Calcul de précision transversale dans l'enquête emploi en France[1]

Pascal ARDILLY(*), Guillaume OSIER(**)

(*) Insee, (**) Insee et Université de Lyon I

1. Généralités sur l'enquête emploi

1.1 Champ et objectifs de l'enquête

L'enquête Emploi (dite « en continu ») est réalisée chaque trimestre par l'INSEE sur un échantillon d'environ 54 000 logements ordinaires (on désigne ainsi tous les logements à usage d'habitation, hormis les communautés). Elle doit répondre à deux objectifs majeurs : étudier le marché du travail en France un trimestre donné (et notamment évaluer le nombre total de chômeurs ainsi que le taux de chômage), et mesurer les variations trimestrielles de l'emploi. Toutes les personnes de 15 ans ou plus, vivant ne serait-ce qu'une partie du temps[2] dans de tels logements, sont interrogées, dès lors qu'il s'agit d'une résidence principale au moment de l'enquête. L'enquête Emploi actuelle a progressivement remplacé, depuis l'été 2001, l'enquête Emploi annuelle, qui était fondée sur une interrogation d'environ 80 000 logements en mars de chaque année.

1.2 Une enquête à caractère « continu »

L'échantillonnage est conçu à un rythme trimestriel. Dans la pratique, l'interrogation d'une personne se fera relativement à une semaine donnée, dite semaine de référence. La situation de cette personne au cours de cette semaine permettra de la ranger dans l'une des trois catégories suivantes : actif occupé, chômeur, ou inactif. Or la collecte est étalée régulièrement tout au long du trimestre afin que toutes les semaines de référence soient représentées de manière identique. Cette approche « en continu » constitue une innovation essentielle de la nouvelle méthode par rapport aux précédentes enquêtes annuelles.

1.3 Un échantillonnage aréolaire et rotatif

Le tirage de l'échantillon Emploi s'appuie sur les données du recensement de la population de mars 1999 (RP99). On échantillonne des aires géographiques parfaitement définies et au sein desquelles on pratique un tirage exhaustif des logements recensés. Il s'agit donc d'un tirage en

[1] Pascal Ardilly et Guillaume Osier, INSEE, 18 Boulevard Adolphe Pinard, 75675 Paris Cedex 14, France.

[2] Les personnes résidant dans une communauté, mais ayant un lien avec un ménage ordinaire (enfant étudiant vivant en foyer universitaire, parent âgé vivant en maison de retraite…) sont également interrogées.

grappes, dit « aréolaire ». Néanmoins, dans certains cas, la prise en compte des logements neufs – c'est à dire des logements dont la construction a été achevée après le RP99 – peut conduire à un échantillonnage direct de ces logements à partir d'informations recueillies sur le terrain par les enquêteurs.

Les intérêts d'un tirage aréolaire sont bien connus : les coûts de déplacement des enquêteurs sont réduits, le taux de réponse est plus fort (grâce à un effet d'entraînement « positif » entre les ménages d'une même zone), le suivi de l'habitat est amélioré (cela permet d'interroger des logements qui ont été oubliés lors du recensement) et facilité (cela permet de prendre en compte les logements neufs sans faire appel à une base de sondage spécifique). En revanche, le tirage aréolaire diminue la précision des estimateurs, à cause de l'« effet de grappe ». Il est souhaitable, pour limiter l'effet de grappe, de disposer d'aires de taille réduite et autant que possible hétérogènes : dans cette enquête, les aires contiennent environ 20 logements.

L'un des objectifs de l'enquête étant de fournir des résultats conjoncturels sur le marché du travail, l'accent a été mis sur la fourniture d'estimations d'évolutions trimestrielles. Afin de réduire la variance de ces estimations, on a choisi de conserver une fraction importante de l'échantillon d'un trimestre sur l'autre, en mettant en place un système d'échantillon rotatif[3] dans lequel chaque aire sera interrogée six trimestres de suite. Le taux de renouvellement a donc été fixé à 1/6 par trimestre.

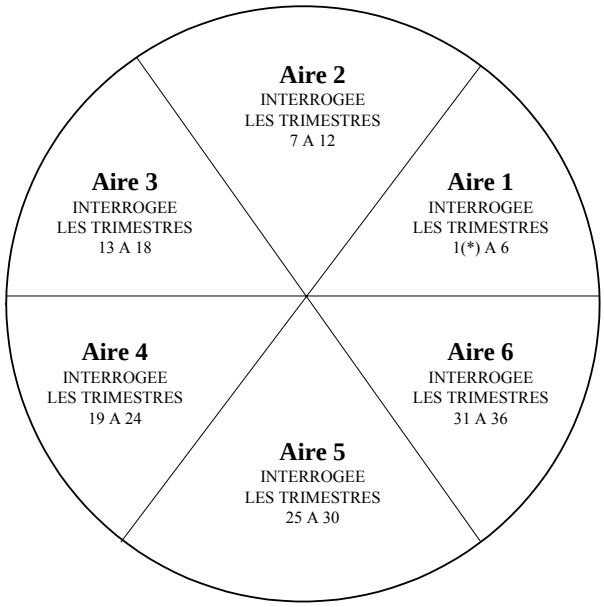
2. Description du plan de sondage

2.1 Un échantillonnage à plusieurs degrés

L'échantillonnage Emploi résulte d'un tirage à plusieurs degrés : ayant constitué un échantillon d'unités primaires (UP), on découpe et on tire des aires au sein de chaque UP. Cependant, certaines UP ont des tailles[4] très importantes, ce qui pourrait entraîner des procédures de découpage longues et coûteuses. C'est pourquoi on peut passer par des unités de tailles intermédiaires, appelées « grandes zones » et « secteurs ». Ainsi, une unité primaire de grande taille est découpée en secteurs. On sélectionne ensuite un secteur dans lequel on découpe puis on tire des aires. Toutefois, si la taille de l'unité primaire est vraiment très grande (plus de 1 000 logements), on commencera d'abord par sélectionner une grande zone au sein de laquelle on tirera un secteur puis des aires.

L'échantillon d'aires doit être géré dans le temps, conformément aux objectifs de rotation. Dans un secteur donné, un trimestre donné, une et une seule aire sera mobilisée, mais un même secteur devra néanmoins donner naissance à plusieurs aires. En effet, compte tenu du caractère rotatif de l'échantillon, chaque aire est interrogée pendant six trimestres consécutifs, à la suite de quoi elle doit être remplacée par une aire ayant des caractéristiques similaires. Chaque secteur devra donc être découpé de manière à donner naissance, in fine, à plusieurs aires de vingt logements en moyenne. Parmi ces aires, six[5] seront retenues pour faire partie de l'échantillon final. Quant aux secteurs eux-mêmes, ils ont été ventilés en six sous-groupes de même taille, lesquels sous-groupes sont sollicités de manière décalée dans le temps, les uns après les autres. Ce processus assez complexe permet d'engager puis d'entretenir assez simplement un échantillon rotatif. Les aires tirées dans un secteur sont finalement numérotées de 1 à 6. Ce numéro va déterminer l'ordre de passage de l'aire dans l'échantillon, comme l'explique le schéma suivant, qui s'applique secteur par secteur :

[3] Néanmoins, la panélisation porte « seulement » sur les logements, et non sur les individus physiques que l'on ne suit pas en cas de déménagement.
[4] Dans tout le texte, on entendra par taille d'une unité le nombre total de logements recensés qu'elle contient.
[5] Ce nombre assure une durée de vie de $6 \times 6 = 36$ trimestres = 9 années à l'échantillon Emploi.



(*) le trimestre 1 correspond au trimestre d'introduction du secteur dans l'échantillon Emploi : c'est la diversité des trimestres d'introduction qui alimente le caractère rotatif de l'échantillon ;

Finalement, le découpage en secteurs est intéressant au moins à deux titres : il permet de diminuer la charge du découpage, en ne découpant en aires que des unités de taille réduite (et malheureusement relativement homogènes), et il minimise les fluctuations aléatoires lors du changement d'aires, puisqu'une aire sortante est remplacée par une aire voisine. Une conséquence directe de cette méthode d'échantillonnage est que les unités primaires devront comporter au moins $6 \times 20 = 120$ logements, afin que l'on puisse y découper 6 aires d'une vingtaine de logements. En revanche, on ne fixe pas de limite supérieure de taille a priori : celle-ci peut atteindre, dans certains cas, plusieurs milliers de logements.

2.1.1 Constitution et tirage des unités primaires

Constitution des UP :

Pour construire la base de sondage des unités primaires, on va partir de communes ou de groupes de communes de plus de 120 logements, en les morcelant le plus finement possible en unités standard sous contrainte qu'elles conservent plus de 120 logements. Les unités ainsi construites vont former les unités primaires. Dans ce processus, on distingue les communes, les districts (qui sont des zones géographiques infra communales « de base » utilisées lors de la collecte du recensement) et les IRIS2000, échelon géographique infra-communal également, composé d'un ensemble de districts et atteignant 2000 habitants (en général).
Pour découper les UP, on part donc de la commune, unité administrative très clairement délimitée sur le terrain. On procédera au préalable, si besoin est, à des groupements de communes pour disposer d'unités ayant au moins 120 logements. Deux cas se présentent alors :

- La commune n'a pas été découpée en IRIS lors du recensement : l'UP est alors la commune (ou le groupe de communes).
- La commune a été découpée en IRIS lors du recensement : on considère alors chacun des IRIS successivement. Deux situations sont alors possibles :
  - L'IRIS ne contient que des districts de plus de 120 logements. Chacun de ces districts constitue alors une UP.
  - L'IRIS contient au moins un district de moins de 120 logements, et c'est alors l'IRIS tout entier qui constitue l'UP.

Tirage de l'échantillon d'UP

Le tirage des UP est stratifié par région et tranche d'unité urbaine (code TU99)[6]. Au sein de chaque strate, les unités primaires seront tirées avec une probabilité proportionnelle à leur nombre de logements recensés. Cette technique permettra de diminuer la variance d'échantillonnage, car les totaux de n'importe quelle variable définis au niveau de l'unité primaire sont à peu près proportionnels au nombre de logements de cette unité. Au total, 2554 UP ont été tirées dans 104 strates.

2.1.2 Découpage et tirage des aires au sein des UP

Au sein de chacune des UP échantillonnées, il s'agit de découper des aires, puis d'en tirer six. Ce découpage se fera en plusieurs étapes :

- Chaque UP sera découpée en secteurs de 120 à 240 logements, ou restera à l'identique si elle a une taille comprise entre 120 et 240 logements. Si besoin est, on découpera préalablement l'unité en grandes zones, et on en tirera une avec une probabilité proportionnelle à sa taille. On découpera ensuite cette grande zone en secteurs.
- Un (seul) des secteurs ainsi découpés sera tiré avec une probabilité proportionnelle à sa taille.
- Le secteur sélectionné sera ensuite découpé en au moins 6 aires de 20 logements en moyenne (la fourchette autorisée en pratique pour la taille des aires sera comprise entre 17 et 23 logements). Les aires découpées devront obligatoirement être d'un seul tenant et devront avoir des limites facilement repérables sur le terrain par un enquêteur (rue, cours d'eau, voie ferrée, étage d'immeuble…). Puis on tire six aires par sondage aléatoire simple, parmi l'ensemble des aires qui ont été découpées dans le secteur (en nombre compris entre 6 et 12). On attribuera in fine aux aires sélectionnées, de manière équiprobable, un numéro de série allant de 1 à 6 qui va définir la façon dont elles se succéderont dans le temps.

Finalement, un trimestre donné, on peut calculer la probabilité d'inclusion d'une aire constituée à l'intérieur du secteur **j** de l'UP **i** appartenant à la strate **h** :

$$\Pi_{hij} = k_h \times \frac{T_i}{\sum_{l \in P_h} T_l} \times \frac{S_{ji}}{T_i} \times \frac{6}{n_{ji}} \times \frac{1}{6} \qquad (1)$$

k_h est le nombre d'unités primaires tirées dans la strate **h**.

T_i est la taille (en nombre de logements recensés) de l'unité primaire **i**.

P_h représente la population des unités primaires de la strate **h**.

S_{ji} est le nombre de logements du secteur **j** tiré dans l'unité **i**.

n_{ji} est le nombre d'aires découpées à l'intérieur du secteur **j** de l'unité **i**.

La moyenne des poids issus de (1) vaut 560 environ, et on constate des écarts entre les moyennes par région (de 282 à 819) et entre les moyennes par catégorie de communes (de 800 environ pour les communes rurales à 350 environ pour les grosses unités urbaines). Ces variations s'expliquent essentiellement par les contraintes de précision régionale imposées par Eurostat, qui ont conduit à augmenter la taille de l'échantillon dans les petites régions.

[6] Le code TU99 vaut 0 en zone rurale, et il augmente de 1 à 8 en fonction de la taille de la population se situant dans l'unité urbaine (TU99=8 pour l'unité urbaine de Paris).

2.1.3 Echantillonnage d'individus :

Tous les individus résidant dans n'importe quel logement recensé dans l'aire sont interrogés. Il peut, en revanche, y avoir un échantillonnage de logements neufs (dans lesquels tous les individus seront interrogés). Plus précisément, si **X** désigne le nombre total de logements neufs contenus dans une aire donnée, la règle est la suivante :

- $X \leq 10$: on interroge tous les logements neufs de la zone.
- $11 \leq X \leq 40$: on tire 10 logements neufs par sondage aléatoire simple (SAS).
- $41 \leq X$: on tire des logements neufs par SAS avec un taux de sondage de 1/4.

Cette méthode d'échantillonnage des logements neufs présente le double avantage de limiter la charge de travail des enquêteurs (une zone géographique peut en effet contenir un très grand nombre de logements neufs) et d'attribuer à ces logements des poids de sondage pas trop différents[7] de ceux des logements recensés. Pour obtenir le poids d'échantillonnage (avant calage) d'un logement (et donc d'un individu physique), il faudra donc distinguer deux cas : soit le logement a été recensé (ou oublié au recensement mais enquêté), et il a in fine le poids de l'aire qui le contient, soit le logement est « neuf ». Dans ce dernier cas, il faut multiplier la probabilité d'inclusion de l'aire - donnée par (1) - par le taux de sondage de ces logements neufs pour obtenir leur probabilité d'inclusion.

2.2 Les repondérations

Les repondérations permettent, d'une part de corriger la non-réponse totale, et d'autre part de limiter les erreurs d'échantillonnage (principe du redressement). Un unique calage, utilisant le logiciel Calmar, a permis de satisfaire simultanément ces deux objectifs en attribuant de nouveaux poids aux individus. On a contraint tous les individus d'un même ménage à avoir le même poids final. Les variables de calage utilisées sont de deux types :

- des variables issues du recensement : nombre de pièces du logement , type de logement, tranche d'unité urbaine ;
- des variables actualisées : tranches d'âge quinquennal par sexe (statistiques issues de l'état civil), caractère recensé ou « neuf » d'un logement.

3. Le logiciel Poulpe

3.1. Plans de sondage traités et résultats produits

POULPE[8] est un logiciel de calcul de précision mis au point par l'INSEE et écrit en langage SAS MACRO. Il permet d'estimer la variance d'estimateurs « simples » ou « complexes » dans le cadre d'un échantillonnage complexe, à probabilités inégales et à plusieurs degrés, comportant plusieurs phases de tirage (ce qui permet par exemple de modéliser la non-réponse), tout en prenant en compte les redressements. Les plans de sondage traités en standard par le logiciel sont :

- Les plans à plusieurs degrés et en une phase (probabilités d'inclusion quelconques) ;
- Les plans à plusieurs degrés et en deux phases, dont la deuxième phase est, soit un tirage post-stratifié soit un tirage de Poisson ;
- Les plans à plusieurs degrés et en trois phases, dont la deuxième phase est un tirage post-stratifié et la troisième phase un tirage de Poisson.

[7] Une trop grande dispersion des poids de sondage est généralement source de variance
[8] Programme Optimal et Universel pour la Livraison de la Précision des Enquêtes.

Poulpe traite évidemment le cas classique des estimateurs de type Horvitz-Thompson :

$$\hat{Y}_{\pi}(s) = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \omega_i \times y_i$$

où π_i désigne la probabilité d'inclusion de l'individu **i**, et ω_i le poids de sondage correspondant, mais il peut aussi traiter les estimateurs non linéaires, soit définis en standard (c'est le cas par exemple des ratios), soit construits à partir des variables linéarisées fournies par l'utilisateur. Pour une liste donnée de variables d'intérêt, Poulpe calculera :

- Les estimations des paramètres d'intérêt, à partir de deux familles distinctes de poids de sondage :
  - Les poids de sondage fournis par l'utilisateur : ces poids ont éventuellement été redressés par un logiciel approprié - comme Calmar. La valeur obtenue pour cette estimation est appelée Sompond.
  - Les poids de sondage recalculés par le logiciel en multipliant les probabilités associées à chaque degré : la valeur de l'estimation obtenue à partir de ces poids est appelée Somlog.
- Une estimation de leurs variances : dans le cas d'une enquête non redressée, le logiciel s'appuiera directement sur les variables d'intérêt. Par contre, si les poids ont été redressés par un calage généralisé, Poulpe s'appuiera sur les résidus issus de la régression linéaire multiple des variables d'intérêt sur les variables de calage.
- Une estimation des bornes de l'intervalle de confiance à 95%.
- L'effet de sondage Deff estimé.

L'intervalle de confiance est estimé à partir des écarts-type estimés. On fait l'hypothèse, justifiée avec des échantillons de grande taille (n ≈ 100 suffit, et ici **n** dépasse 50000…), que les estimateurs $\hat{\theta}$ manipulés suivent des lois de Gauss. L'intervalle de confiance estimé à 95% pour un paramètre θ est donc :

$$\mathrm{IC} = \left[\hat{\theta} - 2\sqrt{\hat{\mathrm{V}}\left(\hat{\theta}\right)}, \hat{\theta} + 2\sqrt{\hat{\mathrm{V}}\left(\hat{\theta}\right)} \right]$$

L'effet de sondage Deff pour un total (Design EFFect) égal par définition au rapport de la variance de l'estimateur d'Horvitz-Thompson du total **Y** sous le plan de sondage adopté, sur la variance de l'estimateur de ce même total que l'on obtiendrait avec un plan aléatoire simple sans remise et de taille fixe, égale à celle de l'échantillon effectivement disponible. Si $\hat{\mathrm{Y}}$ désigne l'estimateur d'Horvitz-Thompson du total **Y**, cela conduit à la définition :

$$\mathrm{DEFF} = \frac{\mathrm{V}_{\mathrm{PLAN}}\left(\hat{\mathrm{Y}}\right)}{\mathrm{V}_{\mathrm{SAS}}\left(\mathrm{N} \cdot \bar{\mathrm{y}}\right)}$$

N désigne la taille de la population et $\bar{\mathrm{y}}$ la moyenne simple sur l'échantillon. Dans le cas d'une enquête en une phase, ce paramètre est estimé ainsi :

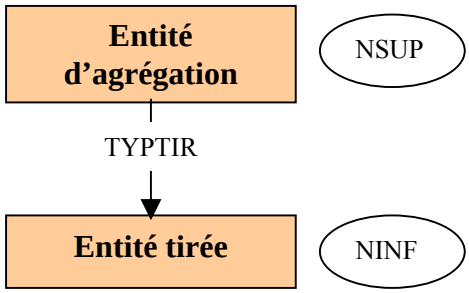
$$\hat{\mathrm{DEFF}} = \frac{\hat{\mathrm{V}}\left(\hat{\mathrm{Y}}\right)}{\frac{1}{\mathrm{n}}\left(1 - \frac{\mathrm{n}-1}{\hat{\mathrm{N}}-1}\right)\hat{\mathrm{N}}\sum_{\mathrm{k} \in \mathrm{s}} \frac{1}{\pi_{\mathrm{k}}}\left(\mathrm{y}_{\mathrm{k}} - \bar{\mathrm{Y}}^{*}\right)^2}$$

avec $\bar{\mathrm{Y}}^{*} = \sum_{\mathrm{k} \in \mathrm{s}} \frac{\mathrm{y}_{\mathrm{k}}}{\pi_{\mathrm{k}}} \Big/ \sum_{\mathrm{k} \in \mathrm{s}} \frac{1}{\pi_{\mathrm{k}}}$ et $\hat{\mathrm{N}} = \sum_{\mathrm{k} \in \mathrm{s}} \frac{1}{\pi_{\mathrm{k}}}$

3.2 Principes de fonctionnement de Poulpe

3.2.1. Modélisation du plan de sondage à plusieurs degrés

Le logiciel va s'appuyer sur un modèle de représentation du plan de sondage à plusieurs degrés. Ce modèle prend la forme d'un «arbre» dont chaque arc représente un sondage « élémentaire », c'est à dire un degré donné de tirage. Ainsi, un plan à plusieurs degrés est un ensemble de sondages élémentaires successifs. Un sondage élémentaire est représenté par un arc du type suivant, joignant deux nœuds (les rectangles grisés) **:**

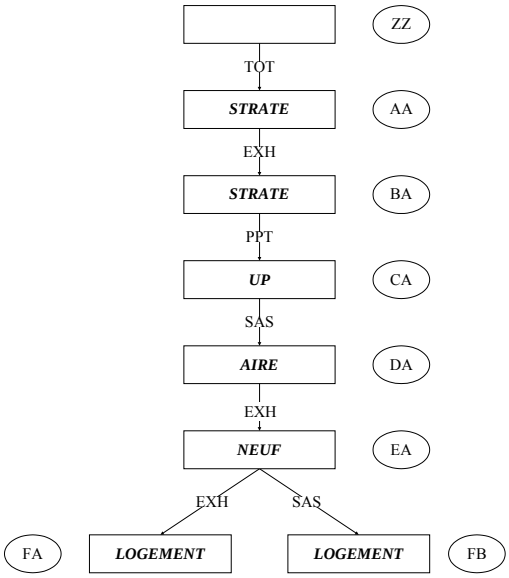


La liaison entre les deux rectangles est caractérisée par un code « TYPTIR » qui précise le type d'échantillonnage utilisé. Les codes actuellement reconnus par Poulpe sont :

- EXH : tirage exhaustif
- SAS : sondage aléatoire simple
- SASEQ : sondage équilibré à probabilités égales
- PPT : tirage à probabilités proportionnelles à la taille
- SYS : tirage systématique

et prochainement EQUIL pour l'échantillonnage équilibré à probabilités inégales. Le code EXH est utilisé pour traduire une stratification. Chaque nœud est identifié par un code (ici NSUP et NINF). Ce code va permettre de faire le lien entre les sondages « élémentaires » successifs[9].

Le plan de sondage adopté pour l'enquête Emploi donne lieu à l'arbre suivant, après la prise en compte des simplifications qui seront exposées au IV :



[9] Tout arbre de sondage devra enfin être surmonté d'un arc spécial imposé qui permettra de récapituler les résultats. Cet arc ne contiendra que les données d'identification des nœuds (NSUP et NINF) et le code TYPTIR qui vaut TOT.

La subdivision finale s'explique ainsi : pour pouvoir distinguer le traitement des logements recensés de celui des logements neufs, on est amené à partager chaque aire en deux strates de logements (repérées par les modalités de la variable « NEUF »). La première strate est composée des logements qui sont recensés (ou oubliés au recensement), tandis que la seconde comprend l'ensemble des logements neufs présents dans l'aire.

b) Utilisation des expressions analytiques de variance

L'arbre précédent étant construit, Poulpe estime les variances de tout plan à plusieurs degrés en s'appuyant sur une formule récursive donnée par Raj[10]**.** On considère un plan de sondage à deux degrés, et on note $\hat{Y}$ l'estimateur d'Horvitz-Thompson du total **Y**. On suppose que les unités secondaires sont tirées indépendamment d'une unité primaire à l'autre. Un estimateur sans biais de la variance de $\hat{Y}$ est :

$$\hat{V}\left(\hat{Y}\right) = f\left(\hat{Y}_i \,\middle|\, i \in s\right) + \sum_{i \in s} \omega_i \hat{V}_i$$

- s : échantillon des unités primaires.
- $f\left(Y_i \,\middle|\, i \in s\right) = \hat{V}_1\left(\sum_{i \in s} \omega_i Y_i\right)$ = variance estimée au premier degré de tirage, dans le cas où les vrais totaux par UP, notés Y_i, sont connus (dans la formule de Raj, on remplacera les vrais totaux Y_i par leurs π-estimateurs $\hat{Y}_i$).
- $\hat{V}_i = \hat{V}_{2|1}\left(\hat{Y}_i\right)$ = estimateur sans biais de la variance au second degré de tirage, dans l'unité primaire **i**.
- $\omega_i = \frac{1}{\pi_i}$ = poids de sondage de l'unité primaire **i**.

De manière générale, Poulpe considère d'abord le dernier degré de tirage correspondant aux feuilles de l'arbre (c'est à dire ici au tirage des logements à l'intérieur d'une aire). Il estime alors la variance générée par ce degré, en utilisant les formules adaptées relatives aux tirages mentionnés sur les arcs partant des feuilles. Cela conduit aux estimateurs $\hat{V}_i$ relatifs aux unités **i** tirées au degré juste supérieur, sachant qu'on dispose par ailleurs des estimations $\hat{Y}_i$ relatives à chaque unité **i**. Le logiciel remonte ensuite au degré encore supérieur : Poulpe évalue la variance associée au tirage des unités **i** en utilisant les formules $f(Y_i | i \in s)$ relatives au type de tirage mentionné sur l'arc. On obtient finalement $f\left(\hat{Y}_i \middle| i \in s\right)$ en remplaçant Y_i (inconnu) par $\hat{Y}_i$ (connu).

La formule de Raj permet donc de décomposer l'estimation de la variance d'un tirage à plusieurs degrés en utilisant les variances estimées au niveau de chacun des degrés : grâce au caractère récursif de la formule, on peut effectuer le calcul global de manière assez simple. Les difficultés causées par les tirages à plusieurs degrés étant ainsi gérées, il reste à traiter les plans complexes à un seul degré. Dans notre cas, l'obstacle se réduit au tirage à probabilités inégales. Pour cela, Poulpe va utiliser une formule proposée par Jean-Claude Deville (Insee), qui ne fait pas intervenir dans son expression les probabilités d'inclusion double π_{ij}, dont le calcul est en général irréalisable en pratique. Pour un plan de sondage **P** de taille fixe **n** et d'entropie très grande, si on note par $\hat{Y}$ l'estimateur d'Horvitz-Thompson du total **Y**, un estimateur de la variance de $\hat{Y}$ est donné par la formule suivante :

$$\hat{V}_P\left(\hat{Y}\right) = \frac{n}{n-1} \sum_{k \in s} \left(1 - \pi_k\right) \left(\frac{y_k}{\pi_k} - \sum_{i \in s} a_i(s) \cdot \frac{y_i}{\pi_i} \right)^2$$

[10] Sampling theory, Mc Graw Hill, 1968.

- π_k est la probabilité d'inclusion de l'individu **k**
- $a_i(s) = \dfrac{1-\pi_i}{\sum_{j\in s}(1-\pi_j)}$

Des simulations semblent accréditer l'idée que les performances de cette formule sont satisfaisantes pour des tailles d'échantillons dépassant la dizaine (sinon la variance sera plutôt sous-estimée).

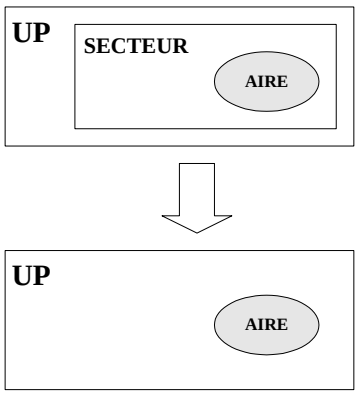
Lorsqu'on traite des estimateurs complexes, si la taille de l'échantillon est suffisamment grande, la variance de l'estimateur « complexe » se confond avec celle de l'estimateur d'Horvitz-Thompson pour la variable linéarisée. On se ramène ainsi à un calcul de variance pour un π-estimateur. Enfin, un résultat bien connu dit qu'en cas de redressement, lorsque la taille **n** est grande, la variance de l'estimateur redressé $\hat{Y}_w$ est approximée par celle de l'estimateur classique de Horvitz-Thompson construit avec une variable d'intérêt égale au résidu de la régression linéaire de **Y** sur **X**.

4. Traitement des principales difficultés méthodologiques rencontrées

Dans la phase de modélisation du plan de sondage de l'enquête Emploi, une difficulté majeure apparaît : un trimestre donné, on tire un seul secteur au sein d'une unité primaire[11], puis une seule aire au sein d'un secteur. Or, il n'est pas possible d'estimer sans biais une variance à partir d'un échantillon de taille 1 : si rien n'est fait, Poulpe attribuera à la variance la valeur zéro, ce qui évidemment la sous-estimera. Il faut donc approcher le véritable plan de sondage par des plans simplifiés ne faisant plus intervenir de tirages d'échantillons de taille 1.

4.1 Elimination du degré « secteur »

Une piste simplificatrice consiste à supprimer le degré « secteur » du plan de sondage :

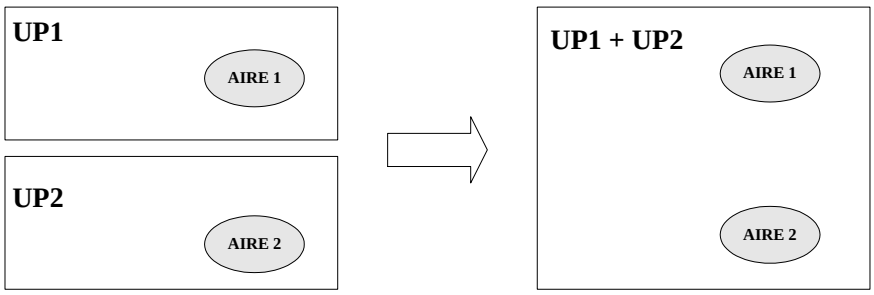


On fera donc « comme si » on tirait directement une aire par sondage aléatoire simple au sein d'une unité primaire. Cette modélisation simplifie le plan de sondage sans détériorer la variance du tirage. En effet, on vérifie que la variance associée au tirage à deux degrés d'une aire à l'intérieur d'un secteur, sous les hypothèses que le secteur est tiré dans l'UP proportionnellement à sa taille et que l'aire est tirée dans le secteur par sondage aléatoire simple, est égale à la variance que l'on obtiendrait si on tirait directement une aire par sondage aléatoire simple au sein de l'unité primaire**.** Le degré « secteur » n'a aucune influence sur la variance du plan de sondage : dans les calculs de précision, on pourra donc l'ignorer.

[11] On ignorera l'étape de tirage des grandes zones , qui est rare et qui relève de toutes façons de la même problématique.

4.2 Regroupement des UP

Malgré la simplification précédente, un problème demeure : au sein d'une unité primaire, on ne tire qu'une seule aire, si bien qu'on a toujours à faire à des échantillons de taille 1. Cependant, on peut regrouper deux par deux les unités primaires échantillonnées en construisant des « pseudo-UP », et imaginer qu'au lieu de tirer un échantillon d'unités primaires, on a tiré directement un échantillon de ces « pseudo-UP ». Au sein des « pseudo-UP », on considérera qu'on a tiré deux aires par sondage aléatoire simple :



La question se pose de savoir quel impact vont avoir ces regroupements d'UP sur la variance estimée. On se place conditionnellement au tirage des UP et on se donne deux unités primaires UP1 et UP2, de tailles respectives N_1 et N_2, au sein de chacune desquelles on tire une aire par sondage aléatoire simple (plan **P**). Soit une variable définie au niveau aire. Notons $\hat{\overline{Y}}$ l'estimateur d'Horvitz-Thompson de la moyenne, et **V** sa vraie variance sous le plan de sondage **P**. Supposons maintenant que l'on tire directement ces deux aires par sondage aléatoire simple au sein de la « pseudo-UP » constituée par le regroupement de UP1 et de UP2 : la variance de ce tirage « approché » est alors naturellement estimée par : $\hat{V}\left(\hat{\overline{Y}}\right) = (1-f)\frac{s^2}{2}$, où s^2 désigne la dispersion sur l'échantillon constitué par les deux aires et **f** représente le taux de sondage des aires. En supposant $N_1 \approx N_2$ (i.e. que UP1 et UP2 ont des tailles voisines) et que N_1 et N_2 sont « suffisamment » grands, on peut montrer :

$$E_P\left(\hat{V}\right) = V + f \cdot V + (1-f) \cdot \left(\frac{\overline{Y}_1 - \overline{Y}_2}{2} \right)^2$$

où $\overline{Y}_1$ (resp. $\overline{Y}_2$) est la vraie moyenne de la variable sur l'UP1 (resp. sur l'UP2). Ce résultat montre notamment que le regroupement des unités primaires va sur-estimer la vraie variance **V** conditionnelle (ce qui apparaît finalement plus « honnête » que de la sous-estimer). En outre, l'augmentation de la variance sera d'autant plus faible que les unités primaires que l'on regroupera auront des caractéristiques similaires. En conclusion, on aura intérêt à réaliser des « pseudo-UP » en regroupant par deux des unités primaires de tailles voisines et dont les moyennes sur des variables « importantes » pour l'enquête Emploi sont proches. Ce regroupement va évidemment influer en théorie sur la variance liée au tirage des UP elles-mêmes, et cela d'une manière complexe. Cependant, on vérifie que l'ordre de grandeur de l'estimateur de variance due au premier degré ne change pas (les probabilités de sélection ne changent pas, les totaux Y_i sont multipliés par deux mais le nombre de termes dans l'estimateur de variance est divisé par quatre).

Le choix des variables de regroupement est certes arbitraire, mais on considérera néanmoins des variables ayant une liaison évidente avec l'emploi. Pour pouvoir regrouper les UP, nous nous sommes appuyés sur des effectifs : cela permet en effet de prendre en compte à la fois les objectifs sur la taille et sur les profils. On a finalement utilisé les effectifs d'artisans, commerçants et chefs d'entreprise, les effectifs d'agriculteurs exploitants, de retraités, de chômeurs, d'ouvriers, et enfin le nombre de résidences secondaires.

Une unité primaire étant caractérisée par un 6-uplet composé des totaux de chacune des six variables précédentes, on cherche à regrouper les unités primaires dont ces totaux sont les plus proches. Une idée naturelle consiste à effectuer une analyse en composantes principales des unités primaires selon les six variables de regroupement, et à conserver les deux premières composantes principales c_1 et c_2. Cela revient à représenter chaque UP par un point (c_1 , c_2) du plan. Partant de là, on effectue un maillage du nuage des points obtenu à partir de ces deux composantes principales. On trie ensuite les UP selon les mailles auxquelles elles appartiennent et on termine en groupant les aires par paquets de deux, dans l'ordre du fichier ainsi trié.

4.3 Détermination du nombre d'aires découpées dans chaque UP

Le nombre d'aires découpées dans une UP doit être connu pour que le logiciel puisse calculer la probabilité d'inclusion d'une aire tirée au sein d'une UP : cette probabilité vaut en effet n/N, où **n** est le nombre d'aires que l'on tire (que Poulpe calcule à partir du fichier des données de l'enquête) et **N** est le nombre total d'aires découpées au sein de l'UP. Malheureusement, la valeur exacte de **N** n'est pas disponible car on n'a jamais découpé complètement les UP en aires. Ce problème a été résolu ainsi : puisque les aires contiennent en moyenne une vingtaine de logements, on a considéré que le nombre d'aires « potentiellement » découpées dans une UP est égal au rapport de la taille de l'UP (en nombre de logements ordinaires recensés), qui est connue, par le nombre moyen de logements que contient une aire, qui est par définition égal à vingt. Ainsi, une UP qui contient 1200 logements sera considérée comme découpée en 1200/20 = 60 aires.

4.4 Détermination du nombre de logements neufs inclus dans une aire

Cette valeur est indispensable pour calculer correctement la probabilité de tirage d'un logement neuf. Le nombre total de logements neufs présents dans une aire donnée n'étant pas disponible, il a fallu approximer cette valeur. En fait, celle-ci pourra être presque toujours déterminée de façon exacte à partir de la table des données de l'enquête. A partir de ce fichier, on connaît en effet le nombre **n** de logements neufs qui ont été tirés au sein d'une aire. Or, à partir de cette valeur, et en considérant la méthode d'échantillonnage des logements neufs décrite au II-1-c, on pourra retrouver, ou au pire « encadrer », la taille **N** de la population totale des logements neufs de l'aire. La règle est la suivante:

n < 10 : N = n
n > 10 : N = 4 × n
n = 10 : on ne sait pas, on pourra seulement conclure que $N \in [10,40]$

Le fait que l'on ne puisse pas déduire **N** lorsque n = 10 n'est pas très grave dans la mesure où on constate que la population des aires dans lesquelles on interroge exactement 10 logements neufs est marginale[12]. Finalement, l'information sur les logements neufs sera connue exactement pour la quasi-totalité des aires qui composent l'échantillon, ce qui constitue un résultat très satisfaisant. Pour les aires où on interroge 10 logements neufs, on choisira N = 40. Ceci constitue un parti de prudence dans la mesure où ce choix conduit à surestimer la variance.

[12] 8 aires au 1er trimestre 2003

5. Mise en œuvre du logiciel

Les calculs réalisés par le logiciel s'appuieront sur les informations issues de trois fichiers distincts :

- La table SAS modélisant le plan de sondage (table MODELE).
- La table SAS géographique (table GEO), contenant les informations nécessaires à Poulpe pour calculer les probabilités d'inclusion locales (c'est à dire relatives à un sondage élémentaire), puis globales, informations qui sont évidemment indispensables à un calcul de variance.
- La table SAS des données de l'enquête (table DATA) : cette table sera enrichie par des variables supplémentaires nécessaires aux calculs de Poulpe.

Ces trois tables devront être construites « à la main » par l'utilisateur. Ce travail est loin d'être évident : il impose notamment à l'utilisateur de maîtriser parfaitement le plan de sondage de son enquête et de disposer des données numériques permettant le calcul des probabilités d'inclusion des unités tirées, ce qui en général n'est pas une mince affaire ! Une fois que ces tables auront été construites, le logiciel pourra être lancé : leur élaboration constitue donc un travail préparatoire d'ampleur dans la mise en œuvre de Poulpe.

5.1 Construction de la table modélisant le plan de sondage

Pour qu'il soit exploité par Poulpe, l'arbre de sondage doit prendre la forme d'une table SAS. Chaque ligne de cette table va contenir les informations relatives à un sondage « élémentaire », c'est à dire à un arc de l'arbre**.** On obtient la table suivante :

NSUP	NINF	TYPTIR	ENTI	ENTAG	T_AILLE
ZZ	AA	TOT			
AA	BA	EXH	STRATE	STRATE	
BA	CA	PPT	STRATE UP	STRATE	nblogt
CA	DA	SAS	STRATE UP AIRE	STRATE UP	
DA	EA	EXH	STRATE UP AIRE NEUF	STRATE UP AIRE	
EA	FA	EXH		STRATE UP AIRE NEUF	
EA	FB	SAS		STRATE UP AIRE NEUF	

Cette table contient des variables obligatoires :

- Identifiants de l'arc (NSUP et NINF) : l'arc est identifié par les codes du nœud supérieur et du nœud inférieur. Ces codes ont deux lettres : par exemple AA, BA,…
- Type de tirage réalisé (TYPTIR) : ce code précise les formules à appliquer pour les calculs. Les codes reconnus par le logiciel sont donnés au III-2-a
- Identifiants de l'entité tirée (ENTI) et de l'entité d'agrégation (ENTAG), incluant notamment la liste des noms des entités qui ont été tirées aux degrés supérieurs.

et des variables spécifiques : lorsque le tirage est à probabilités inégales, en pratique proportionnel à la taille, il faut ajouter le nom de la variable qui sert à désigner la taille (T_AILLE).

5.2 Construction de la table géographique

Le fichier géographique apporte les informations nécessaires au calcul des probabilités d'inclusion. Ces renseignements sont absents du fichier des données de l'enquête et doivent donc être obtenus à partir de sources extérieures, ici les fichiers du recensement de la population. Le fichier géographique est structuré de façon à permettre un appariement ultérieur avec la table DATA des données de l'enquête : les unités sondées devront donc être reconnues dans les tables

DATA et GEO par des identifiants communs (c'est à dire des variables SAS ayant le même nom et le même format). Une ligne du fichier géographique caractérisera une unité sondée **i**. On trouvera également :

- Une variable « effectif » : elle donne le nombre total d'unités de degré immédiatement inférieur comprises dans l'unité **i** (variable NNN) ;
- Une variable « taille » : elle donne la taille de l'unité sondée (en nombre total de logements recensés), utilisée dans le cas d'un sondage à probabilités proportionnelles à la taille (variable TAILUNI).
- Une variable (AUXNIV) permettra de savoir à quel degré de tirage l'unité est sollicitée

Voici un extrait de la table géographique (on se souviendra que les UP sont en fait des « pseudo-UP » issues de regroupements d'UP) :

STRATE	UP	AIRE	AUXNIV	NNN	TAILUNI
235	02	1	3	21	21
235	02		2	112	2259
235			1	28	58991

Par exemple l'aire 235-02-1 contient 21 ménages, l'UP 235-02 contient 112 aires et 2259 ménages et la strate 235 contient 28 UP et 58991 ménages.

On tire des aires par sondage aléatoire simple à l'intérieur des UP. Le tirage des aires dans l'UP 235-02 va conduire au calcul du taux de sondage : la taille de la population **N** est égale à 112 et la taille de l'échantillon est calculée à partir du fichier DATA des données de l'enquête.

On tire des UP proportionnellement à leur taille à l'intérieur des strates. Le tirage de l'UP 235-02 dans la strate 235 va conduire au calcul de la probabilité d'inclusion $\pi_i = n\frac{X_i}{X}$, où **n** est la taille de l'échantillon d'UP (obtenue à partir du fichier DATA des données de l'enquête), X_i vaut 2259 (logements) et **X** vaut 58991 (logements).

5.3 Enrichissement de la table des données de l'enquête

La table DATA des données de l'enquête rassemble tous les individus tirés, répondants comme non-répondants. Si l'enquête est en deux phases, cette table doit contenir l'échantillon première phase. En effet, lorsque l'enquête se décompose en deux, voire trois phases, la connaissance de la taille de l'échantillon première phase est indispensable pour l'estimation de variance. On précisera, pour chaque individu, son identifiant complet. Pour pouvoir rapprocher la table DATA et la table MODELE, il faut introduire une variable précisant la position de chaque ménage dans l'arbre modélisant le plan de sondage : ce lien sera établi par le code de la feuille de l'arbre à laquelle est rattachée l'observation. On introduit donc ce code dans une nouvelle variable spécifique du fichier de données. Par exemple, avec notre plan sondage, tout logement recensé aura le code de feuille FA et tout logement neuf aura le code FB (voir III-2-a).

La table DATA sera également complétée par d'autres variables, en fonction du type de tirage réalisé :

- Dans le cas d'une enquête en plusieurs phases, la table DATA devra contenir une variable supplémentaire indiquant à quelle phase de tirage le logement est tiré (1,2 ou 3).
- Dans le cas d'une deuxième phase post-stratifiée (nous n'avons pas retenu cette modélisation), l'utilisateur devra rentrer dans la table la variable définissant les post-strates.

- Dans le cas d'une deuxième ou troisième phase de Poisson, il faudra préciser les probabilités de tirage (il s'agit des probabilités de réponse) de chaque logement.
- Si les données de l'enquête ont été redressées par CALMAR, il faudra inclure les variables de calage, afin de permettre le calcul des résidus.

Il a donc fallu injecter dans cette table différentes informations individuelles, dont les principales sont les suivantes :

5.3.1 Injection d'une variable de phase

Pour pouvoir prendre en compte la non-réponse dans le calcul de variance, on est amené à considérer l'enquête comme une enquête en deux phases pour laquelle l'échantillon de première phase est constitué des logements échantillonnés et l'échantillon de deuxième phase est constitué des logements qui ont effectivement répondu à l'enquête (totalement ou partiellement[13]). On a créé une variable de « phase » qui vaut 1 si le logement est non répondant, et 2 s'il est répondant.

5.3.2 Injection du code de feuille

Cette variable indique, pour chaque logement tiré, l'identifiant de la feuille de l'arbre de sondage à laquelle il se rapporte. Dans notre cas elle vaut 'FA' si le logement est un logement recensé (ou oublié au recensement) et 'FB' sinon.

5.3.3 Injection des probabilités de réponse estimées des logements

On a modélisé la non-réponse par un tirage de deuxième phase de Poisson. Il a fallu insérer dans la table des logements échantillonnés la probabilité de réponse de chaque logement (cette information n'est cependant utilisée par Poulpe que pour les logements répondants). Pour cela, nous avons déterminé, par un modèle logistique, un ensemble de variables qualitatives explicatives de la non-réponse. En croisant les modalités de ces différentes variables, on partitionne en « cellules » la population des logements. La probabilité de réponse d'un logement est estimée par le taux de réponse calculé dans la cellule qui le contient (cette estimation se justifie par le fait qu'il s'agit de l'estimateur du maximum de vraisemblance de la probabilité dans un tirage de Bernoulli). La probabilité de réponse des logements recensés dépend des variables suivantes, collectées au recensement de 1999 : nombre de pièces du logement, type de logement, nombre de personnes âgées de 24 ans ou moins, nombre de personnes âgées de 60 ans ou plus, région, et enfin tranche d'unité urbaine. Pratiquement, pour éviter d'avoir un nombre de cases trop élevé, on a regroupé certaines modalités de ces variables. Pour traiter la non réponse des logements neufs, les cellules résultent du croisement des variables ZEAT[14] et TU99.

Un ajustement « manuel » des regroupements a succédé à cette phase automatique, afin d'éviter que certaines cellules ne présentent de trop faibles effectifs tirés. Finalement, on obtient une classification de la population échantillonnée en une centaine de cellules homogènes. A titre d'information, les taux de réponse par cellule varient entre 68,9% et 94,8%, avec une moyenne égale à 84,7% et un écart-type de 0,06.

[13] Partiellement signifie que seulement certaines personnes du logement n'ont pas répondu à l'enquête. Dans ce cas, il y a une repondération qui consiste à multiplier le poids initial commun aux individus du logement par l'inverse du taux de réponse des individus dans ce logement. Cette forme de non-réponse ne sera pas prise en compte dans la modélisation du plan de sondage, mais ses effets numériques sont totalement négligeables car il y a seulement une centaine de logements concernés chaque trimestre.

[14] Variable caractérisant des regroupements de régions contiguës.

5.4 Lancement de Poulpe

Concrètement, Poulpe se présente sous la forme d'une succession d'écrans de dialogue, au travers desquels on entrera des paramètres venant alimenter des macros SAS. Quatre (ou cinq) grandes étapes sont nécessaires au logiciel pour réaliser un calcul de variance.

ETAPE 1 : Enrichissement et contrôle de l'arbre décrivant le plan de sondage

Poulpe enrichit la table MODELE en lui adjoignant de nouvelles variables précisant en particulier la structure de l'arbre de sondage et l'ordre dans lequel les arcs devront être traités.

ETAPE 2 : Calcul des probabilités d'inclusion

Dans cette étape, le logiciel commence par calculer les probabilités d'inclusion (globales) de première phase en faisant le produit des probabilités d'inclusion locales (c'est à dire celles relatives aux sondages « élémentaires »). Pour cela, Poulpe utilise la table des données de l'enquête qui fournit les tailles d'échantillon, et la table géographique qui fournit à chaque degré la taille des unités tirées et la taille des unités d'agrégation. S'il y a plusieurs phases, la probabilité d'inclusion finale calculée par le logiciel sera égale au produit des probabilités d'inclusion relatives à chacune des phases du tirage. C'est cette valeur qui permettra de produire l'estimation Somlog, dont l'intérêt essentiel est, par comparaison avec Sompond, de détecter d'éventuelles erreurs dans le fichier géographique (voir III-1).

ETAPE 3 : Choix des variables d'intérêt pour lesquelles on souhaite une estimation de variance.

ETAPE 4 : Estimation de variance pour des estimateurs de totaux

Dans cette étape, Poulpe estime la variance du total des variables d'intérêt qui ont été déclarées à l'étape précédente. Pour cela, le logiciel s'appuiera sur l'arbre de sondage et mobilisera la formule de Raj et les expressions analytiques « élémentaires » de variance.

ETAPE 5 : Estimation de variance pour des estimateurs de paramètres « complexes ».

Poulpe estimera directement la variance d'un ratio en linéarisant l'estimateur. L'utilisateur devra définir les ratios en respectant la syntaxe suivante :

%div(tx,chômeur,actif) ;%listfonc(tx) ;

La macro %div définit le paramètre complexe tx comme étant le rapport du nombre de chômeur sur le nombre d'actifs : on cherche ainsi à obtenir une estimation de la variance du taux de chômage tx. La macro %listfonc indique au logiciel que c'est sur le paramètre tx que l'on souhaite une estimation de variance. S'il s'agit de statistiques autres que des ratios, l'utilisateur doit faire figurer directement la variable linéarisée dans la table des données DATA.

A chacune de ces cinq étapes correspondra une macro qui construira des tables SAS intermédiaires. C'est l'ensemble des tables construites aux étapes préparatoires 1, 2 et 3 qui va permettre à Poulpe d'effectuer le calcul de précision lors des étapes 4 et/ou 5.

6. Quelques résultats numériques :

Ces estimations portent sur le premier trimestre 2003. Elles ne concernent, pour l'instant, que des paramètres « transversaux », c'est-à-dire considérés un trimestre donné et définis sur la population courante. Pour cet objectif, le schéma rotatif n'apporte aucune complication, et il est d'ailleurs totalement « transparent ». L'estimation des évolutions sera plus complexe et viendra ultérieurement.

6.1 Précisions associées à des variables d'intérêt fondamentales

6.1.1 Estimations de totaux

VARIABLE	TOTAL ESTIME	ECART-TYPE ESTIME	INTERVALLE DE CONFIANCE DE NIVEAU 95% POUR LE TOTAL		CV (%)	DEFF (%)
			BORNE INFERIEURE	BORNE SUPERIEURE		
NOMBRE DE CHOMEURS AU NIVEAU NATIONAL						
ENSEMBLE	**2 684 701**	**56 992**	**2 572 996**	**2 796 406**	**2.1**	**1.85**
HOMME	**1 289 136**	**35 458**	**1 219 638**	**1 358 634**	**2.7**	**1.63**
FEMME	**1 395 565**	**34 244**	**1 328 447**	**1 462 683**	**2.4**	**1.37**
15-29 ANS	**918 009**	**30 190**	**858 836**	**977 182**	**3.3**	**1.66**
HOMME	**472 643**	**19 663**	**434 104**	**511 182**	**4.2**	**1.37**
FEMME	**445 366**	**18 157**	**409 778**	**480 954**	**4.1**	**1.29**
30-49 ANS	**1 305 894**	**33 650**	**1 239 940**	**1 371 848**	**2.6**	**1.41**
HOMME	**576 036**	**20 735**	**535 395**	**616 677**	**3.6**	**1.31**
FEMME	**729 858**	**24 002**	**682 815**	**776 901**	**3.3**	**1.29**
+50 ANS	**460 798**	**21 785**	**418 099**	**503 497**	**4.7**	**1.56**
HOMME	**240 457**	**15 082**	**210 897**	**270 017**	**6.3**	**1.54**
FEMME	**220 341**	**12 959**	**194 941**	**245 741**	**5.9**	**1.16**
NOMBRE DE CHOMEURS PAR REGION						
ILE DE France	**541 076**	**34 361**	**473 727**	**608 425**	**6.3**	**3.28**
RHONE –ALPES	**221 330**	**14 382**	**193 141**	**249 519**	**6.5**	**1.26**
AUVERGNE	**43 763**	**6 308**	**31 398**	**56 127**	**14.4**	**1.29**
NORD - PAS DE CALAIS	**219 867**	**15 418**	**189 648**	**250 086**	**7.0**	**1.44**
NOMBRE D'ACTIFS OCCUPES						
ENSEMBLE	**24 387 066**	**94 629**	**24 201 594**	**24 572 538**	**0.4**	**1.94**
HOMME	**13 318 005**	**55 914**	**13 208 413**	**13 427 597**	**0.4**	**1.76**
FEMME	**11 069 061**	**62 353**	**10 946 848**	**11 191 274**	**0.6**	**1.63**

6.1.2 Estimations de ratios

VARIABLE	RATIO ESTIME (%)	ECART-TYPE ESTIME (%)	INTERVALLE DE CONFIANCE DE NIVEAU 95% POUR LE RATIO		CV (%)	DEFF (%)
			BORNE INFERIEURE	BORNE SUPERIEURE		
TAUX DE CHOMAGE AU NIVEAU NATIONAL (%)						
ENSEMBLE	**9.9**	**0.22**	**9.5**	**10.3**	**2.2**	**1.96**
HOMME	**8.8**	**0.26**	**8.3**	**9.3**	**2.9**	**1.68**
FEMME	**11.2**	**0.28**	**10.6**	**11.7**	**2.5**	**1.46**
15-29 ANS	**16.9**	**0.57**	**15.8**	**18.0**	**3.4**	**1.64**
HOMME	**15.9**	**0.68**	**14.5**	**17.2**	**4.3**	**1.34**
FEMME	**18.2**	**0.76**	**16.7**	**19.7**	**4.2**	**1.37**
30-49 ANS	**8.6**	**0.24**	**8.2**	**9.1**	**2.8**	**1.47**
HOMME	**7.1**	**0.28**	**6.6**	**7.7**	**3.9**	**1.33**
FEMME	**10.4**	**0.35**	**9.7**	**11.0**	**3.7**	**1.35**
+50 ANS	**7.1**	**0.33**	**6.4**	**7.7**	**4.6**	**1.58**
HOMME	**6.8**	**0.43**	**5.9**	**7.6**	**6.3**	**1.57**
FEMME	**7.4**	**0.42**	**6.6**	**8.2**	**5.7**	**1.14**
TAUX DE CHOMAGE PAR REGION (%)						
ILE DE France	**10.0**	**0.59**	**8.8**	**11.1**	**5.9**	**2.65**
NORD-PAS DE CALAIS	**12.6**	**0.92**	**10.8**	**14.4**	**7.3**	**1.70**
RHONE-ALPES	**8.6**	**0.52**	**7.6**	**9.7**	**6.0**	**1.22**
AUVERGNE	**7.6**	**0.91**	**5.9**	**9.4**	**12.0**	**0.78**

Les résultats numériques obtenus pour des totaux et pour des ratios sont cohérents. Pour les estimations au niveau national, on vérifie que plus l'effectif de la population concernée est faible, plus le coefficient de variation (CV) est grand. Pour les estimations au niveau régional, le coefficient de variation augmente si la région est petite (Auvergne). On remarque que l'effet de sondage pour l'estimation du nombre total de chômeurs en Ile de France est très fort (3.28), traduisant un fort effet de grappe au niveau des aires découpées en Ile de France. Ce phénomène se manifeste également lorsqu'on estime le taux de chômage en Ile de France (DEFF de 2.65).

6.2 Influence du nombre de variables de calage

Nous avons eu l'occasion de tester l'efficacité des calages et l'influence du nombre de variables de calage (NVAR) sur la précision des résultats. Il y a au total 5 variables et 57 modalités - donc 57 effectifs - utilisées pour le calage complet (voir II-2). De façon un peu surprenante, la précision associée à un nombre restreint de modalités de calage n'est pas vraiment sensible à la liste des modalités retenues (une sélection de modalités conduit aux tableaux ci dessous).

6.2.1 Pour des totaux

VARIABLE	AVEC TOUTES LES VARIABLES DE CALAGE	SANS CALAGE	NVAR = 5	NVAR = 10	NVAR = 20	NVAR = 30	NVAR = 40
NOMBRE TOTAL DE CHOMEURS AU NIVEAU NATIONAL							
ENSEMBLE	**56 992**	**67 559**	**58 600**	**58 516**	**58 604**	**59 309**	**59 029**
HOMME	**35 458**	**39 328**	**36 445**	**36 493**	**36 803**	**36 317**	**36 224**
FEMME	**34 244**	**40 531**	**35 878**	**35 825**	**35 231**	**35 412**	**35 434**
15-29 ANS	**30 190**	**36 357**	**33 189**	**31 648**	**31 458**	**30 364**	**30 502**
HOMME	**19 663**	**22 211**	**21 055**	**19 830**	**19 810**	**19 743**	**19 833**
FEMME	**18 157**	**21 637**	**20 042**	**19 872**	**19 679**	**18 298**	**18 308**
30-49 ANS	**33 650**	**37 919**	**34 669**	**34 588**	**34 467**	**34 698**	**34 646**
HOMME	**20 735**	**22 079**	**21 364**	**21 282**	**21 296**	**21 096**	**21 094**
FEMME	**24 002**	**26 664**	**24 762**	**24 759**	**24 474**	**24 426**	**24 398**
+50 ANS	**21 785**	**22 965**	**22 456**	**22 326**	**21 850**	**22 068**	**21 913**
HOMME	**15 082**	**16 121**	**15 749**	**15 661**	**15 396**	**15 222**	**15 221**
FEMME	**12 959**	**13 280**	**13 225**	**13 205**	**13 106**	**13 044**	**13 045**
NOMBRE DE CHOMEURS PAR REGION							
ILE DE France	**34 361**	**34 512**	**34 216**	**34 546**	**34 546**	**34 788**	**34 687**
RHONE –ALPES	**14 382**	**14 374**	**14 261**	**14 216**	**14 222**	**14 238**	**14 259**
AUVERGNE	**6 308**	**6 346**	**6 329**	**6 327**	**6 313**	**6 305**	**6 307**
NORD - PAS DE CALAIS	**15 418**	**15 214**	**15 232**	**15 201**	**15 211**	**15 246**	**15 257**
NOMBRE D'ACTIFS OCCUPES							
ENSEMBLE	**94629**	**308165**	**186212**	**162651**	**126043**	**97030**	**107181**
HOMME	**55914**	**172818**	**108295**	**86846**	**57499**	**56329**	**66831**
FEMME	**62353**	**150246**	**101212**	**98209**	**91721**	**63879**	**64265**

6.2.2 Pour des ratios

VARIABLE	AVEC TOUTES LES VARIABLES DE CALAGE	SANS CALAGE	NVAR = 20
TAUX DE CHOMAGE AU NIVEAU NATIONAL (%)			
ENSEMBLE	**0.22**	**0.23**	**0.23**
HOMME	**0.26**	**0.26**	**0.27**
FEMME	**0.28**	**0.30**	**0.30**
15-29 ANS	**0.57**	**0.57**	**0.58**
HOMME	**0.68**	**0.68**	**0.69**
FEMME	**0.76**	**0.78**	**0.77**
30-49 ANS	**0.24**	**0.25**	**0.25**
HOMME	**0.28**	**0.29**	**0.29**
FEMME	**0.35**	**0.36**	**0.36**
+50 ANS	**0.33**	**0.34**	**0.34**
HOMME	**0.43**	**0.45**	**0.44**
FEMME	**0.42**	**0.42**	**0.42**
TAUX DE CHOMAGE PAR REGION (%)			
ILE DE France	**0.59**	**0.59**	**0.59**
RHONE -ALPES	**0.52**	**0.52**	**0.52**
AUVERGNE	**0.90**	**0.90**	**0.90**
NORD - PAS DE CALAIS	**0.92**	**0.92**	**0.92**

De façon générale, le calage diminue (logiquement) la variance dans le cas où le paramètre est un total défini sur l'ensemble de la population : cela se vérifie très clairement pour le nombre total de chômeurs et pour le nombre total d'actifs occupés (mais le gain est beaucoup plus spectaculaire pour le nombre d'actifs que pour le nombre de chômeurs). On peut néanmoins noter que l'écart type estimé n'est pas une fonction rigoureusement décroissante de la taille de l'échantillon, parce que d'une part, le plan est complexe, et d'autre part Poulpe manipule des résidus estimés et non des résidus exacts. Pour les sous-populations croisant sexe et âge, on note une certaine amélioration de la qualité lorsqu'on cale. Le gain reste néanmoins modéré parce qu'en dehors des variables sexe et âge, le calage porte sur des totaux relatifs à la population entière et non pas à la population de sexe et âge concernés. En revanche, quand l'estimation porte sur un domaine (par exemple les régions) ou sur un ratio (par exemple le taux de chômage), il n'y a pas d'amélioration.

* * * * *