudinaux en deux : les répondants longitudinaux (qui ont validé l'enquête Emploi en t et $t+1$) et les non-répondants longitudinaux (les autres).

3.3.2 Calcul des marges

Supposons avoir classé le fichier en t en quatre catégories d'individus : R_t (les répondants longitudinaux), NR_t (les non répondants longitudinaux), HC_t (les hors champ) et I_t (les incertains). Les marges de notre population d'inférence seront calculées à partir des individus de R_t, NR_t et I_t, bien que ce dernier sous fichier puisse contenir des individus ne faisant pas partie de la population d'inférence. Pour pallier le problème, les poids de ces individus seront diminués afin de tenir compte de l'incertitude portant sur eux. Plus précisément, on cherche pour chaque sous-population $\omega \subset \Omega$ (par exemple, les hommes entre 20 et 24 ans en t) , un scalaire δ^w tel que

$$\left(\sum_{i\in R_t\cap w} d_i\right) + \left(\sum_{i\in NR_t\cap w} d_i\right) + \left(\sum_{i\in I_t\cap w} \delta^w \cdot d_i\right) = N_t^w - \Delta_t^w,$$

où N_t^w est le nombre d'individus de w à la date t, et Δ_t^w est le nombre d'individus de w qui sont morts ou qui ont quitté le champ de l'enquête emploi entre t et $t+1$. Ces nombres sont fournis par le département de la démographie de l'Insee.

Au final les marges en t sur notre population d'inférence seront calculées de la façon suivante

$$X_t = \left(\sum_{i\in R_t} d_i \cdot X_i\right) + \left(\sum_{i\in NR_t} d_i \cdot X_i\right) + \sum_{w\subset\Omega}\left(\sum_{i\in I_t\cap w} \delta^w d_i \cdot X_i\right).$$

4 Correction de la non-réponse et calage

4.1 Notations

A partir de bases de sondage de logements ordinaires français (RP et fichier TH), est tiré un échantillon S de logements que l'on pondère de la façon suivante :

$$\forall l \in S, \qquad p_l = \mathbb{P}(l \in S)^{-1}$$

où p_l désigne bien entendu le poids associé au logement l de l'échantillon S (poids défini par l'inverse de la probabilité de sélection du logement dans l'échantillon). Cette méthode de pondération est inspirée par l'utilisation des estimateurs de Horwitz-Thompson.

Une fois cet échantillon de logement constitué, les enquêteurs prennent contact avec les résidences tirées. Cependant, certains logements n'offrent pas de réponse, ou des réponses négatives aux enquêteurs. Divers motifs expliquent cette situation. En effet, le logement échantillonné peut être vacant ou être injoignable car situé dans un immeuble dont l'entrée est protégée par un digicode. Il peut être momentanément inoccupé pour cause de départ en vacances. Ou encore, il peut s'agir d'un petit logement occupé par un seul individu qui est très peu présent en journée. De fait, la non réponse au niveau logement n'est pas uniforme. En notant R_{log} l'ensemble des logements de S où les enquêteurs ont pu interroger au moins une partie des ses habitants en première interrogation, on posera

$$\mathbb{P}(l \in R_{log} | l \in S) = \Phi_l^S.$$

Par ailleurs, dans chaque logement l qui répond, seuls certains individus du champ de l'enquête Emploi (voire même aucun) sont des répondants longitudinaux. Certains habitants peuvent tout simplement refuser de répondre à l'enquête. En outre, quelques habitants du logement peuvent aussi avoir quitté le logement entre les deux dates d'enquête. D'autres peuvent s'être lassés de répondre, ou encore être tout simplement injoignables lors des périodes de collecte (vacances par exemple). Ainsi, en notant R_{indiv} l'ensemble des répondants longitudinaux à notre enquête, on posera

$$\mathbb{P}((l,k) \in R_{indiv} | l \in R_{log}) = \Phi_{(l,k)}^{R_{log}}.$$

Cette fois, chaque individu de R_{indiv} est identifié, pour plus de lisibilité, à l'aide d'un couple (l,k).

Finalement, l'échantillon panel dont on dispose est constitué des individus habitant dans une résidence principale qui a pu être approchée par un enquêteur, et qui ont répondu longitudinalement à l'enquête. Par conséquent, en utilisant les notations proposées plus haut, on obtient

$$\begin{aligned} \mathbb{P}((l,k) \in R_{indiv}) &= \mathbb{P}(l \in S) \cdot \mathbb{P}(l \in R_{log} | l \in S) \cdot \mathbb{P}((l,k) \in R_{indiv} | l \in R_{log}) \\ &= p_l^{-1} \cdot \Phi_l^S \cdot \Phi_{(l,k)}^{R_{log}}. \end{aligned}$$

Cette quantité pourra être modélisée de façon classique sous la forme

$$p_l^{-1} \cdot \Phi(T_l' \cdot \lambda) \cdot \Phi(Z_{(l,k)}' \cdot \mu).$$

où Φ désigne une fonction de lien du type *logit* ou *probit*. T_l et $Z_{(l,k)}$ désignent quant à eux des vecteurs de covariables relatifs respectivement au logement l et à l'individu (l,k).

4.2 Calage ou estimateur des moments ?

La théorie des sondages nous dit que, pour toutes séries (X_l^1) et $(X_{(l,k)}^2)$ dont les totaux sur la population entière sont X^1 et X^2, les estimateurs de Horwitz-Thompson définis par

$$\sum_{l\in R_{log}} \frac{p_l}{\Phi_l^1} X_l^1 \qquad \text{et} \qquad \sum_{(l,k)\in R_{indiv}} \frac{p_l}{\Phi_l^1 \cdot \Phi_{(l,k)}} X_{(l,k)}^2$$

estiment sans biais, respectivement X_1 et X_2.

Par conséquent, si le système

$$\left\{ \begin{array}{rcl} \displaystyle\sum_{l\in R_{log}} \frac{p_l}{\Phi(T_l' \cdot \lambda)} X_l^1 & = & X^1 \\ \displaystyle\sum_{(l,k)\in R_{indiv}} \frac{p_l}{\Phi(T_l' \cdot \lambda) \cdot \Phi(Z_{(l,k)}' \cdot \mu)} X_{(l,k)}^2 & = & X^2 \end{array} \right.$$

admet au moins une solution $(\hat{\lambda}, \hat{\mu})$, le couple ainsi trouvé pourra être vu comme un estimateur des moments. Il permet en outre de réaliser un calage.

Si jamais ce système n'admet pas de solution, on peut toujours trouver un couple $(\tilde{\lambda}, \tilde{\mu})$ qui minimise la quantité

$$\left(\sum_{l\in R_{log}} \frac{p_l}{\Phi(T_l' \cdot \lambda)} X_l^1 - X^1 \right)^2 + \left(\sum_{(l,k)\in R_{indiv}} \frac{p_l}{\Phi(T_l' \cdot \lambda) \cdot \Phi(Z_{(l,k)}' \cdot \mu)} X_{(l,k)}^2 - X^2 \right)^2 .$$

4.3 Mise en œuvre du calage

4.3.1 Méthode de Newton-Raphson

Considérons une application $f : U \subset \mathbb{R}^p \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^1$, de jacobien J_f. Supposons que l'équation $f(x) = 0$ admette au moins une solution x^* sur l'ouvert U. La méthode de Newton-Raphson consiste à construire une suite (x_n) de la façon suivante

1. choisir une valeur initiale x_0 pas trop éloignée de x^*,
2. x_n fixé, construire x_{n+1} tel que

$$J_f(x_n)(x_{n+1} - x_n) = -f(x_n),$$

cette dernière équation étant en fait un système linéaire.

Dans de bonnes conditions (existence d'une solution x^*, valeur de x_0 bien choisie, hypothèses de régularité de f suffisantes), la suite (x_n) ainsi construite converge vers x^*.

4.4 Solutions et pseudo-Solutions de systèmes linéaires

La construction par récurrence de la suite de Newton-Raphson nécessite la résolution de systèmes linéaires. Voici quelques rappels sur ce type de problème.

Considérons un système linéaire à n équations et p inconnues $Ax = b$; avec $A \in \mathcal{M}_{n,p}(\mathbb{R})$, $x \in \mathcal{M}_{p,1}(\mathbb{R})$ et $b \in \mathcal{M}_{n,1}(\mathbb{R})$. Dans la tentative de résolution d'un tel système, deux cas se distinguent généralement.

1. si le système est sur-déterminé ($n > p$), le système a peu de chances d'avoir une solution. Une pseudo-solution au système peut tout de même être fournie par la technique des moindres carrés

$$x^* = Argmin_x \|Ax - b\|^2.$$

Si ${}^tAA \in GL_p(\mathbb{R})$, on obtient de façon explicite $x^* = ({}^tAA)^{-1} \cdot {}^tAb$.

2. Si le système n'est pas sur-déterminé ($n \leq p$), le système $Ax = b$ a de bonnes chances d'avoir au moins une solution. Si $A{}^tA \in GL_n(\mathbb{R})$, une solution du système est donnée par $x^* = {}^tA(A{}^tA)^{-1} \cdot b$. Notons que si $n < p$, le système a en général une infinité de solutions.

3. de manière générale, et quand les matrices considérées sont inversibles on notera les pseudo-inverses de A de la façon suivante

$$\begin{array}{ll} A^+ = ({}^tAA)^{-1} \cdot {}^tA & \text{si } n > p, \\ A^+ = {}^tA \cdot (A{}^tA)^{-1} & \text{si } n \leq p. \end{array}$$

Ces pseudo-inverses permettent de fournir une pseudo-solution, ou une solution si elle existe au système $Ax = b$. C'est

$$x^* = A^+ b.$$

4.4.1 Algorithme utilisé

On considère la fonction f définie par

$$f : (\lambda, \mu) \longrightarrow \begin{pmatrix} \displaystyle\sum_{l \in R_{log}} \frac{p_l}{\Phi(T'_l \cdot \lambda)} X^1_l - X^1 \\ \displaystyle\sum_{(l,k) \in R_{indiv}} \frac{p_l}{\Phi(T'_l \cdot \lambda) \cdot \Phi(Z'_{(l,k)} \cdot \mu)} X^2_{(l,k)} - X^2 \end{pmatrix}.$$

On fixe $\varepsilon > 0$ et $(\lambda_0, \mu_0) = (0, 0)$.
Tant que $\|f(\lambda_n, \mu_n) - (X^1, X^2)\| > \varepsilon$, faire

$$(\lambda_{n+1}, \mu_{n+1}) = (\lambda_n, \mu_n) - J_f(\lambda_n, \mu_n)^+ \cdot f(\lambda_n, \mu_n).$$

Si la suite ainsi définie ne semble pas converger, deux possibilités restent possibles

1. essayer une autre valeur pour (λ_0, μ_0),
2. en notant $g(\lambda, \mu) = \|f(\lambda, \mu)\|_2^2$, chercher un minimum de g en résolvant $g'(\lambda, \mu) = 0$ avec la méthode de Newton-Raphson décrite ci-dessus.

5 Application de la méthode à l'enquête Emploi

Les pondérations calculées dans cette partie correspondent à la population d'inférence des présents au premier trimestre 2009 (T1 2009) et présents au premier trimestre 2010 (T1 2010).

5.1 Liste des variables au niveau du logement

Au niveau logement, le choix s'est porté sur des variables permettant à la fois de corriger la non réponse et de caler certaines estimations. Les variables sélectionnées dans ce document sont les suivantes :

- la tranche d'unité urbaine (en 5 modalités) : des études montrent que la réponse au niveau logement est plus forte en secteur rural ;
- le statut d'occupation du logement : les locataires de meublés sont plus difficilement joignables et sont en outre plus enclins à déménager ;
- le type de ménage qui habite le logement (en 5 modalités) : on sait que les jeunes célibataires sont plus difficiles à joindre que les autres ;
- l'habitat en zone urbaine sensible : il paraît plausible que l'interrogation soit plus difficile dans ces quartiers ;
- le caractère « maison individuelle » ;
- l'année d'achèvement de construction de l'immeuble (en 5 modalités) ;

Toutes les marges de ces variables ont été calculées à partir de l'enquête Emploi en utilisant les 6 vagues disponibles au T1 2009 et des pondération `extri11` calculées en coupe. Il s'agit donc d'un « auto-calage ».

5.2 Liste des variables au niveau individuel

Au niveau individuel, une stratégie différente a été adoptée. Aux 28 variables individuelles qui ont servi au calage, il a été rajouté un certain nombre de variables permettant d'expliquer à la fois la non réponse et l'attrition individuelle. Pour ces variables, on ne pouvait pas ou on ne souhaitait faire de calage. En particulier, deux variables de transition sur le marché du travail sont dans le modèle de non réponse. Or, elles sont aussi des variables d'intérêt de cette étude : elles sont par conséquent non admissibles à un calage !

Les variables qui ont servi au calage sont les suivantes :

- le sexe ;
- le statut d'activité en 2009 (Actif Occupé, Chômeur, Inactif) ;
- le statut d'activité en 2010 ;
- le niveau de diplôme en (4 modalités) ;
- la région d'habitation (découpage du territoire en huit grandes zones d'études et d'aménagement) ;
- la semaine de référence (13 semaines de référence au T1 2009) : en effet, la notion de chômeur BIT est une notion hebdomadaire, et le taux de chômage BIT calculé trimestriellement est le la moyenne des 13 taux de chômage hebdomadaires.

Les marges de ces variables ont été calculées à partir des 6 vagues de l'enquête Emploi au T1 2009 (à part bien sûr les marges d'activité au T1 2010). Notons que les marges ont été ajustées des variations démographiques (voir partie 3.3.2 Calcul des marges).

En plus de ces variables, pertinentes dans l'étude de la non réponse, nous avons rajouté les variables suivantes :

- l'âge quinquennal ;
- le groupe social des actifs occupés (on a regroupé les agriculteurs, les artisans, commerçants et chefs d'entreprise) ;
- le chômage de longue durée (au chômage depuis plus d'un an au T1 2009) ;
- le souhait d'avoir un autre emploi en plus ou en remplacement de l'emploi actuel pour les actifs occupés ;
- la transition de l'emploi au T1 2009 vers le non-emploi au T1 2010 ;

– la transition du non emploi au T1 2009 vers l'emploi au T1 2010.

Au total ce sont donc 47 variables qui ont servi au calage et 65 qui ont servi à l'explication de la non réponse. On est donc dans un modèle « sur-dimensionné » au niveau des paramètres à estimer.

5.3 Estimation des paramètres

En reprenant les notations introduites plus tôt dans ce document, la méthode de Newton a été initialisée à partir de $(\lambda_0, \mu_0) = 0_{\mathbb{R}^{64}}$ et il a fallu 11 itérations avant que l'algorithme ne s'arrête (l'algorithme s'arrêtait quand toutes les marges estimées étaient à un distance plus petite que 1 des « vraies » marges). La fonction de lien qui a été utilisée est la fonction *logit*. On peut voir Table 1 l'estimation finale du vecteur explicatif de la non réponse obtenu grâce à notre méthode.

Dans l'ensemble, les coefficients obtenus sont conformes à nos attentes. Soulignons que les coefficients relatifs aux transitions de l'emploi vers le non emploi et du non emploi vers l'emploi sont relativement importants et sont de signes contraires. Ceci s'explique convenablement dans le cadre de l'enquête Emploi. En effet, sur le marché du travail, certains individus, au chômage au T1 2009 et ayant retrouvé un emploi avant le T1 2010 ont du déménager. Alors que d'autres personnes qui se retrouvent subitement au chômage déménagent moins, par crainte de ne pas réussir à se re-loger.

En outre, la probabilité de répondre pour un logement de référence est donnée par le « *logit* de la constante au niveau Logement » : on obtient 0.53. Ce chiffre, *a priori* faible et qui semble contradictoire avec les taux de réponse de l'EEC, ne prend pas en compte le champ des logements de l'EEC. En effet, il s'agit de 53 % de réponse en moyenne par rapport à l'ensemble des logements de l'échantillon, et non pas par rapport aux logements ordinaires considérés comme résidences principales. Par ailleurs, quand on fait la moyenne (non pondérée) des probabilités de réponse au niveau individuel sur notre échantillon, on trouve : 0.90. Dans ce type de protocole où la non-réponse peut se décomposer en deux parties, il semble au final que la difficulté principale est de joindre le logement lors de l'approche initiale.

5.4 Nouvelle distribution des poids

Dans un calage classique, l'idée sous-jacente est que la distribution des poids calés doit être la plus proche possible de la distribution initiale des poids. La distance considérée dans ce paradigme, est la distance moyenne du rapport des poids à 1. Rappelons que le choix de la distance est équivalent au choix de la fonction de lien, et que les utilisateurs utilisent généralement une distance contrôlant les rapports de poids (par exemple en imposant une borne supérieure à ce rapport). La théorie dit qu'en l'absence de non réponse, le choix de la fonction de lien (ou de la distance) importe peu, car sous certaines conditions de régularité, ces méthodes sont asymptotiquement équivalentes. Ici, nous sommes dans un cadre légèrement différent. L'objectif est d'abord de corriger la non réponse afin d'obtenir un jeu de pondérations se rapprochant le plus possible du cadre de Horwitz-Thompson que l'on sait non biaisé. Aucun contrôle n'a été imposé pour limiter supérieurement les poids.

Coefficients Logement		Coefficients Individu	
Constante	0.13	Constante	6.96
tu1	0.27	ag15	0.72
tu2	0.27	ag20	0.05
tu3	0.18	ag25	-0.5
tu4	0.11	ag30	-0.12
tu5	*ref*	ag35	-0.28
Typmen1	-0.2	ag40	0.49
Typmen2	0.03	ag45	-0.04
Typmen3	*ref*	ag50	0.25
Typmen4	0.02	ag55	0.63
Typmen5	-0.01	ag60	0.81
Zus	-0.25	ag65	*ref*
PropAcc	*ref*	Femme	0.71
Prop	0.13	AO-09	0.51
Loc	-0.04	CHO-09	-1.47
Meubl	-0.55	INAC-09	*ref*
Grat	0.1	AO-10	-0.48
Mi	-0.03	CHO-10	-1.47
aai4	*ref*	INAC-10	*ref*
aai5	0.03	Dip0	0.32
aai6	0.17	Dip1	*ref*
aai7	0.29	Dip2	0.44
aai8	-1.03	Dip3	0.05
		CS12	0.75
		CS3	0.11
		CS4	*ref*
		CS5	-1.9
		CS6	1
		CHOanc	-0.17
		AOsou	-0.14
		TRANSe-ne	0.59
		TRANSne-e	-0.4

Table 1 - Vecteur explicatif de la non réponse

On peut voir Figure 2, la distribution initiale des poids qui provient de la construction de l'échantillon (en traits pointillés) et la distribution finale des poids qui a été obtenue avec la méthode proposée dans ce document (en trait plein). On peut voir aussi sur ce graphique la « distribution intermédiaire » des poids obtenue avec seulement la correction de la non-réponse au niveau ménage. On s'aperçoit une nouvelle fois en observant ce graphique que la correction de la non réponse au niveau logement est plus importante que la correction individuelle, due essentiellement à l'attrition.

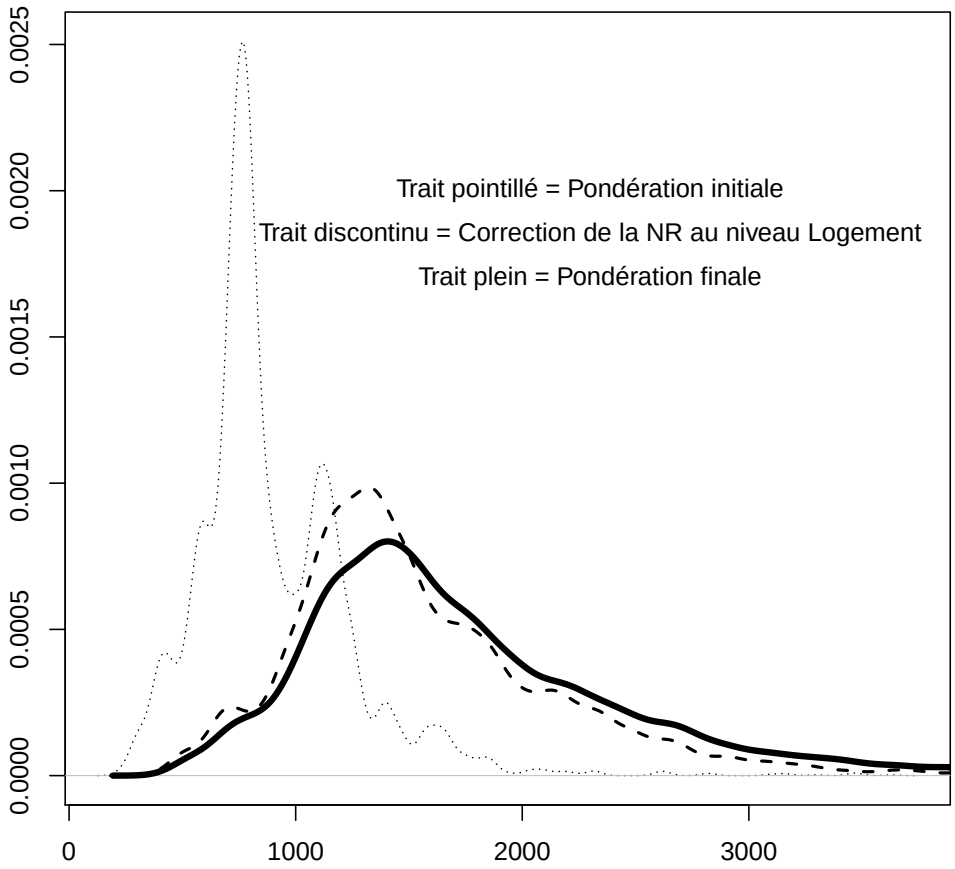

Figure 2 - Distribution des poids, avant et après calage

En regardant de plus près la distribution du rapport des poids dans la Table 2, on observe que le rapport moyen est un peu plus grand que deux, que le rapport minimum vaut 1.4 (avec la méthode que nous proposons, les poids calés sont tous supérieurs aux poids d'échantillonnage) et que le rapport le plus grand vaut 18.5. Cet individu, vraisemblablement atypique, représente un peu plus de 20 individus « moyens ». Notons que seules 19 observations (sur un total de 26 574) présentent un rapport des poids plus grand que 10.

Statistiques		Quantiles	
Moyenne	2.06	$q_{5\%}$	1.5
Médianne	1.8	$q_{10\%}$	1.6
Ecart-Type	0.75	$q_{25\%}$	1.7
Minimum	1.4	$q_{75\%}$	2.1
Maximum	18.5	$q_{90\%}$	2.9
		$q_{95\%}$	3.4
		$q_{99\%}$	5.1

Table 2 - Rapport des poids : Poids calés / Poids de sondage

Lecture : Le rapport moyen des poids est de 2.06 ; 95 % des rapports de poids sont plus petits que 3.4

6 Applications aux transitions sur le marché du travail

6.1 Première évaluation de la pondération

Afin de savoir si le jeu de pondération que nous avons obtenu permet d'être cohérent avec d'autres sources, on compare nos résultats sur le statut matrimonial légal (variable `MATRI`) avec ceux de l'Enquête Annuelle de Recensement. Notons tout de même qu'il existe une différence de champ entre les deux tables de données : dans l'EAR 2010, il n'était pas possible de sélectionner *stricto sensu* les personnes qui étaient dans le champ de l'enquête au 1er janvier 2009.

	Célibataire		Marié(e)		Veuf(ve)		Divorcé(e)	
Région	EAR	EEC	EAR	EEC	EAR	EEC	EAR	EEC
Région Parisienne	38,7	40,1	51,0	49,6	1,6	1,6	8,7	7,8
Bassin Parisien	32,5	28,7	56,0	58,6	2,3	3,0	9,2	9,7
Nord	31,0	31,4	57,2	56,8	2,8	2,3	9,0	9,4
Est	31,0	31,3	57,4	56,6	2,1	1,6	9,5	10,4
Ouest	32,2	23,0	57,2	68,7	2,1	1,7	8,4	6,6
Sud-Ouest	35,2	41,4	52,6	47,3	1,9	1,1	10,2	10,1
Centre-Est	33,1	30,5	55,5	58,1	1,8	2,0	9,6	9,4
Méditerranée	33,9	33,8	52,7	52,6	1,9	2,0	11,5	11,6
France Métropolitaine	34,0	33,1	54,6	55,8	2,0	1,9	9,4	9,2

Table 3 - Statut matrimonial légal au T1 2010 - individus âgés entre 25 et 64 ans au 31 décembre 2010

Les résultats que nous obtenons à ce stade (Table 3) sont assez cohérents avec ceux du recensement, tout du moins d'un point de vue national. Pour les zones Ouest et Sud-Ouest, les divergences sont assez fortes et non expliquées pour le moment... Une utilisation de cette méthode de pondération semble donc légitime sur la population métropolitaine.

6.2 Matrices de transition

Comme nous l'avons évoqué au début de ce document, l'un des intérêts majeurs des pondérations longitudinales est de construire des matrices de transition sur le marché du travail pour la population des Présent-Présent. La Table 4 montre la matrice de transition, avant calage, entre les 3 grands états. Les individus pris en compte dans ce tableau sont ceux de la population d'inférence des présents au T1 2009, présents au T1 2010 et âgés entre 20 et 64 ans au 31 décembre 2009. Cette matrice est à comparer à celle présentée dans la Table 5.

On voit par exemple dans la Table 5, que 6.7 % des individus inactifs au T1 2009 sont devenus chômeurs au T1 2010. Et ce taux varie en fonction du sexe : il s'élève à 8.8 % chez les hommes quand il est seulement de 5.8 % chez les femmes.

Quand on compare ces deux matrices de transition, on voit que les changements les plus importants concernent les deuxième et troisième lignes. La « diagonale Chômeur » est plus importante avec les pondérations longitudinales, alors que la « diagonale Inactif » l'est moins. Cela est très certainement lié au fait que les chômeurs répondent longitudinalement moins souvent (d'autant moins s'ils sont en plus chômeurs de longue durée) que les inactifs. Ce fait est confirmé par la Table 1.

	Actif Occupé 2010		Chômeur 2010		Inactif 2010	
Actif Occupé 2009	**92.9**		**3.0**		**4.1**	
	93.3	92.4	3.1	2.8	3.6	4.8
Chômeur 2009	**37.8**		**45.4**		**16.8**	
	34.5	41.4	49.7	40.6	15.8	18.0
Inactif 2009	**7.9**		**4.9**		**87.1**	
	7.5	8.2	5.7	4.4	86.8	87.4

Table 4 - Transitions sur le marché du travail des 20-64 ans - Pondération initiale

Lecture : 92.9 % des individus qui étaient actifs occupés au T1 2009 sont toujours actifs occupés au T1 2010. 4.4 % des femmes qui étaient inactives au T1 2009 sont au chômage au T1 2010.

	Actif Occupé 2010		Chômeur 2010		Inactif 2010	
Actif Occupé 2009	**93.0**		**3.3**		**3.7**	
	93.4	92.7	3.6	3.1	3.0	4.2
Chômeur 2009	**36.0**		**50.5**		**13.5**	
	33.1	40.1	55.1	43.9	11.8	18.0
Inactif 2009	**9.1**		**6.7**		**84.2**	
	9.0	9.3	8.8	5.3	82.2	85.4

Table 5 - Transitions sur le marché du travail des 20-64 ans - Pondération après calage

Lecture : 36 % des individus chômeurs au T1 2009 sont actifs occupés au T1 2010.
55.1 % des hommes qui étaient au chômage au T1 2009 sont toujours au chômage au T1 2010.

Références

[1] de Foucauld J-B et al., *Emploi, chômage, précarité. Mieux mesurer pour mieux débattre et mieux agir*. Rapport du Cnis, 2008.

[2] Christine M., *La construction de l'échantillon de la future enquête Emploi en continu à partir du recensement de 1999*. Acte JMS, 2002.

[3] Deville J-C., *La correction de la non-réponse par calage généralisé*. Acte JMS, 2002.

[4] Freyssinet J. et al. *Niveaux de vie et inégalités sociales*. Rapport du Cnis, 2006.

[5] Loonis V. *La construction du nouvel échantillon de l'enquête Emploi en continu à partir des fichiers de la taxe d'habitation*. Acte JMS, 2009.

[6] Naud J-F. *Pondération longitudinale avec panels combinés, Enquête sur la dynamique du travail et du revenu*. Document de recherche StatCan, 2004.

[7] Nouël de Buzonniere C., Jauneau Y., *Transitions annuelles au sens du BIT sur le marché du travail*. Document de travail DSDS N°F1107, 2011.