Méthodologie statistique

M 2016/02

Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu

Emmanuel Gros – Karim Moussallam

Document de travail

Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2016/02

Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu

Emmanuel Gros * - Karim Moussallam **

Les auteurs remercient Dominique Place – qui a initié les réflexions méthodologiques sur le calcul de précision dans l'Enquête Emploi en Continu –, Josiane Le Guennec – qui s'est investie dans les formules et programmes de calcul de variance dans le cadre de ses travaux sur la concaténation d'enquêtes successives dans l'enquête Emploi pour des estimations sur petits domaines – et Sébastien Faivre – pour sa relecture attentive.

* Insee – DMCSI – Division Sondages
18 boulevard Adolphe Pinard - 75675 PARIS CEDEX 14
** Au moment de la rédaction de ce document de travail, Insee – DMCSI – Division Sondages
18 boulevard Adolphe Pinard - 75675 PARIS CEDEX 14

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101
18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 33 - Fax : 33 (1) 41 17 61 97 - CEDEX - E-mail :-DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.
Working papers do not reflect the position of INSEE but only their author's views.

Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu

Emmanuel Gros * – Karim Moussallam **

Résumé

L'Enquête Emploi en Continu, qui constitue l'une des principales enquêtes ménages de l'Insee ainsi que l'une des pièces centrales du dispositif statistique de connaissance de l'emploi et du chômage, résulte d'un plan de sondage complexe, qui rend particulièrement ardu les calculs de précision portant sur les statistiques issues de cette enquête. Ce document de travail représente l'aboutissement de plusieurs années de travail sur le sujet.

La première partie, intitulée « Compendium », revient sur le cadre méthodologique de l'Enquête Emploi en Continu et présente d'une part les principales options méthodologiques retenues pour les calculs de précision, tant en transversal qu'en longitudinal, et d'autre part les principaux résultats obtenus sur l'année 2012. Les parties suivantes exposent beaucoup plus en détail les différentes formules de variance, leur justification et les résultats obtenus.

Nota bene : les estimations – taux de chômage, variation trimestrielle du nombre de chômeurs – présentées dans ce document de travail correspondent à des résultats bruts issus d'estimations directes réalisées de manière ad hoc pour les calculs de variance. Elles ne correspondent donc pas aux estimations officielles diffusées par l'Insee, ni au niveau national – du fait notamment de l'absence de désaisonnalisation – ni au niveau régional – pour lequel les estimations officielles (les « taux de chômage localisés »), également désaisonnalisées, sont obtenues selon une méthode composite mobilisant les informations de l'Enquête Emploi en Continu et des clefs de répartitions fondées sur les demandeurs d'emploi en fin de mois (DEFM).

Les résultats portent sur la période 2003-2012, avant la rénovation de l'enquête Emploi qui a eu lieu au 1er janvier 2013.

Mots clés : Enquête Emploi en Continu, estimation de variance transversale, estimation de variance longitudinale.

Abstract

The Labour Force Survey is one of the major household surveys conducted by INSEE as well as a centrepiece of the French employment and unemployment statistical system. It results from a complex sample design, which makes variance estimation particularly difficult. This working paper constitutes the outcome of years of work on this subject.

The first part, titled "Compendium", draws the methodological framework of the Labour Force Survey and presents on the one hand the main methodological choices made for the variance estimation, boss for cross-sectional and longitudinal estimations, and on the other hand the main results for 2012. The following sections set out the different formulas, their justification and the different results obtained with more details.

Nota bene : the estimates – unemployment rates, quarterly variation of the number of unemployed – presented in this paper correspond to raw results from direct estimates performed on an ad hoc basis for the variance estimation. They do not tally with the official estimates released by INSEE, neither at the national level – in particular due to the lack of seasonal adjustment – not at the regional level – for which official estimates (the "localised unemployment rates"), also seasonally adjusted, are computed according to specific estimators mobilizing the ongoing Labour Force Survey and allocation keys based on end-of-month jobseekers.
The results refer to the period 2003-2012, before the updating of the questionnaire at the beginning of 2013.

KeyWords : Labour Force Survey, cross-sectional variance estimation, longitudinal variance estimation.

* DMCSI – Département des Méthodes Statistiques – Division Sondages
emmanuel.gros@insee.fr
** Au moment de la rédaction de ce document de travail, Insee – DMCSI – Division Sondages
karim.moussallam@travail.gouv.fr

Table des matières

Compendium 1
Partie I Méthode d'estimation de la précision transversale 10
I.1 principe général de l'estimation 10
I.2 estimation de la variance de première phase 12
I.3 stratégie de collapse des unités primaires 15
I.4 quelques résultats de l'estimation de variance transversale 18
Partie II Méthode d'estimation de la précision longitudinale et évaluation par rapport aux objectifs d'Eurostat 21
II.1 décomposition de la covariance 21
II.2 estimation de la covariance de première phase 21
II.3 deux applications du calcul de variance longitudinale 23
II.4 rappel des objectifs de précision d'Eurostat 26
II.5 résultats obtenus pour les critères d'Eurostat 28
Annexe A estimation d'une forme quadratique de totaux par strate 36
A.1 principe général 36
A.2 application au 'collapse' d'unités primaires 38
Annexe B résultats détaillés de l'estimation transversale 39
Annexe C estimation transversale sur la période de l'échantillonnage RP1999 42
C.1 variance de première phase sur l'échantillon RP 42
C.2 partage des poids RP-TH 42
Annexe D une interprétation du facteur de proportionnalité pour la variance de la moyenne annuelle 43
Annexe E évaluation d'une prise en compte partielle des covariances intra-unité primaire . . . 44
Annexe F covariance temporelle entre estimateurs du chômage 47
Annexe G estimation longitudinale sur la période de transition entre les échantillonnages RP et TH 48

Compendium

L'Enquête Emploi en Continu (EEC) – qui constitue l'unique source d'information permettant de mettre en œuvre la mesure de l'activité suivant les concepts du Bureau international du Travail (BIT) – est l'une principales enquêtes de l'Insee ainsi que l'une des pièces centrales du dispositif statistique de connaissance de l'emploi et du chômage. S'inscrivant dans le cadre des enquêtes sur les forces de travail réalisées dans tous les pays de l'Union européenne (« Labour Force Survey »), elle permet en effet à la France de comparer le niveau et l'évolution de son taux de chômage avec ceux des autres pays européens, mais aussi avec ceux des pays membres de l'organisation internationale du Travail (OIT). Elle apporte en outre de nombreuses informations[1] sur l'état du marché du travail et son évolution et fournit également des données originales sur la structure par professions, sur l'activité des femmes ou des jeunes, sur la durée du travail et sur les emplois temporaires.

En conséquence, la mesure de la précision des statistiques issues de cette enquête constitue un enjeu majeur, afin d'une part de permettre une interprétation juste et complète des différents résultats de l'EEC, et d'autre part de pouvoir quantifier dans quelle mesure la France répond aux exigences de précision imposées par Eurostat sur la mesure des statistiques de chômage. En s'appuyant sur des premières réflexions méthodologiques relatives au calcul de précision dans l'EEC initiées à partir de 2008 par Dominique Place, Karim Moussallam a mené un travail complet et rigoureux d'élaboration et de mise en œuvre des formules de calcul de variance présentées dans ce document de travail.

L'objectif de ce compendium est de présenter de manière synthétique, d'une part les principales options méthodologiques retenues pour les calculs de précision dans l'EEC, tant en transversal qu'en longitudinal, et d'autre part les principaux résultats obtenus sur l'année 2012. **Dans ce document de travail, le terme « transversal » renvoie à des estimations de statistiques pour un trimestre donné à partir du seul échantillon relatif à ce trimestre ; le terme « longitudinal » fait quant à lui référence à des estimations faisant intervenir au moins deux échantillons trimestriels : en pratique, il s'agit d'estimations de précisions relatives à des moyennes annuelles ou à des variations d'un trimestre à l'autre.** Les différentes formules de variance sont présentées et démontrées dans les parties I et II.

1. L'enquête Emploi en Continu

C'est une enquête trimestrielle conduite[2] sur un échantillon d'environ 67 000 logements ordinaires identifiés comme résidences principales en France métropolitaine, ce qui conduit en moyenne chaque trimestre à environ 110 000 personnes de 15 ans ou plus répondantes, réparties dans environ 57 000 ménages.

Ses objectifs majeurs sont de fournir des informations quantitatives et structurelles sur le marché du travail français, notamment en évaluant chaque trimestre la population active, le nombre de chômeurs et le taux de chômage au sens du Bureau International du Travail ainsi que les évolutions trimestrielles du chômage. Le règlement européen n° 577/98 définit deux critères de précision sur les estimations de chômage :

[1] Volume de la population active, nombre de chômeurs et taux de chômage, caractéristiques des personnes présentes sur le marché du travail, de celles qui arrivent ou le quittent, etc.

[2] Depuis sa refonte engagée en 2008 et achevée au quatrième trimestre de 2011, qui s'est traduite en particulier par un augmentation de 50 % de la taille de l'échantillon.

- pour un groupe de chômeurs représentant 5 % de la population d'âge actif, le coefficient de variation pour l'estimation des moyennes annuelles ne doit pas excéder 8 % au niveau des régions de plus de 300 000 habitants ;

- l'écart type des estimations de variations au niveau national entre deux trimestres consécutifs ne doit pas dépasser 2 % du nombre de chômeurs.

Ces deux critères sont contradictoires : pour satisfaire le premier, la solution optimale est de renouveler complètement l'échantillon entre deux trimestres pour interroger le plus de personnes différentes en une année, tandis que pour le second, c'est l'approche panel qui est la plus adaptée. Le schéma de rotation choisi constitue un compromis entre ces deux contraintes, tout en sachant que la première est délicate à respecter pour les régions de taille réduite. Les ménages sont interrogés six trimestres de suite, et chaque trimestre, un sixième de l'échantillon est renouvelé. Les première et dernière interrogations se font en face en face tandis que les interrogations intermédiaires se font par téléphone.

Une contrainte pèse aussi fortement sur la collecte : les personnes sélectionnées doivent être interrogées au sujet de leur situation sur le marché du travail lors d'une semaine de référence fixée au préalable, pendant les deux semaines qui suivent celle-ci. L'échantillon est réparti de manière uniforme selon toutes les semaines d'un trimestre[3] et la collecte se réalise en continu durant toutes les semaines d'un trimestre.

1.1. Le plan de sondage de l'EEC (exposé plus en détails dans [1])

Fin 2008, les échantillons trimestriels utilisés sur la période 2009 à 2018 ont été sélectionnés en une seule fois[4] dans le fichier de taxe d'habitation 2006 selon un plan de sondage relativement complexe, visant à permettre le remplacement progressif de l'échantillon issu du RP 1999 par celui tiré dans la TH et à assurer le renouvellement au sixième de l'échantillon chaque trimestre. En première approximation, on peut toutefois considérer que, pour un trimestre donné, l'échantillon de l'EEC est issu du plan de sondage à deux degrés suivant :

❶ au premier degré, sélection de secteurs – unités primaires regroupant six ou sept unités secondaires appelées « grappes » – selon un plan de sondage stratifié par région avec allocations proportionnelles, et au sein de chaque région, tirage équilibré sur différents critères[5], avec des probabilités proportionnelles au nombre de résidences principales des secteurs ;

❷ au second degré, sélection d'une grappe – ensemble de logements contigus comprenant une vingtaine de résidences principales – par secteur par tirage équilibré[6] avec mise en commun[7] des phases d'atterrissage.

---

[3] Un trimestre EEC normal comprend treize semaines mais comme une année fait 52 semaines plus un ou deux jours selon qu'elle est bissextile ou non, on ajoute tous les cinq ou six ans une semaine au quatrième trimestre pour rattraper le décalage avec l'année civile.

[4] Chaque année, les échantillons impliqués dans les enquêtes trimestrielles à venir sont mis à jour à partir du dernier fichier de taxe d'habitation connu. Cette mise à jour consiste en un double processus, d'une part d'enrichissement par les logements nouveaux apparus depuis la dernière mise à jour, et d'autre part de rafraîchissement des informations auxiliaires relatives aux logements et nécessaires au repérage sur le terrain et aux opérations de correction de la non-réponse.

[5] La liste détaillée des variables d'équilibrage de premier degré est disponible dans [1] au §5.1.

[6] La liste détaillée des variables d'équilibrage de second degré est disponible dans [1] aux §5.2 à 5.4.

[7] L'échantillonnage équilibré stratifié avec mise en commun des phases d'atterrissage est décrit en détail dans [2]. Le principe en est le suivant :

- on essaye d'abord d'équilibrer par strate : on réalise une phase de vol indépendamment dans chaque strate, en équilibrant sur les variables auxiliaires ;

La construction des unités primaires et secondaires a été réalisée selon une logique ascendante : toutes les adresses du fichier de la taxe d'habitation ont été réparties en des grappes comprenant chacune une vingtaine de logements déclarés comme résidences principales, en suivant les références cadastrales, les noms des voies et la numérotation dans les voies. Ensuite, ces grappes ont été regroupées en secteurs par 6 ou 7. Ce sont ces secteurs[8] qui constituent les unités primaires du plan de sondage.

1.2. Redressements et estimateurs

Les probabilités de sélection des secteurs sont proportionnelles à leur taille, mesurée par le nombre de résidences principales au sens de la TH. Soit un secteur i appartenant à la région h, sa probabilité d'inclusion est $\pi_{Ii} = M_h N_i / N_h$, avec M_h le nombre de secteurs à sélectionner dans la région h et N_i (respectivement N_h) le nombre de logements déclarés comme principaux dans les fichiers fiscaux pour le secteur i (respectivement la région h). Puis, étant donné que l'on sélectionne au second degré une grappe par secteur, la probabilité de sélection d'une grappe k au sein d'un secteur i tiré au 1er degré est égales à $\pi_{k|i} = 1/n_i$ avec n_i le nombre de grappes dans le secteur i. Au final, la probabilité d'inclusion d'une grappe k dans l'échantillon trimestriel de l'EEC est donc $\pi_k = \pi_{Ii}\pi_{k|i} = \frac{M_h N_i}{N_h n_i}$, et l'estimateur initial d'un total T_Y est le π-estimateur usuel : $\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}$.

Chaque trimestre, les fluctuations d'échantillonnage et la non-réponse sont corrigées par une procédure de calage de niveau logement en une étape, impliquant les unités répondantes et hors champ et effectuée par rang d'interrogation. Plus précisément, pour un trimestre donné, l'échantillon des logements répondants ou hors champ[9] est découpé en six sous-échantillons, caractérisés chacun par le rang d'interrogation des grappes qui le composent (un sous-échantillon constitué des grappes entrantes, un sous-échantillon constitué des grappes en deuxième interrogation, etc.) ; chaque sous-échantillon est ensuite repondéré de manière séparée par calage sur un ensemble[10] de variables avec des marges relatives à l'ensemble de la population ; enfin, les poids ainsi obtenus sont divisés par 6 pour aboutir aux poids finaux w_k. Au final, l'estimateur utilisé est l'estimateur calé $\hat{Y}_w = \sum_{k \in R} w_k y_k$, R désignant l'ensemble des répondants.

2. Principes généraux et éléments théoriques relatifs aux méthodes d'estimation de précision proposées pour l'EEC

2.1. Hypothèse simplificatrice relative au plan de sondage de second degré

Dans le plan de sondage retenu en première approximation pour décrire la sélection d'un échantillon trimestriel de l'EEC et exposé en 1.1, la sélection des grappes de logements n'est

---

– quand il n'est plus possible d'équilibrer par strate, on recherche un équilibrage global : on réunit les unités qui n'ont encore été ni échantillonnées ni rejetées pendant les différentes phases de vol, puis on réalise une dernière phase de vol sur ces unités avant d'effectuer une phase d'atterrissage.

[8] Dans les zones peu denses, les secteurs ont été regroupés par deux ou trois dans le cadre des frontières cantonales pour constituer ce qu'on appellera des bisecteurs (même contenant trois secteurs).

[9] Le fait que les unités hors-champ participent au calage s'explique par le fait que certaines marges de calage sont calculées à partir de la base de sondage.

[10] Décrites en détail dans [3].

pas strictement indépendante entre les secteurs du fait des tirages équilibrés avec mise en commun des phases d'atterrissage. Cette absence d'indépendance entre les tirages de second degré rendant inextricables les calculs de variance, nous avons dû négliger cet équilibrage de second degré. On suppose donc dans la suite que les tirages des grappes sont bien indépendants entre les secteurs, ce qui permet l'application des formules usuelles pour les plans de sondages à plusieurs degrés, alors qu'une estimation dans le cadre général d'un plan à deux phases, qui serait approprié ici, est impossible à mettre en œuvre.

2.2. Principes centraux des méthodes d'estimation de variance proposées

Les méthodes d'estimation de variance proposées reposent essentiellement sur l'application itérée des deux formules fondamentales suivantes :

- d'une part la formule de décomposition de la variance :

  > pour un estimateur $\hat{Y}$ calculé sur un échantillon sélectionné en deux phases, on a
  >
  > $$V(\hat{Y}) = V_1[E_{2|1}(\hat{Y})] + E_1[V_{2|1}(\hat{Y})]$$
  >
  > l'indice 1 faisant référence à l'aléa de tirage de l'échantillon de première phase et l'indice 2 à l'aléa de tirage de l'échantillon de seconde phase ;

- d'autre part la propriété suivante, relative aux tirages à deux degrés avec tirages des unités secondaires indépendants entre les unités primaires, et démontrée en annexe A :

  > Soit un plan de sondage à deux degrés, au sein duquel d'une part le plan de sondage de premier degré conduit à un échantillon S_I de m unités primaires sélectionnées selon des probabilités d'inclusion π_i et à des estimateurs sans biais $\hat{T}_i$ des vrais totaux T_i par unités primaires, et d'autre part les unités secondaires sont tirées indépendamment d'une unité primaire à l'autre. Si l'on dispose d'une forme quadratique $Q(T_1,...,T_m) = \sum_{i\in S_I} q_i T_i^2 + \sum_{(i,j)\in S_I,\, i\neq j} q_{ij} T_i T_j$ permettant d'estimer sans biais sur l'échantillon S_I la variance relative au premier degré de sondage en fonction des vrais totaux T_i par unité primaire, alors on montre que :
  >
  > $$E[Q(\hat{T}_1,...,\hat{T}_m)] = Q(T_1,...,T_m) + \sum_{i\in S_I} q_i V(\hat{T}_i)$$
  >
  > Par conséquent, si l'on dispose d'un estimateur sans biais de la variance de $\hat{T}_i$ liée au second degré de sondage au sein de l'unité primaire i, on peut estimer sans biais à partir de l'échantillon final S la variance relative au premier degré de sondage grâce à la formule suivante :
  >
  > $$\hat{V}_1(\hat{T}) = Q(\hat{T}_1,...,\hat{T}_m) - \sum_{i\in S_I} q_i \hat{V}(\hat{T}_i),$$

L'utilisation combinée de ces deux propriétés permet d'établir la formule d'estimation de variance de Rao – que l'on trouve également démontrée dans [4], page 40 :

Formule de Rao

Avec les notations précédemment introduites, la variance de l'estimateur d'Horvitz-Thompson $\hat{T}$ d'un vrai total T dans le cadre d'un sondage à deux degrés avec tirages des unités secondaires indépendants entre les unités primaires est estimée sans biais par :

$$\hat{V}(\hat{T}) = Q(\hat{T}_1,..., \hat{T}_m) + \sum_{i \in S_I} \left(\frac{1}{\pi_i^2} - q_i \right) \hat{V}(\hat{T}_i)$$

Cette formule permet de calculer, via une approche récursive, la précision d'estimateurs dans le contexte de plan de sondage à n degrés, sous réserve qu'à chaque degré les tirages d'unités « secondaires » soient effectués indépendamment d'une unité « primaire » à l'autre, et dès lors que l'on dispose de formules de variance permettant d'estimer sans biais les variances relatives à chaque degré.

2.3. Grandes lignes de la méthode d'estimation de variance transversale

La formule de variance retenue pour la précision transversale – i.e. relative à un trimestre donné – de l'EEC prend en compte le plan de sondage à deux degrés de l'échantillon trimestriel, l'étape de correction de la non-réponse totale – modélisée par un mécanisme poissonnien – et l'effet du calage sur marges, en procédant comme suit :

- Le calage de l'estimateur $\hat{Y}_w$ est pris en compte de manière usuelle, en faisant porter le calcul de variance non pas sur la variable Y elle-même mais sur les résidus de sa régression, pondérée par les poids de sondage, sur les variables de calage ;

- la non-réponse à l'enquête emploi est modélisée par un mécanisme poissonnien. Dans ce cadre, la réponse est indépendante entre les logements conditionnellement à l'échantillon de l'enquête. On peut donc voir la phase de non-réponse comme un troisième degré de sondage : au sein de chaque logement de l'échantillon, tirage bernoullien de zéro ou une unité selon une probabilité de tirage correspondant à la probabilité de réponse du logement, et tirages indépendants entre les logements. On peut donc appliquer la formule de Rao pour prendre en compte ce degré de sondage supplémentaire, par rapport au plan de sondage initial à deux degrés. C'est l'approche qui a été retenue (cf. formule « e I.1.2 » page 11). Toutefois, on montre qu'omettre le terme de variance relatif à ce troisième degré de sondage[11] lié à la non-réponse conduit à un estimateur majorant de la variance des estimations transversales. Dans une optique conservatrice, cette simplification à été effectuée et conduit à la formule « e I.1.3 » page 11) ;

- le plan de sondage à deux degrés de l'échantillon trimestriel est pris en compte grâce à la formule de Rao :

  – l'estimateur retenu pour la variance relative au plan de sondage de premier degré est, au sein de chaque région, l'estimateur de Deville-Tillé suivant, qui

[11] Dans l'application récursive de la formule de Rao, il s'agit du second terme, de forme générique $\sum_{i \in S_I} \left(\frac{1}{\pi_i^2} - q_i \right) \hat{V}(\hat{T}_i)$.

fournit une approximation de la variance de l'estimateur d'Horvitz-Thompson dans le cadre d'un sondage à probabilités inégales π_k de taille n équilibré sur q variables :

$$\hat{V}_{DT}(\hat{Y}_\pi) = \sum_{k \in S} c_k (z_k - \tilde{z}_k)^2$$

où $z_k = \frac{y_k}{\pi_k}$, $\tilde{z}_k$ est le prédicteur de z_k par la régression, pondérée par les poids de sondage, sur les variables d'équilibrage, et c_k une pondération. Plusieurs pondérations ont été testées, et sont détaillées au §I.2. La pondération finalement retenue, pour des raisons de simplicité et au vu de l'impact relativement mesuré du choix de ce paramètre, est la pondération de Hajek $c_k = (1 - \pi_k)\frac{n}{n-q}$;

- le second degré de tirage étant assimilé à un SAS de une grappe parmi six ou sept, une estimation directe de la variance de second degré au sein de chaque UP est impossible. Pour pallier cette difficulté, une procédure de « collapse[12] » – regroupement de deux ou trois unités primaires permettant ensuite une estimation de la variance de second degré au sein des unités primaires ainsi regroupées – a été mise en œuvre et conduit à un estimateur majorant de la variance de second degré (second terme de la formule « e I.1.1 »). Différentes stratégies de regroupement des unités primaires ont été testées, et sont détaillées au §I.3. Le choix de la stratégie de collapse s'avère en pratique sans incidence notable sur la qualité des estimations de précision.

2.4. Grandes lignes de la méthode d'estimation de variance longitudinale

On s'intéresse ici à l'estimation de la précision d'estimations faisant intervenir plusieurs échantillons trimestriels distincts de l'EEC. En pratique, on s'intéresse principalement aux estimations de précisions relatives à des moyennes annuelles ou à des variations d'un trimestre à l'autre. Dans tous les cas, l'estimation de la précision de telles quantité se décompose en calculs de variances transversales d'une part et de covariances d'autre part (cf. §II.1). Par exemple, la variance d'une variation d'un trimestre à l'autre pour une variable Y est égale à :

$$V(\hat{Y}_{w,t} - \hat{Y}_{w,t-1}) = V(\hat{Y}_{w,t}) + V(\hat{Y}_{w,t-1}) - 2\,\mathrm{cov}(\hat{Y}_{w,t}, \hat{Y}_{w,t-1})$$

Les variances transversales sont estimées au moyen de la méthode précédemment exposée. Pour l'estimation des covariances entre estimateurs relatifs à différents trimestres, une approche de type « Ultimate Cluster » a été retenue : seules[13] les covariances liées au premier degré du plan de sondage sont prises en compte, au moyen d'une formule s'inspirant de l'estimateur de variance de Deville-Tillé utilisé pour calculer la variance de premier degré. Ce point est détaillé au §II.2.

---

[12] Cette technique de collapse permettant l'estimation de variance dans le cas d'un plan de sondage aléatoire simple stratifié avec sélection d'une unité par strate découle d'un cas particulier d'application de la formule de Rao (cf. annexe A.2 pour plus de détails). Elle correspond également à l'approche décrite par Cochran dans [5], §5A.12.

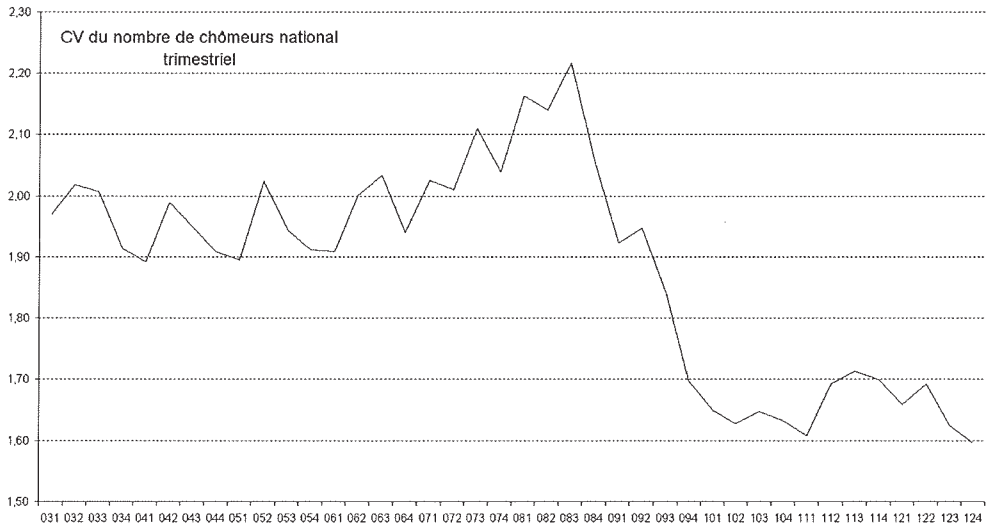
[13] Le calage reste toutefois pris en compte, puisque les formules sont toujours appliquées aux résidus des régressions des variables d'intérêt sur les variables de calage.

3. Principaux résultats[14].

3.1. Précision transversale

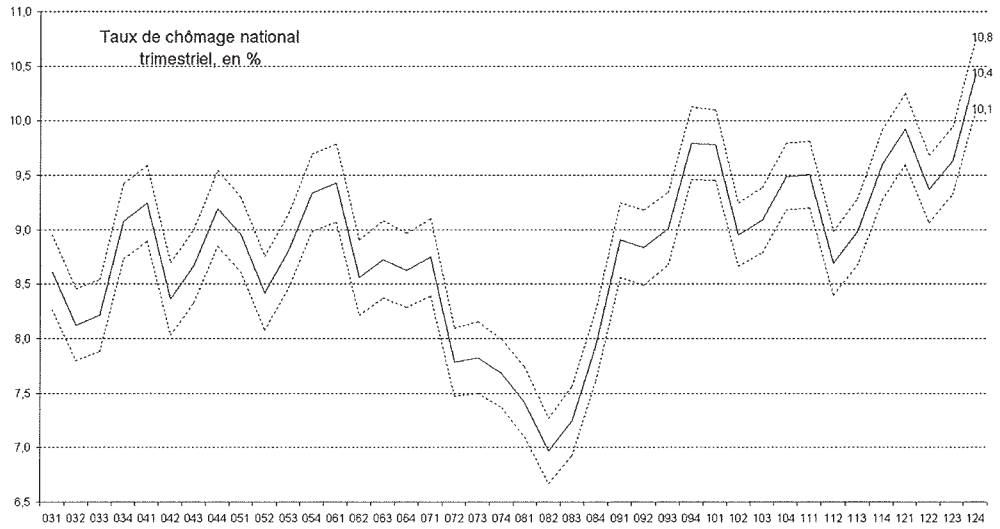
Le graphique 1 donne l'évolution du coefficient de variation de l'estimation trimestrielle du nombre de chômeurs issue de l'EEC, sur la période s'étalant du 1er trimestre 2003 au dernier trimestre 2012. On constate très clairement un gain de précision lié à l'augmentation de 50 % de la taille de l'échantillon entre fin 2008 et fin 2011. Sur l'année 2012, les coefficients de variation des estimations trimestrielles de nombre de chômeurs évoluent entre 1,6 % et 1,7 %.

Graphique 1 : coefficient de variation de l'estimation trimestrielle du nombre de chômeurs, en %



Autre représentation de la précision transversale de l'EEC, le graphique 2 représente le taux de chômage trimestriel mesuré par l'enquête sur la période 2003-2012 et l'intervalle de confiance à 95 % associé.

Graphique 2 : taux de chômage trimestriel et intervalle de confiance associé

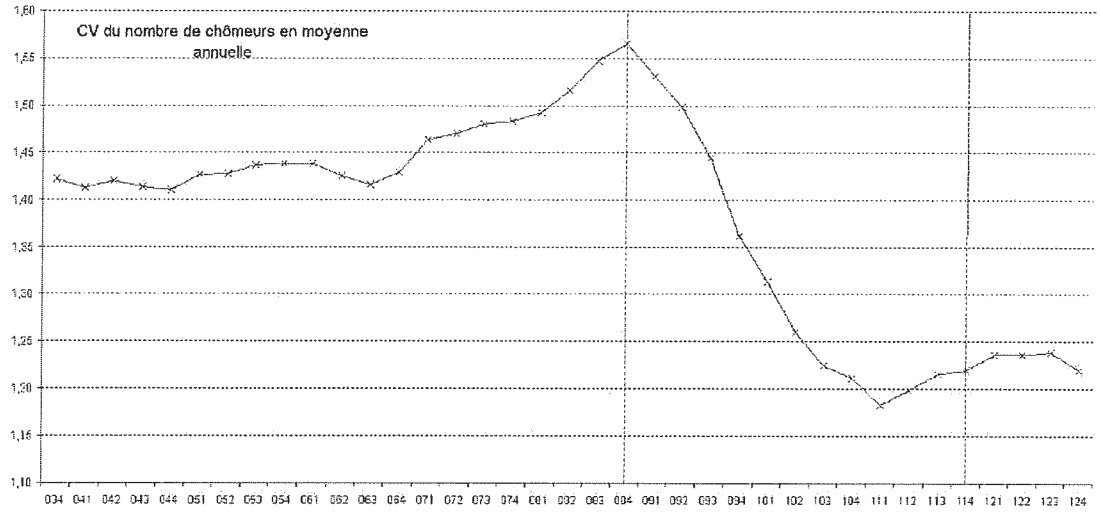


[14] Pour mémoire, ces résultats constituent des résultats bruts issus d'estimations directes réalisées de manière *ad hoc* pour les calculs de variance. Ils ne correspondent donc pas aux estimations officielles diffusées par l'Insee.

3.2. Précision longitudinale

Le graphique 3 donne l'évolution du coefficient de variation de la moyenne annuelle nationale du nombre de chômeurs, toujours sur la période 2003-2012. On retrouve tout aussi clairement que dans le cas des estimations trimestrielles le gain de précision lié à l'augmentation de 50 % de la taille de l'échantillon opérée entre fin 2008 et fin 2011.

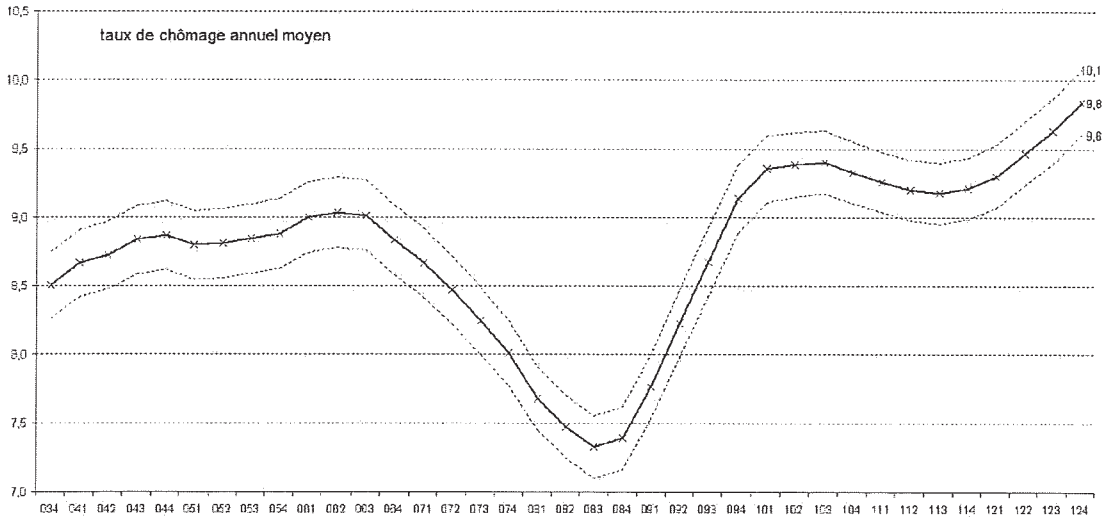
Graphique 3 : CV de la moyenne annuelle nationale du nombre de chômeurs, en %



Note : Au dénominateur, la moyenne annuelle d'un trimestre t est calculée par la somme des estimations des trimestres t-3 à t divisée par 4.

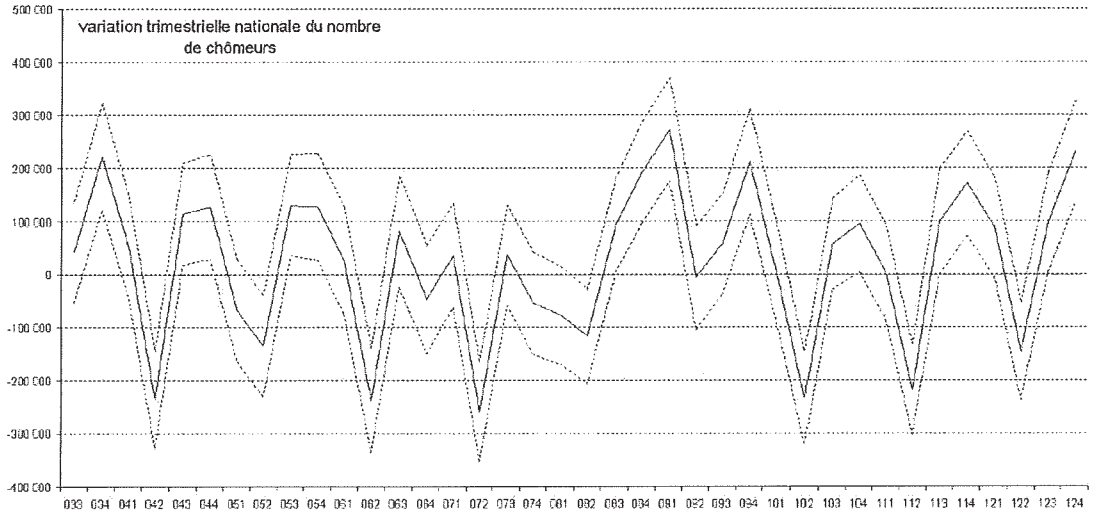
Le graphique 4 fournit une autre représentation de la précision de la moyenne annuelle, en représentant le taux de chômage moyen annuel mesuré par l'enquête sur la période 2003-2012 avec son intervalle de confiance à 95 %.

Graphique 4 : intervalle de confiance du taux de chômage annuel moyen, en %



Enfin, le graphique 5 décrit l'évolution par trimestre de l'intervalle de confiance à 95% pour la variation trimestrielle annuelle du nombre de chômeurs. On constate qu'une variation nationale de 100 000 chômeurs est à la limite de la significativité.

Graphique 5 : Intervalle de confiance de la variation trimestrielle nationale du nombre de chômeurs



Dernier point relatif à la précision longitudinale – détaillé au paragraphes II.4 et II.5 –, les calculs de précisions menés dans le cadre de ce document de travail ont permis de vérifier que les estimations issues de l'EEC satisfont systématiquement l'objectif d'Eurostat en termes de précision des variations trimestrielles nationales[15], et très fréquemment l'objectif d'Eurostat en termes de précision des moyennes annuelles régionales[16] – seules quelques régions ne satisfont pas chaque trimestre cette contrainte.

Bibliographie

[1] Loonis V., « La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation», *Actes des 10èmes Journées de Méthodologie Statistique*, mars 2009, Insee.

[2] Chauvet G., « Échantillonnage Équilibré Stratifié », *Série des documents de travail du CREST*, N° 2008-08.

[3] Juillard M., « Calculs des pondérations dans l'enquête Emploi en continu (France métropolitaine) : description de l'existant en date du 1er novembre 2012 », *note N° 2396/DG75-F230 du 15 novembre 2012*, Insee.

[4] Caron N., Deville J.C. et Sautory O., « Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE », *document de travail Insee M9806*, 1998, Insee.

[5] Cochran W., « Sampling Techniques, 3rd edition », *Wiley*, 1977.

[15] Pour une sous-population de 5 % de la population active (d'un pays de plus de 20 millions d'habitants), l'écart-type de la variation trimestrielle ne doit pas dépasser 2 % de la sous-population.

[16] Pour un groupe de chômeurs représentant 5 % de la population active, l'écart type de la moyenne annuelle ne doit pas dépasser 8 % de la sous-population. Les régions de moins de 300 000 habitants sont exclues du champ de cette contrainte.

Partie I
Méthode d'estimation de la précision transversale

La première section de cette partie présente la méthode d'estimation de la variance transversale de l'enquête Emploi, c'est-à-dire celle qui exploite une seule enquête trimestrielle. Les propriétés de cette estimation sont étudiées sur la période 2003T1-2012T4 dans la deuxième section. Le plan de sondage de l'enquête Emploi tirée dans les fichiers de la Taxe d'Habitation (TH) est décrit dans [VL].

I.1 principe général de l'estimation

L'estimateur de variance est défini sous forme analytique, c'est-à-dire par une fonction explicite des valeurs de la variable d'intérêt observées sur l'échantillon.

– Dans la suite, $\widehat{Y}^n$ désigne l'estimateur d'Horvitz-Thompson du total de la variable d'intérêt y en fonction de l'échantillon de n-ième phase, noté s_n. π_n est la probabilité d'inclusion à l'ordre n : $\pi_n(i_n) = proba(i_n \in s_n)$, $\pi_n(i_n, j_n) = proba\{\{i_n, j_n\} \subset s_n\}$ et la probabilité d'inclusion conditionnelle est notée $\pi_{n|m}(i_n) = proba(i_n \in s_n | s_m)$. Q_n désigne la forme quadratique choisie pour estimer la variance de n-ième phase sur l'échantillon de n-ième phase : l'estimateur de $Var\left\{\widehat{Y}^n\right\} = V_n$ est de la forme $\widehat{V}_n^{\,n} = Q_n[y_n(s_n)]$, où y_n est le total de y sur une unité de n-ième phase. Enfin, pour $m > n$, $\widehat{y}_n^{\,m}(i_n)$ désigne une estimation de $y_n(i_n)$ sur l'échantillon $s_m(i_n)$.

– Pour l'enquête Emploi, la première phase est celle du tirage des secteurs [1]. Le choix de la forme quadratique Q_1, qui estime donc la variance de première phase, est étudié plus loin.

– L'estimation sur l'échantillon s_2 des grappes de logements (unités de deuxième phase) approxime la deuxième phase du tirage par un sondage stratifié par unité primaire, donc l'assimile à un deuxième degré d'échantillonnage. Cette approximation permet de décomposer la variance ainsi :

$$Var\left[\widehat{Y}^2\right] = V_2 = Var\left[\widehat{Y}^1\right] + E\left\{Var\left[\widehat{Y}^2\middle|s_1\right]\right\}$$

$$\widehat{V}_2^{\,2} = \widehat{V}_1^{\,2} + \widehat{Var\left[\widehat{Y}^2\middle|s_1\right]}$$

$$= Q_1\left(\widehat{y}_1^{\,2}(s_1)\right) + \sum_{s_1}\left(\frac{1}{\pi_1^2} - q_1\right)\widehat{Var\left(\widehat{y}_1^{\,2}\middle|s_1\right)} \qquad \text{où } q_1 \text{ désigne le coefficient diagonal de } Q_1$$

(car $E\left\{Q_1\left[\widehat{y}_1^{\,2}(s_1)\middle|s_1\right]\right\} = Q_1(y_1) + \sum_{s_1} q_1 Var\left(\widehat{y}_1^{\,2}\middle|s_1\right)$, cf annexe A.1)

De plus, comme l'échantillonnage EEC tire à probabilités égales une seule unité secondaire par unité primaire, la loi conditionnelle de $s_2(i_1)$ est uniforme sur i_1 et donc :

$$Var\left(\widehat{y}_1^{\,2}(i_1)\middle|s_1\right) = Var\left\{\frac{y_2[s_2(i_1)]}{\pi_{2|1}}\middle|s_1\right\} = \frac{1}{\pi_{2|1}^{\,2}} Var_{i_1}(y_2) = \frac{1}{\pi_{2|1}^{\,2}}\frac{1}{|i_1|}\sum_{i_2 \in i_1}\left(y_2(i_2) - \frac{\sum_{i_2 \in i_1} y_2}{|i_1|}\right)^2$$

(où $|i_1|$ désigne le cardinal en unités secondaires de l'unité primaire i_1)

1. Plus précisément, la première phase de l'enquête Emploi tire des grappes de secteurs, de 1 à 3 secteurs. Le terme de 'grappe de secteurs' est utilisé ici de préférence à celui de bisecteur [VL].

→ En notant $\widetilde{i_1}$ l'élément générique d'une partition formée des collapses (cf annexe A.2) de deux ou trois unités primaires et $x^+ = \max(x, 0)$, un estimateur majorant la variance (non conditionnelle) de deuxième phase peut être construit ainsi [2] :

$$\widehat{V}_2^{\,2} = Q_1\left(\widehat{y_1}^{\,2}(s_1)\right) + \sum_{\widetilde{i_1}} \left|\widetilde{i_1}\right| \widetilde{\wp}^2_{\widetilde{i_1}} \left\{\sqrt{(1-q_1\pi_1^2)^+}\,\frac{y_2}{\pi_2}\right\} \qquad \textbf{(e I.1.1)}$$

$$= Q_2\left[y_2(s_2)\right]$$

– estimation sur l'échantillon s_3 des logements répondants [3] : Le tirage intra-grappe (pour les logements neufs) et la réponse des logements enquêtés sont considérés comme constituant un processus indépendant entre les grappes conditionnellement à la deuxième phase. Par conséquent :

$$V_3 = Var\left[\widehat{Y}^2\right] + E\left\{Var\left[\widehat{Y}^3 \middle| s_2\right]\right\}$$

$$\widehat{V}_3^{\,2} = \widehat{V}_2^{\,2} + Var\left[\widehat{Y}^3 \middle| s_2\right]$$

$$= Q_2\left[y_2(s_2)\right] + \sum_{s_2} \frac{1}{\pi_2^2} Var\left(\widehat{y_2}^{\,3} \middle| s_2\right)$$

$$\widehat{V}_3^{\,3} = Q_2\left(\widehat{y_2}^{\,3}(s_2)\right) + \sum_{s_2} \left(\frac{1}{\pi_2^2} - \overbrace{q_2}^{3}\right) Var\left(\widehat{y_2}^{\,3} \middle| s_2\right) \qquad \textbf{(e I.1.2)}$$

→ Si $q_1\pi_1^2 \leqslant 1$ alors le coefficient diagonal de la forme quadratique de deuxième phase vaut, pour $i_2 \in i_1 \subset \widetilde{i_1}$:

$$q_2 = q_1 \frac{1}{\pi_{2|1}^{\,2}} + \left|\widetilde{i_1}\right| \frac{1}{\left|\widetilde{i_1}\right|} \frac{1-q_1\pi_1^2}{\pi_2^2} = \frac{1}{\pi_2^2}$$

Donc sous cette hypothèse ($q_1\pi_1^2 \leqslant 1$) le deuxième terme de (**e I.1.2**) est nul. Dans le cas contraire, $q_1\pi_1^2 > 1$ et alors $q_2 = q_1 \dfrac{1}{\pi_{2|1}^{\,2}} > \dfrac{1}{\pi_2^2}$ et le deuxième terme de (**e I.1.2**) est négatif. Au final, dans les deux cas (**e I.1.3**) donne un estimateur majorant, fonction de l'échantillon des logements répondants.

$$\boxed{\widehat{V}_3^{\,3} = Q_2\left(\widehat{y_2}^{\,3}(s_2)\right) = Q_1\left(\widehat{y_1}^{\,3}(s_1)\right) + \sum_{\widetilde{i_1}} \left|\widetilde{i_1}\right| \widetilde{\wp}^2_{\widetilde{i_1}} \left\{\sqrt{(1-q_1\pi_1^2)^+}\,\frac{\widehat{y_2}^{\,3}}{\pi_2}\right\}} \qquad \textbf{(e I.1.3)}$$

remarques :

- Pour l'estimateur Horvitz-Thompson de la variance, $q_1 = \dfrac{1-\pi_1}{\pi_1^2} < \dfrac{1}{\pi_1^2}$.
- Si $Q_1 \geqslant Q_1^*$, la forme quadratique de l'estimateur Horvitz-Thompson de la variance, alors $q_1\pi_1^2$ peut être remplacé par $1 - \pi_1$ dans (**e I.1.3**), avec un effet majorant.

– Pour le calcul de $\widehat{y_2}^{\,3}$, la probabilité de réponse est estimée par le rapport entre le poids de tirage du logement et son poids final, c'est-à-dire par l'inverse du facteur de calage [4].

- Dans le cadre de la formulation (**e I.1.3**), l'enjeu de la méthode d'estimation est de choisir la forme quadratique de première phase et le mode de collapse des unités primaires.

2. L'annulation des coefficients $1 - q_1\pi_1^2 < 0$ vise à éviter le risque que les collapses ne mènent à sous-estimer la variance. L'impact de cette troncature s'avère sensible.

3. Les logements hors champ (non principaux) doivent être considérés comme répondants, avec une valeur nulle pour la variable d'intérêt.

4. Le calage de l'enquête Emploi traite la non-réponse.

• Usuellement, le calage de l'estimateur $\widehat{Y}$ est pris en compte dans l'estimateur de variance en remplaçant la variable d'intérêt par le résidu de sa régression sur les variables de calage. Ici, cette régression est à effectuer au niveau logement.

I.2 estimation de la variance de première phase

– L'option retenue pour l'estimateur de première phase est l'approximation de Deville et Tillé [DT]. Elle correspond à la forme quadratique positive (**e I.2.4**), où $()^-$ désigne l'inverse généralisée.

$$\widehat{V}^{DT} = Q\left[y\left(s\right)\right] = \min_{\beta} \sum_s c\left(\frac{y}{\pi} - \frac{x'}{\pi}\beta\right)^2 = \sum_s c\left(\frac{y}{\pi}\right)^2 - \sum_s c\frac{yx'}{\pi^2}\left[\sum_s c\frac{xx'}{\pi^2}\right]^- \sum_s c\frac{yx}{\pi^2} \qquad \textbf{(e I.2.4)}$$

Cette dernière expression montre que son coefficient diagonal est :

$$q = \frac{c}{\pi^2}\left(1 - \frac{c}{\pi^2}x'\left[\sum_s c\frac{xx'}{\pi^2}\right]^- x\right) \quad \left(\in [0, \frac{c}{\pi^2}[\right)$$

Il est parfois utile de reformuler la forme quadratique (**e I.2.4**) $Q\left[y\left(s\right)\right] = \Delta\left[\frac{y}{\pi}(s)\right]$. Le coefficient diagonal de la forme quadratique Δ est noté $\delta\left(c\right) = c\left(1 - \frac{c}{\pi^2}x'\left[\sum_s c\frac{xx'}{\pi^2}\right]^- x\right)$.

– L'estimateur de variance (**e I.2.4**) peut être calculé simplement par régression de la variable d'intérêt par les variables d'équilibrage, pondérée par c/π^2.

– Il ne prend pas en compte la variance d'atterrissage de l'échantillonnage TH. Donc il sous-estime vraisemblablement la variance, sous réserve de l'effet de la pondération c. Une étude a évalué une estimation alternative de la variance basée sur des probabilités d'inclusion double des grappes de secteurs approximées par réplication de l'échantillonnage, selon la méthode de Breidt et Chauvet [BC]. Ceci modifierait le calcul de la forme quadratique Q_1 qui intervient dans la formule (**e I.1.3**). La prise en compte par cette méthode de l'équilibrage du tirage des grappes de logements est une autre éventualité à explorer, si la difficulté liée au tirage d'une seule unité secondaire par unité primaire peut être contournée.

– choix de la pondération des résidus d'équilibrage :

- la pondération 'de Hajek' (ou de Deville [AMYT]) (c_h) : $c_h = (1-\pi)\frac{|s|}{|s|-|x|}$, où $|s|$ est la taille d'échantillon et $|x|$ la dimension d'équilibrage.
- pondération calée sur la trace de l'estimateur sans biais (c_2) : L'estimateur sans biais d'Horvitz-Thompson de la variance est donné par la forme quadratique $\sum_{i,j\in s} \frac{y}{\pi}(i)\frac{y}{\pi}(j)\frac{\pi\left(i,j\right)-\pi\left(i\right)\pi\left(j\right)}{\pi\left(i,j\right)}$. Son coefficient diagonal, comme fonction de y/π, est $1-\pi$. Il en découle la pondération calée sur la trace de cet estimateur :

$$c_2 = (1-\pi)\sum_s(1-\pi) \Big/ \sum_s \delta\,(1-\pi)$$

– Le programme d'estimation de la variance transversale EEC est également paramétré de façon à remplacer le coefficient $q_1\pi_1^2$ dans (**e I.1.3**) par $\mu_1 \in \{1-\pi_1, 1, \delta\left(c_1\right)\}$.

- Le choix $\mu_1 = \delta\left(c_1\right)$ $(= q_1\pi_1^2)$ est préconisé, parce qu'il devrait limiter le biais de l'estimateur.
- $\mu_1 = 1 - \pi_1$ assimile Q_1 à l'estimateur de variance d'Horvitz-Thompson.

- $\mu_1 = 1$ supprime le deuxième terme de (**e I.1.3**). C'est une option de type 'ultimate cluster'.

trois remarques sur la pondération de Hajek :

– S'il existait une pondération c^* telle que Q_{c^*} coincide avec l'estimateur de variance d'Horvitz-Thompson, le coefficient diagonal vérifierait $\delta(c^*) = 1-\pi$. Comme $\delta(1-\pi) = (1-\pi)l$ avec $l = 1-\frac{1-\pi}{\pi^2}x'\left\{\sum_s (1-\pi)\frac{xx'}{\pi^2}\right\}^- x$, une option possible est de 'corriger' le poids initial $1-\pi$ par l'inverse de $\bar{l}^s$ [5]. Or cette moyenne vaut $(|s| - |x|) / |s|$. En effet, pour toute pondération de l'échantillon c_s :

$$\sum_s 1 - \frac{c}{\pi^2}x'\left[\sum_s c\frac{xx'}{\pi^2}\right]^- x = |s| - \text{trace}\left\{\left[\sum_s c\frac{xx'}{\pi^2}\right]^- \sum_s c\frac{xx'}{\pi^2}\right\} = |s| - \text{rang}(x_s) \quad [6]$$

(L'indiçage est implicite dans cette formulation.)

– Avec une seule contrainte d'équilibrage et pour un sondage aléatoire simple, la pondération de Hajek donne l'estimateur de variance d'Horvitz-Thompson ($\frac{N^2\left(1-\frac{|s|}{N}\right)}{|s|}\frac{1}{|s|-1}\sum_s\left(y - \frac{\sum y(s)}{|s|}\right)^2$).

– Cependant, en théorie l'estimateur de variance ainsi pondéré parait susceptible de sous-estimer la variance, même sans phase d'atterrissage. En effet, sa trace peut s'écrire (avec $u = x / \pi$) :

$$\sum_s \delta(c) = \sum_s c - \sum_s c^2 u'\left(\sum_s cuu'\right)^- u = \sum_s c - |s|\,Cov_s\left\{c, cu'\left(\sum_s cuu'\right)^- u\right\} - \sum_s c\sum_s cu'\left(\sum_s cuu'\right)^- u/|s|$$

$$= \sum_s c\frac{|s|-|x|}{|s|} - |s|\,Cov_s\left\{c, cu'\left(\sum_s cuu'\right)^- u\right\}$$

Or la covariance 'empirique' entre c et $cu'\left(\sum_s cuu'\right)^- u$ pourrait être supposée > 0. C'est le cas notamment lorsque $u = 1$ (une seule contrainte de taille d'échantillon fixe) et que c n'est pas constant sur l'échantillon, et plus généralement lorsque $(c, u'\left(\sum_s cuu'\right)^- u)$ est indépendant. C'est également vrai pour le calage sur la population totale $x = 1$, car $Cov_s\left(1-\pi, \frac{1-\pi}{\pi^2}\right) > 0$, vu que ce sont deux fonctions décroissantes de $\pi \in]0,1[$ [7]. Cette hypothèse $Cov_s\left\{c, cu'\left(\sum_s cuu'\right)^- u\right\} > 0$ entrainerait une tendance à sous-estimer la variance.

- En pratique, la trace de l'estimateur de la pondération de Hajek sur l'échantillon effectif des grappes de secteurs excède la trace d'Horvitz-Thompson dans toutes les régions hors Corse. Ce résultat, cohérent avec le fait que la pondération de Hajek donne des estimations de variance en général plus élevées que les pondérations calées sur la trace de l'estimateur sans biais (cf graphique suivant et Tableau B.2 en annexe), suggère que l'hypothèse n'est pas vérifiée.

---

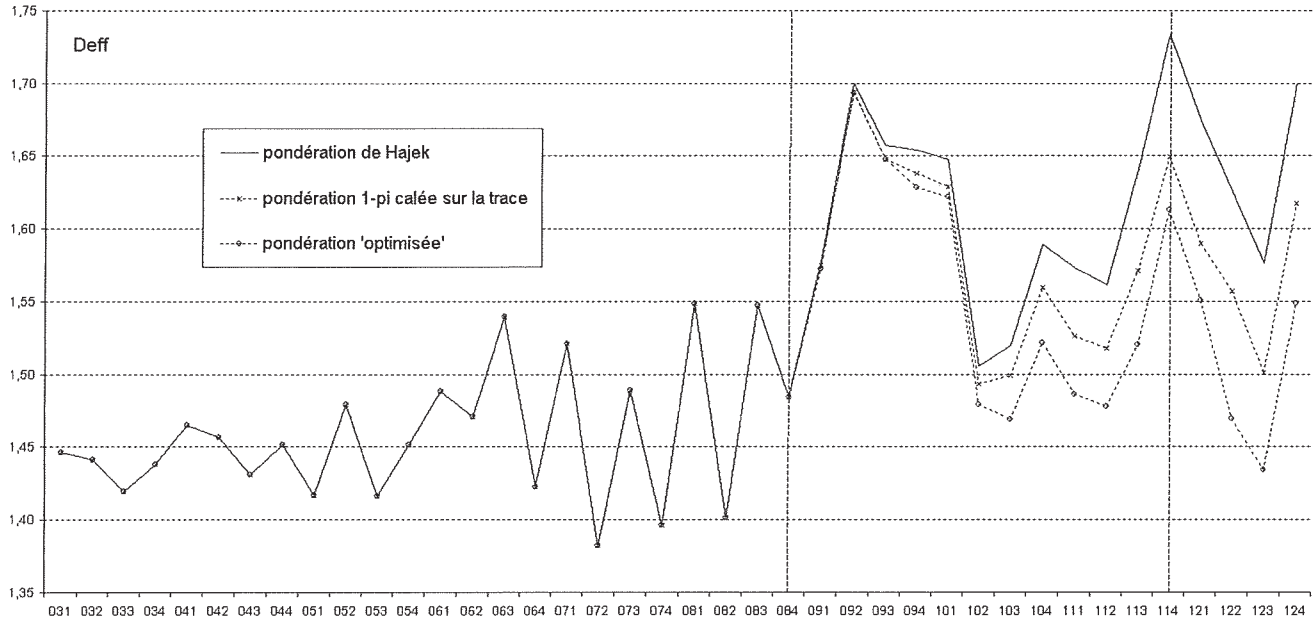
5. $\bar{l}^s$ est la moyenne sur l'échantillon du coefficient multiplicateur l. On remarque cependant que $\overline{\left(\frac{1}{l}\right)}^s \geqslant \frac{1}{\bar{l}^s}$.

6. Si A est une matrice symétrique alors $\text{trace}(AA^-) = \text{rang}(A)$. Soient $\mathscr{U}$ et $\mathscr{V}$ deux bases orthonormées de $\text{Im}(A)$ et de $\text{Ker}(A)$ alors $\text{trace}(AA^-) = \sum_{u\in\mathscr{U}} u'AA^-u + \sum_{v\in\mathscr{V}} v'AA^-v$, qui est égal à $\sum_{u\in\mathscr{U}} u'AA^-u = \sum_{u\in\mathscr{U}} u'u = \text{rang}(A)$, car $AA^-A = A \iff AA^- = \text{id sur Im}(A)$.

7. Si f et g sont deux fonctions décroissantes alors $Cov_p(f,g) = \iint (f(x)-g(y))(f(x)-g(y))dp(x)\,dp(y) \geqslant 0$ (inégalité de Tchebychev).

• L'enjeu de la pondération des résidus d'équilibrage n'est pas négligeable : jusqu'à 5% du Deff du taux de chômage (Graphique 1).

Graphique 1 – Impact de la pondération des résidus d'équilibrage sur le Deff estimé du taux de chômage



Note :

– *Les deux traits verticaux délimitent la période de transition de l'échantillonnage, pour le passage à la TH.*

– *La pondération 'optimisée' est obtenue en minimisant* $\sum_s \{\delta(c) - (1-\pi)\}^2$ *par récursion sur* $c - \delta(c) + 1 - \pi$ *puis en calant la trace* $\sum_s \delta$ *sur* $\sum_s 1 - \pi$.

• La simplicité du calcul de la pondération de Hajek, les résultats comparables à ceux des autres pondérations et surtout son usage courant (notamment dans [AMYT]) mènent à retenir cette pondération pour le calcul de variance de l'enquête Emploi.

I.3 stratégie de collapse des unités primaires

- La stratégie du collapse des unités primaires (secteurs TH) vise à déterminer une partition $\widetilde{s_1}$ de s_1, l'échantillon de secteurs, telle que pour $\widetilde{i_1} \in \widetilde{s_1}$ $2 \leqslant \left|\widetilde{i_1}\right| <= 3$ et $E\left\{\sum_{\widetilde{i_1}} \left|\widetilde{i_1}\right| \widetilde{\mathcal{S}}^2_{\widetilde{i_1}} \left\{\sqrt{(1-q_1\pi_1^2)^+}\frac{y_2}{\pi_2}\right\} \middle| s_1\right\}$ soit supérieur à $\sum_{i_1 \in s_1} (1-q_1\pi_1^2)\frac{1}{\pi_2^2} Var_{i_1}(y_2)$ tout en étant le plus petit possible. Cependant cet objectif idéal n'est pas opérationnel, en raison d'une part de la multiplicité des variables d'intérêt y de l'enquête, d'autre part du fait que seul le résidu de la régression de calage importe pour la variance de l'estimateur calé.
- tri préalable 'optimal' :

  Les secteurs [8] sont fusionnés dans l'ordre du tri par la variable $\sqrt{(1-\mu_1)^+}\frac{1}{\pi_2}$, où :

  ▷ $\mu_1^{TH} = c\left(1 - \frac{c}{\pi^2}x'\left[\sum_s c\frac{xx'}{\pi^2}\right]^- x\right) (= \delta(c))$

  ▷ $\frac{1}{\pi_2}$ est le poids de tirage de la grappe de logements (calculé par `wgrp/pbisec_i`)

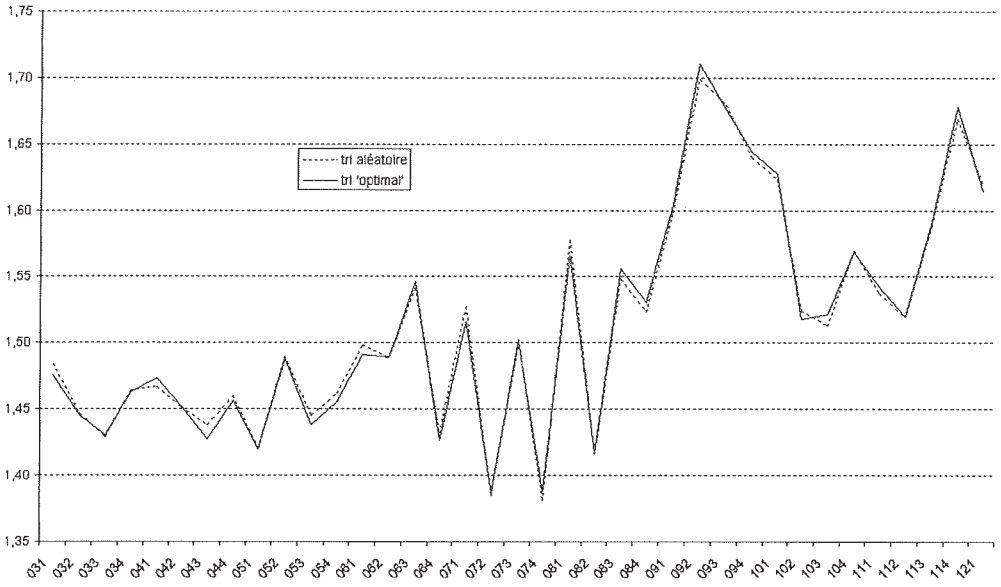
  Le regroupement est effectué, dans l'ordre de tri, par couples successifs ou par le triplet final en cas de nombre impair d'unités primaires dans la strate de collapse.

NB : Sur toute la période, l'ordre de grandeur du terme Q_1 de l'équation (**e I.1.3**) s'avère 10 fois plus grand que le terme $\mathcal{S}^2$, ce qui limite l'incidence de la stratégie de collapse.

8. et non pas les grappes de secteurs TH

→ L'incidence sur l'estimation de variance du tri 'optimal' préalable au collapse, comparé à un tri aléatoire, s'avère faible, et pas toujours à la baisse (Graphique 2).

Graphique 2 – Impact du tri préalable au collapse sur le Deff estimé du taux de chômage



Note : Les regroupements des unités primaires ont été effectués au sein des strates de tirage de première phase.

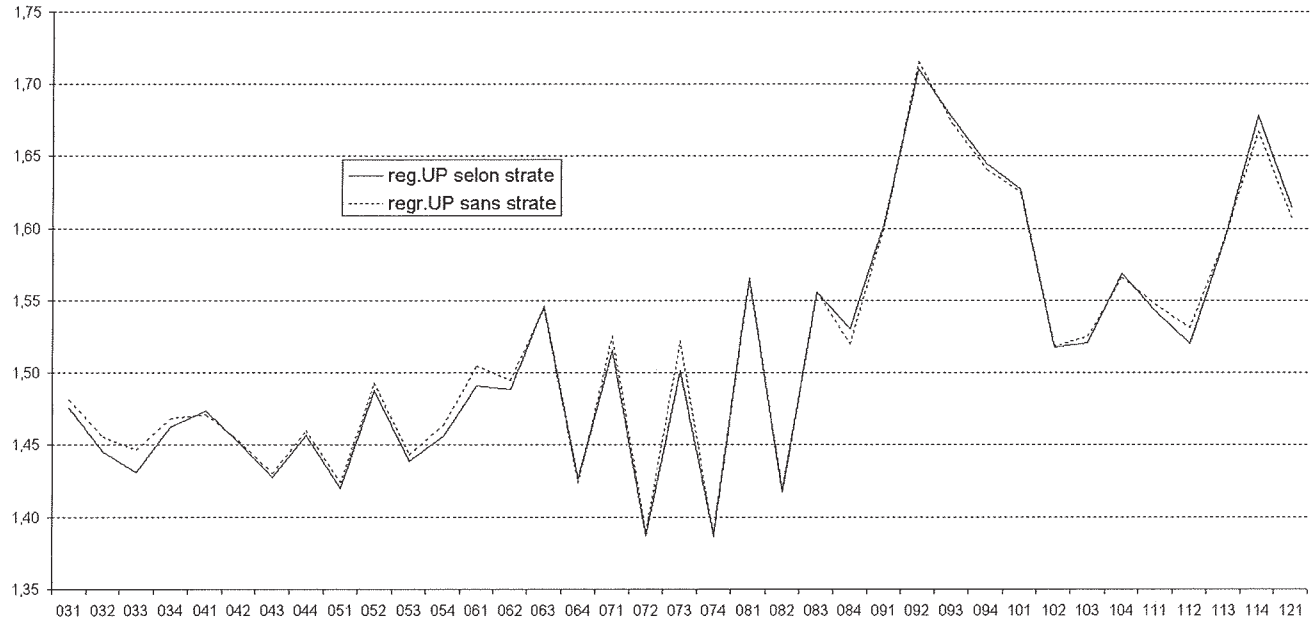
Le tri par la région préalablement au collapse et la suppression de la stratification donne une série de Deff peu différente et plutôt préférable (Graphique 3)

Graphique 3 – Comparaison du tri 'optimal' avec stratification au tri régional sans stratification

- stratification du collapse :

L'incidence de la stratification du collapse par les strates de première phase (région pour l'échantillon TH) est faible pour la pondération c_2 (Graphique 4), comme pour la pondération c_h. L'effet est en général favorable, c'est-à-dire qu'il limite la surestimation due au collapse.

Graphique 4 – Impact d'une stratification régionale du collapse sur l'estimation du Deff du taux de chômage



Note : Pour l'échantillon RP, le respect des strates de tirage induit des singletons, qui sont regroupés avec des couples d'unités primaires déjà agrégées.

→ Une stratification régionale du collapse des unités primaires parait sans effet sensible au niveau national, mais sans doute préférable pour les calculs de précision régionaux. C'est l'option choisie par défaut dans le programme livré.

I.4 quelques résultats de l'estimation de variance transversale

Cette partie fournit des éléments pour étudier la cohérence de la série des estimations de variance pour les variables 'Eurostat'. Les choix suivants sont appliqués :

- pondération de Hajek, pour les deux échantillonnages (RP et TH) ; $\mu_1 = \delta(c_h)$
- stratification régionale pour le regroupement des UPs

→ Le coefficient de variation du nombre de chômeurs baisse d'environ 2.0 avant 2008T4 à 1.7 en fin de période (Tableau B.1, en annexe). A Deff constant, l'augmentation de 50% de la taille de l'échantillon laissait espérer une baisse un peu plus prononcée, de 18%.

– Alors que le Deff du taux de chômage parait globalement stable sur la période d'estimation par l'échantillon RP, il subit une forte croissance au début de la période d'introduction de l'échantillon TH, et des fluctuations plus amples (Graphique 5).

Graphique 5 – estimation du Deff du taux de chômage par trimestre

Note :

– *formule d'estimation du Deff :* $\widehat{Deff} = \dfrac{\widehat{Var\left[\widehat{Y}\right]}}{\widehat{N}^2\left(1-\dfrac{|s|}{\widehat{N}}\right)\dfrac{1}{|s|}\left\{\dfrac{\widehat{\sum y^2}}{\widehat{N}-1}-\dfrac{\widehat{Y}^2-\widehat{Var\left[\widehat{Y}\right]}}{\widehat{N}(\widehat{N}-1)}\right\}}$

(y est en fait remplacé par le résidu de calage, pour fournir un Deff de l'estimateur calé.)

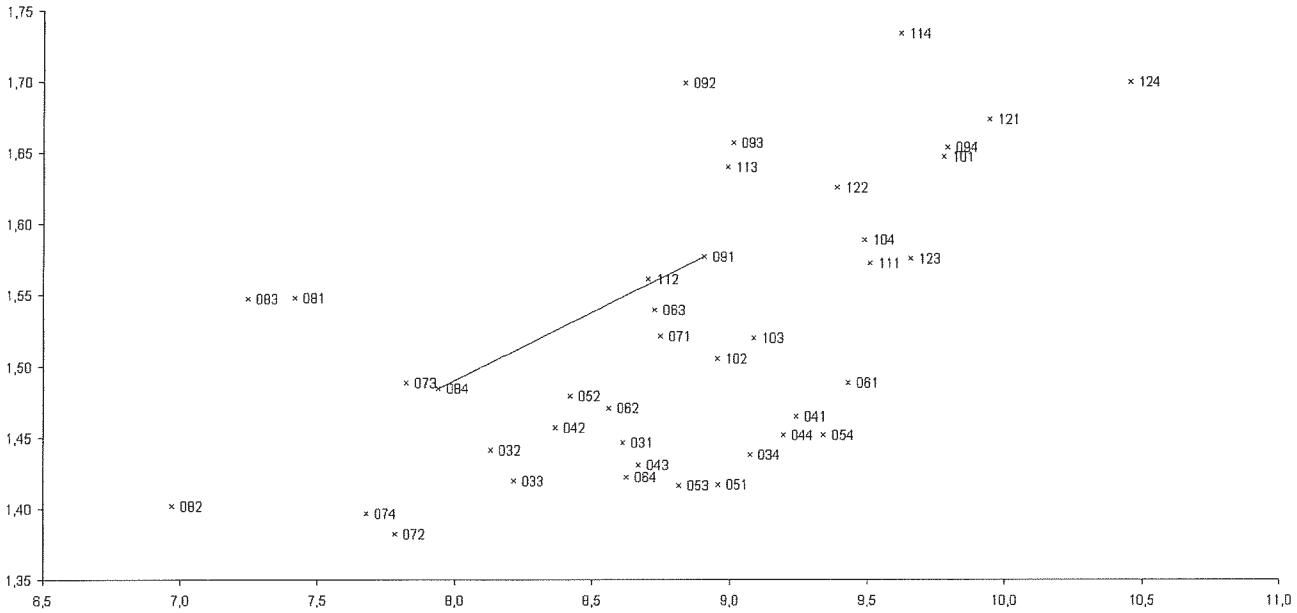
– *$\widehat{N}$ est calculé par cumul du poids final des logements répondants, y compris les hors champ. Comme les univers de logements des deux sources diffèrent, et que le partage des poids est limité à une partie des groupes de rotation, l'espérance de cet estimateur n'a pas un sens clair. Ce changement d'univers affecte aussi le Deff.*

– *L'estimation des deux derniers trimestres 2012 a été réalisée sur des données datées du 25 février 2013. Ces données évoluent avec la production des enquêtes, même pour les trimestres passés.*

– *Les fluctuations de haute fréquence entre 063 et 083 pourraient s'expliquer par la présence de points extrêmes. Ainsi il apparait un point très excentré sur le graphique de dispersion* $\widehat{y_1}(074) \times \widehat{y_1}(084)$.

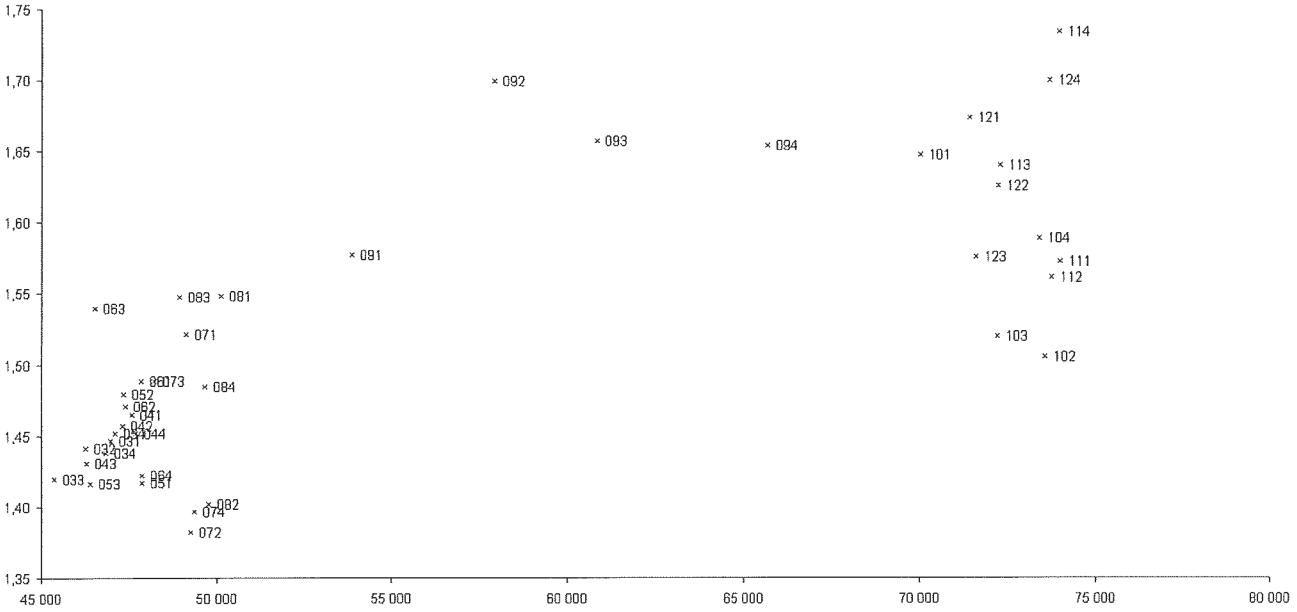
– Sur la période 'RP', il n'y a pas de corrélation entre le Deff et le niveau du taux de chômage (Graphique 6). Néanmoins, de fortes variations du chômage, comme entre 084 et 091, pourraient accroître la dispersion des résidus d'équilibrage et expliquer les Deff les plus élevés.

Graphique 6 – Deff du taux de chômage en fonction du niveau de celui-ci, selon le trimestre (aat)



– L'accroissement du nombre de logements répondants ne parait pas avoir causé une hausse 'technique' du Deff, vu l'absence de lien discernable sur les périodes 'mono-source' (Graphique 7).

Graphique 7 – Deff du taux de chômage en fonction du nombre de logements répondants



– En revanche, un calcul de la taille moyenne des grappes de logements indique que celle-ci a augmenté d'environ 22 à 27 entre les deux échantillons. (Le calcul a été effectué par somme du rapport entre le poids de tirage du logement et celui de la grappe de logements, pour chacune de celles-ci, puis moyenne simple [9].) Donc la hausse du Deff pourrait s'expliquer par une taille des grappes de logements TH plus grande que celle des aires de l'échantillonnage RP. Sous cette hypothèse, la méthode d'estimation de la variance ne serait pas en cause.

• Le Tableau B.2 en annexe permet de mesurer l'impact du choix des paramètres du programme : la pondération des résidus d'équilibrage et le coefficient μ_1 appliqué au deuxième terme de variance (intra-UP). Si le choix $\mu_1 = \delta(c_1)$ semble s'imposer par son absence de biais conditionnel (dans le cadre des approximations utilisées pour l'estimation de variance), il mène pour certaines variables à des surestimations très sensibles de la variance totale. D'autre part, l'utilisation d'une pondération optimisée pour minimiser la distance à $1 - \pi$ n'améliore pas uniformément la qualité de l'estimation de variance.

9. Ces résultats sont confirmés par un calcul alternatif :
   – La moyenne du nombre de logements TH sur la base de sondage des grappes de logements vaut 26.7.
   – Le rapport du nombre de logements RP 1999 sur le nombre d'aires de l'échantillon des secteurs RP vaut 20.2.

Partie II
Méthode d'estimation de la précision longitudinale et évaluation par rapport aux objectifs d'Eurostat

Cette deuxième partie décrit la méthode d'estimation de la variance longitudinale de l'enquête Emploi. L'exploitation longitudinale concerne les estimations en fonction d'au moins deux échantillons trimestriels. Il s'agit notamment des évolutions entre deux trimestres et des moyennes annuelles.

II.1 décomposition de la covariance

– L'estimation de la précision d'une moyenne annuelle ou d'une variation trimestrielle se décompose en calcul de variances transversales d'une part et de covariances d'autre part. En effet :

→ pour une moyenne annuelle (de totaux d'une variable d'intérêt y) :

$$Var\left(\widehat{\overline{Y}}_t\right) = Var\left\{\sum_{l=0}^{3}\widehat{Y}_{t-l} \Big/ 4\right\}$$

$$= \frac{1}{16}\left\{\sum_{l=0}^{3} Var\left(\widehat{Y}_{t-l}\right) + 2\sum_{\tilde{t}=t-2}^{t}\sum_{l=1}^{\tilde{t}-(t-3)} Cov\left(\widehat{Y}_{\tilde{t}}, \widehat{Y}_{\tilde{t}-l}\right)\right\}$$

→ pour une variation :

$$Var\left(\widehat{Y}_t - \widehat{Y}_{t-l}\right) = Var\left(\widehat{Y}_t\right) + Var\left(\widehat{Y}_{t-l}\right) - 2Cov\left(\widehat{Y}_t, \widehat{Y}_{t-l}\right)$$

– Pour l'EEC, l'échantillon de première phase, noté s_1 et composé de secteurs, est pérenne [10].

– L'échantillon de deuxième phase, noté s_2 et composé de grappes de logements TH, est renouvelé par sixième chaque trimestre. Chaque grappe sortante est remplacée par une grappe différente (disjointe) du même secteur géographique.

– Le conditionnement de la covariance par l'échantillon de première phase, pour l'estimateur d'Horvitz-Thompson (identique à l'estimateur par expansion pour le plan de sondage étudié ici) donne :

$$Cov\left(\widehat{Y}_{t-l}, \widehat{Y}_t\right) = C = Cov\left\{E\left(\widehat{Y}_{t-l}\middle|s_1\right), E\left(\widehat{Y}_t\middle|s_1\right)\right\} + E\left[Cov\left(\widehat{Y}_{t-l}, \widehat{Y}_t\middle|s_1\right)\right]$$

$$= Cov\left(\widehat{Y}_{t-l}^{1}, \widehat{Y}_t^{1}\right) + E\left[Cov\left(\widehat{Y}_{t-l}, \widehat{Y}_t\middle|s_1\right)\right] \qquad \textbf{(e II.1.5)}$$

(Dans la suite, $y_{1,t}$ désigne le total sur une unité de première phase de la variable y à la date t.)

II.2 estimation de la covariance de première phase

– Q_1 désigne la forme quadratique retenue pour estimer la variance transversale de première phase sur l'échantillon correspondant. Sa notation est confondue avec celle de la matrice correspondante :

$$\widehat{Var\left[\widehat{Y}^1\right]} = Q_1\left(y_1\left(s_1\right)\right) = y_1\left(s_1\right)' Q_1 y_1\left(s_1\right)$$

10. Ce n'est pas vrai sur la période de transition, avec la sortie de secteurs RP et l'entrée de secteurs TH.

– La covariance entre les estimateurs de première phase des dates $t-l$ et t s'exprime ainsi :

$$Cov\left\{\widehat{Y_{t-l}}^{1}, \widehat{Y_t}^{1}\right\} = C_1 = \frac{-Var\left\{\widehat{Y_t}^{1} - \widehat{Y_{t-l}}^{1}\right\} + Var\left[\widehat{Y_t}^{1}\right] + Var\left[\widehat{Y_{t-l}}^{1}\right]}{2}$$

$$\Longrightarrow \widehat{C_1}^{1} = \frac{-Q_1\left\{(y_{1,t} - y_{1,t-l})(s_1)\right\} + Q_1\left[y_{1,t}(s_1)\right] + Q_1\left[y_{1,t-l}(s_1)\right]}{2} \qquad \textbf{(e II.2.6)}$$

$$= y_{1,t-l}(s_1)'Q_1 y_1(s_1) \qquad \textbf{(e II.2.7)}$$

• La validité de cette expression de l'estimation de la covariance de première phase repose sur la pérennité de l'échantillon d'unités primaires.

• Le choix de l'approximation de Deville-Tillé donne la forme quadratique qui estime la variance de première phase :

$$\widehat{V}_1^{1} = Q_1(y_1(s_1)) = \left\| \frac{y_1}{\pi_1} - \frac{x_1'}{\pi_1}\widehat{\beta}_1 \right\|_{c_1}^2$$

(La pondération c_1 retenue est celle de Hajek : $c_1 = \dfrac{|s_1|}{|s_1| - |x_1|}(1 - \pi_1)$.)

Sa matrice s'exprime ainsi : $Q_1 = \left(\mathrm{id} - \mathrm{proj}^{\frac{c_1}{\pi_1^2}}_{\mathrm{Im}(x_1)}\right)' \mathrm{diag}\left(\dfrac{c_1}{\pi_1^2}\right)\left(\mathrm{id} - \mathrm{proj}^{\frac{c_1}{\pi_1^2}}_{\mathrm{Im}(x_1)}\right)$.

Par suite, (**e II.2.7**) peut se calculer par une somme pondérée de produits des résidus aux deux dates :

$$\widehat{\epsilon}_{1,t-l}'\mathrm{diag}\left\{c_1 / \pi_1^2\right\}\widehat{\epsilon}_{1,t} \qquad \textbf{(e II.2.8)}$$

où $\widehat{\epsilon}_{1,t}$ est le résidu de la régression de $y_{1,t}$ par les variables d'équilibrage de première phase (x_1). Cette régression est pondérée par c_1 / π_1^2 et stratifiée par région.

– De la décomposition de la covariance (**e II.1.5**) et avec le recours à l'hypothèse approximative d'indépendance conditionnelle des tirages par unité primaire, il découle que la covariance entre les estimateurs finaux (avant calage) peut s'estimer théoriquement sur l'échantillon de première phase par :

$$\widehat{C}^{1} = y_{1,t-l}(s_1)'Q_1 y_{1,t}(s_1) + \sum_{s_1} \frac{1}{\pi_1^2} Cov\left\{\widehat{y}_{1,t-l}, \widehat{y}_{1,t} | s_1\right\}$$

– En exploitant à nouveau l'indépendance approximative des échantillons finaux entre unités primaires, conditionnellement à l'échantillon de première phase, une expression 'théorique' d'un estimateur de la covariance sur l'échantillon final s'en déduit :

$$\widehat{C} = \widehat{y}_{1,t-l}(s_1)' Q_1 \widehat{y}_{1,t}(s_1) + \sum_{s_1}\left(\frac{1}{\pi_1^2} - q_1\right) \widehat{Cov}\left\{\widehat{y}_{1,t-l}, \widehat{y}_{1,t} | s_1\right\} \qquad \textbf{(e II.2.9)}$$

(q_1 désigne un élément générique de la diagonale de Q_1. Cette formule est générale : $\widehat{y}_{1,t}$ tient compte du tirage de troisième phase.)

$\Longrightarrow$ Par manque d'alternative clairement préférable (Annexe E et [JR]), l'option retenue est d'omettre le terme de covariance intra-UP. Avec ce choix, la formule d'estimation des covariances se simplifie en (**e II.2.10**). Cet estimateur est de type 'Ultimate Cluster' [11].

$$\widehat{Cov}\left(\widehat{Y}_{t-l}, \widehat{Y}_t\right) = \widehat{y}_{1,t-l}(s_1)' Q_1 \widehat{y}_{1,t}(s_1) \qquad \textbf{(e II.2.10)}$$

---

11. Ce terme fait référence à l'estimateur approché de la variance d'un sondage en deux phases sous la forme :

$$\widehat{Var}^{2}\left(\widehat{Y}^{2}\right) = Q_1\left(\widehat{y_1}^{2}(s_1)\right), \text{ où } Q_1\left\{y_1(s_1)\right\} \text{ estime } Var\left(\widehat{Y}^{1}\right)$$

remarque : Si l'estimateur de la variance transversale est du type :

$$\widehat{Var}\left(\widehat{Y}_t\right) = Q_1\left[\widehat{y}_{1,t}(s_1)\right] + \sum_{s_1}\left(\frac{1}{\pi_1^2} - q_1\right)^{+} Var\left[\widehat{y}_{1,t}|s_1\right] \qquad \textbf{(e II.2.11)}$$

et si $Q_1 \geqslant 0$ alors l'estimateur de covariance 'Ultimate Cluster' (**e II.2.10**) assure la condition de cohérence entre les estimations des covariances et des variances (**e II.2.12**) :

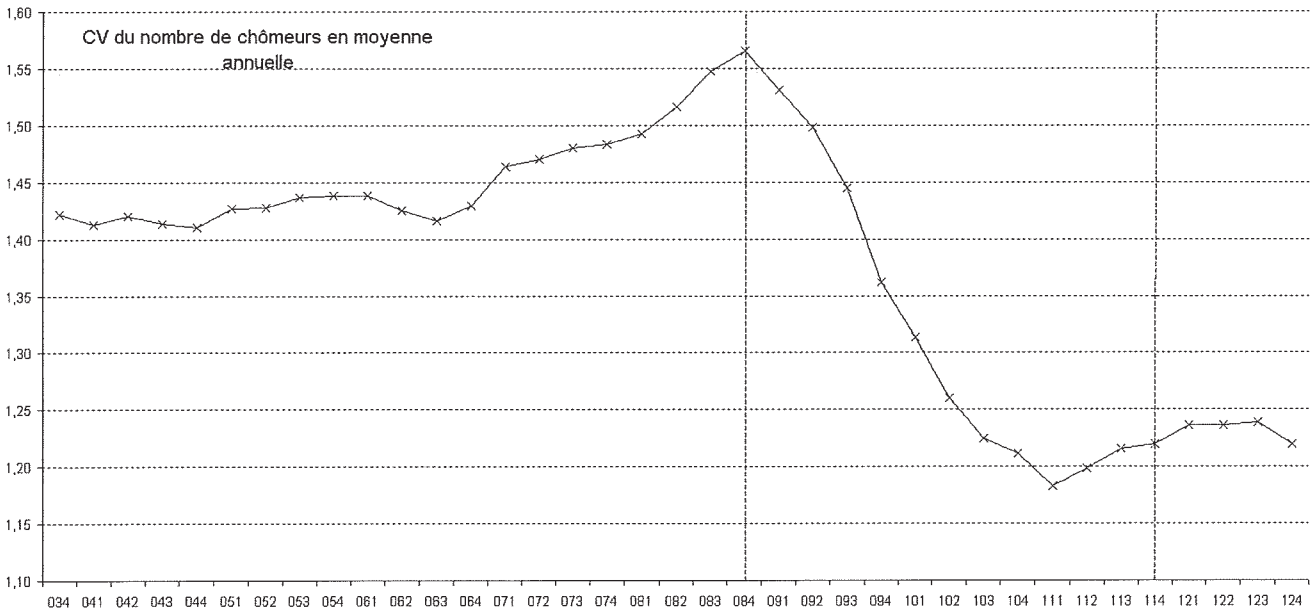
$$\begin{aligned}\left|\widehat{Cov}\left(\widehat{Y}_{t-l},\widehat{Y}_t\right)\right| &= \left|\widehat{y}_{1,t-l}(s_1)'Q_1\widehat{y}_{1,t}(s_1)\right| \\ &\leqslant \sqrt{Q_1\left[\widehat{y}_{1,t-l}\right]Q_1\left[\widehat{y}_{1,t}\right]} \\ &\leqslant \sqrt{\widehat{Var}\left(\widehat{Y}_{t-l}\right)\widehat{Var}\left(\widehat{Y}_t\right)} \qquad \textbf{(e II.2.12)}\end{aligned}$$

Cette propriété garantit que $\widehat{Var}\left(\widehat{Y}_t - \widehat{Y}_{t-l}\right) \geqslant 0$, pourvu que l'estimateur de la variance transversale soit toujours positif ($Q_1 \geqslant 0$).

II.3 deux applications du calcul de variance longitudinale

• Le Graphique 8 décrit le coefficient de variation de la moyenne annuelle nationale du nombre de chômeurs. Il montre clairement le gain en précision lié au doublement de l'échantillon entre fin 2008 et fin 2011. Toutefois, une légère détérioration est observée sur la période la plus récente.

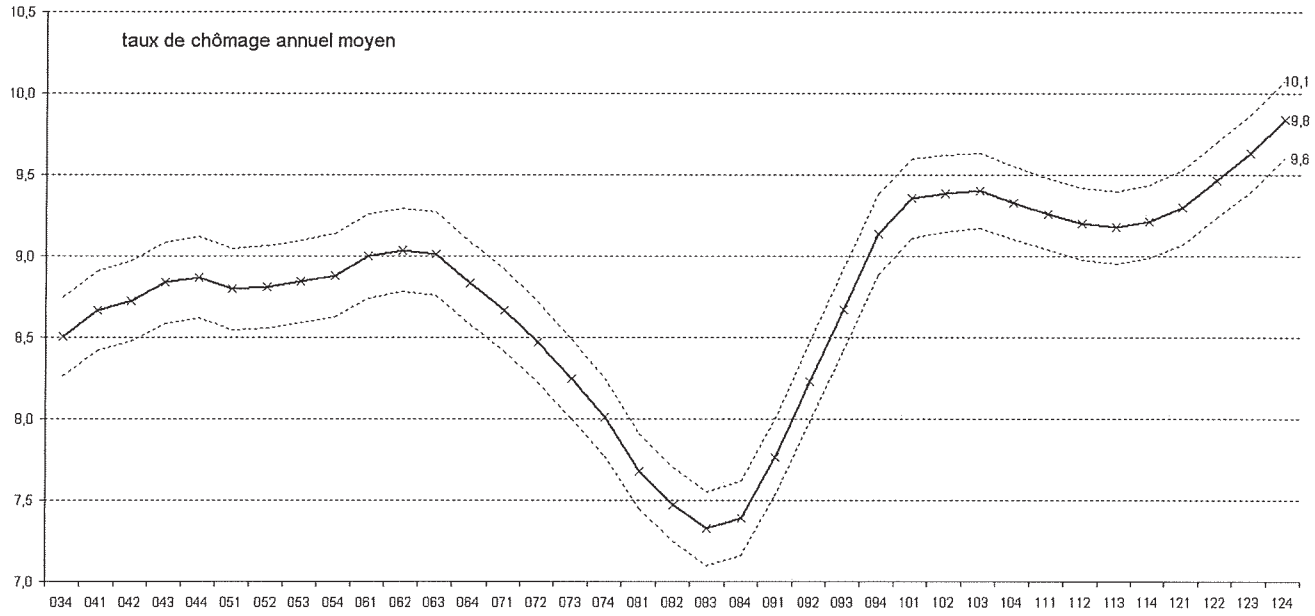
Graphique 8 – Coefficient de variation de la moyenne annuelle nationale du nombre de chômeurs, en %



Note : Au dénominateur, la moyenne annuelle d'un trimestre t est calculée par la somme des estimations des trimestres t-3 à t divisée par 4.

Comme une autre représentation de la précision de la moyenne annuelle, le Graphique 9 décrit l'intervalle de confiance à 95% pour le taux de chômage mesuré en moyenne annuelle.

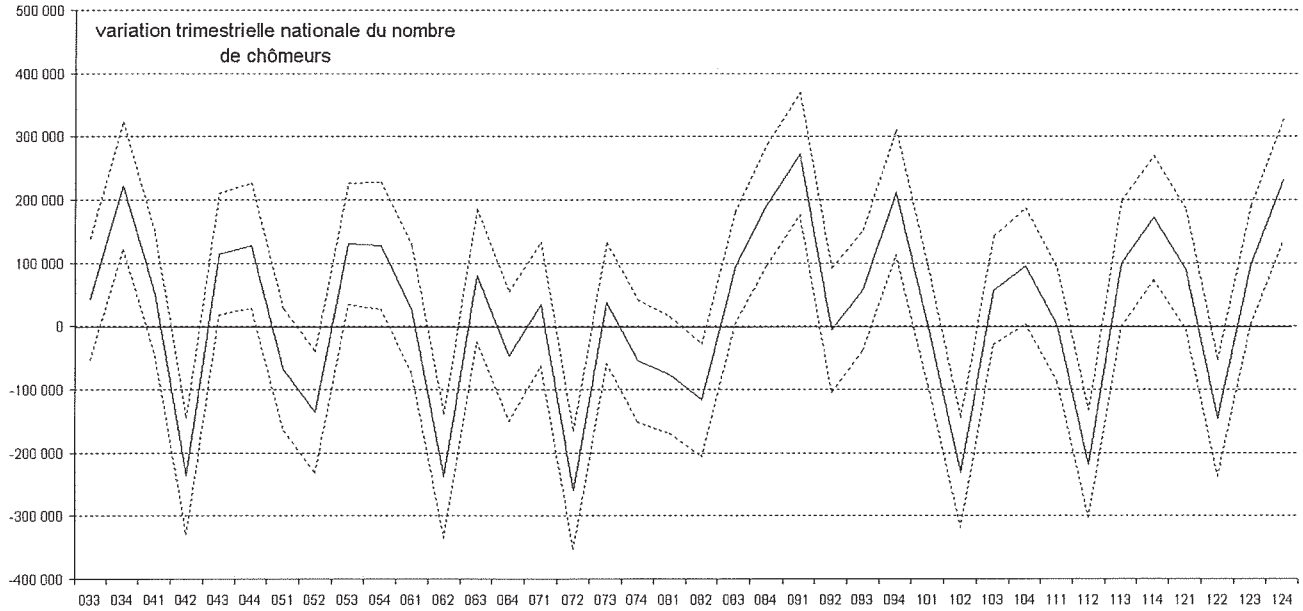
Graphique 9 – Intervalle de confiance du taux de chômage annuel moyen, en %



Note : L'intervalle de confiance est calculé en ajoutant et en retranchant deux fois l'écart-type estimé. Il est représenté par les deux courbes en pointillés. Pour mémoire, le taux de chômage annuel présenté ici constitue un résultat brut issu d'estimations directes réalisées de manière ad hoc pour les calculs de variance, et ne correspond donc pas au taux de chômage officiel diffusé par l'Insee.

– Le graphique suivant décrit l'évolution par trimestre de l'intervalle de confiance à 95% pour la variation trimestrielle annuelle du nombre de chômeurs. Il apparait qu'une variation nationale de 100 000 chômeurs est à la limite de la significativité.

Graphique 10 – Intervalle de confiance de la variation trimestrielle nationale du nombre de chômeurs



Note : Pour mémoire, les variations trimestrielles nationales du nombre de chômeurs présentées ici sont issues d'estimations directes réalisées de manière ad hoc pour les calculs de variance, et ne correspondent donc pas aux résultats de l'enquête Emploi diffusés par l'Insee.

La suite de la deuxième partie de la note analyse les résultats du programme d'estimation de la variance longitudinale pour les objectifs de précision d'Eurostat.

II.4 rappel des objectifs de précision d'Eurostat [12]

① précision des moyennes annuelles régionales : Pour un groupe de chômeurs représentant 5% de la population active d'une région, l'écart type de la moyenne annuelle ne doit pas dépasser 8% de la sous-population. Les régions de moins de 300 000 habitants sont exclues du champ de cette contrainte [13].

→ L'explicitation proposée par le groupe d'experts d'Eurostat se formule ainsi :

$$\frac{\sigma\left(\widehat{\overline{U}}_t\right)}{\overline{P}_t} \leqslant \sqrt{\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\left(1-\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\right)/2969} \qquad \textbf{(e II.4.13)}$$

où $\overline{P}_t$ est la (moyenne annuelle de la) population de 15 à 74 ans de la région, supposée connue.

② précision des variations trimestrielles nationales : Pour une sous-population de 5% de la population active (d'un pays de plus de 20 millions d'habitants), l'écart-type de la variation trimestrielle ne doit pas dépasser 2% de la sous-population.

→ L'interprétation par le groupe Eurostat du deuxième objectif de précision est exprimée par l'équation (**e II.4.14**).

$$\frac{\sigma\left(\Delta\widehat{U}_t\right)}{P_t} \leqslant \sqrt{\frac{\widehat{U}_t}{P_t}\left(1-\frac{\widehat{U}_t}{P_t}\right)\Big/\,47500} \qquad \textbf{(e II.4.14)}$$

où $\overline{P}_t$ est la population nationale de 15 à 74 ans, supposée connue.

La suite de cette section fournit une justification des deux relations de proportionnalité spécifiées par ces contraintes. Un complément est fourni en Annexe D.

notations :

- u_t est l'indicatrice du chômage à la date t, par individu
- $N = P$ est la taille de la population, supposée pérenne
- $n = |s|$ est la taille de l'échantillon d'individus, supposée constante dans le temps

---

- hypothèse : $Var\,(u_t) = Var\,(u)$ est constante en fonction du temps.
- Les expressions des deux bornes Eurostat peuvent se justifier à partir de l'hypothèse de base sur les covariances inter-temporelles (**e II.4.16**). L'Annexe F propose des justifications de cette hy-

---

12. On s'intéresse ici aux deux principaux objectifs de précision d'Eurostat relatifs à des estimations de nombre de chômeurs, obtenues en exploitant plusieurs échantillons trimestriels complets de l'EEC. Eurostat impose aussi des objectifs de précision pour des statistiques dites "structurelles", calculées une fois l'an *via* une procédure d'estimation spécifique à partir de l'empilement de sous-échantillons – pour un trimestre donné, unités entrantes dans l'échantillon – des quatre échantillons trimestriels de l'année. Le programme de calcul de variance présenté ici ne permettant pas de calculer la précision de telles statistiques, ces objectifs de précision relatifs à ces variables structurelles ne sont donc pas abordés dans ce document de travail.

13. La Corse, dont la population ne dépasse le seuil de 300 000 habitants qu'à partir de 2008, n'a pas été prise en compte dans les calculs de précision, menés sur la période 2003-2012.

pothèse.

$$Cov\left(\widehat{U}_{t-l},\widehat{U}_t\right) \cong f(t-l,t)\frac{N^2}{n}Cov\left(u_{t-l},u_t\right) \quad \textbf{(e II.4.15)}$$

$$\cong g(t-l,t)\frac{N^2}{n}Var(u) \quad \textbf{(e II.4.16)}$$

Le facteur de proportionnalité $g(t-l,t)$ dépend du sondage, notamment par le taux de recouvrement entre les deux échantillons, $o(t-l,t)$. Les Deff sont également impliqués, mais il parait difficile d'expliciter cette dépendance. De plus, la corrélation inter-temporelle $Corr\left(u_{t-l},u_t\right)$ intervient, vu l'équation précédente. Donc la conjoncture du chômage peut influer sur le facteur de proportionnalité $g(t-l,t)$. C'est pourquoi Eurostat prévoit de tenir compte des circonstances économiques pour le diagnostic de non-respect des contraintes de précision (tolerance in compliance assessment).

- Comme $u \in \{0,1\}$ [14], $Var(u) = E(u)(1-E(u)) = \frac{U}{P}(1-\frac{U}{P})$.

① La variance de la moyenne annuelle est proportionnelle à cette variance :

$$Var\left(\widehat{\overline{U}}_t\right) \cong \sum_{l,\widetilde{l}=0}^{3}\frac{1}{16}g\left(t-l,t-\widetilde{l}\right)\frac{N^2}{n}Var(u)$$

$$= G(t)\frac{P^2}{n}\frac{U}{P}(1-\frac{U}{P})$$

– La première contrainte de précision d'Eurostat peut se formuler ainsi :

$$\sigma\left(\widehat{\overline{U}}_t\right) \leqslant 0.08U \text{ lorsque } U = 0.05P$$

$\Longrightarrow$ Les deux dernières équations mènent à traduire la condition ainsi :

$$\sqrt{G(t)\frac{P^2}{n}0.05(1-0.05)} \leqslant 0.08\times 0.05P$$

$$\Longleftrightarrow \sqrt{\frac{G(t)}{n}} \leqslant \frac{0.08\times 0.05}{\sqrt{0.05(1-0.05)}} \cong \frac{1}{\sqrt{2969}}$$

② Pour la variation trimestrielle :

$$Var\left(\widehat{U}_t-\widehat{U}_{t-1}\right) = Var\left(\widehat{U}_t\right)+Var\left(\widehat{U}_{t-1}\right)-2Cov\left(\widehat{U}_t,\widehat{U}_{t-1}\right)$$

$$= \{g(t,t)+g(t-1,t-1)-2g(t-1,t)\}\frac{N^2}{n}Var(u)$$

$$= H(t)\frac{N^2}{n}Var(u)$$

$$\Longrightarrow Var\left(\widehat{U}_t-\widehat{U}_{t-1}\right) = \frac{H(t)}{n}\frac{U}{P}(1-\frac{U}{P})P^2$$

$\Longrightarrow$ La deuxième contrainte de précision d'Eurostat équivaut alors à :

$$\sqrt{\frac{H(t)}{n}0.05(1-0.05)P^2} \leqslant 0.02\times 0.05P$$

$$\Longleftrightarrow \sqrt{\frac{H(t)}{n}} \leqslant \frac{0.02\times 0.05}{\sqrt{0.05(1-0.05)}} = \frac{1}{\sqrt{47500}}$$

14. Ceci n'est vrai qu'au niveau individuel. C'est pourquoi le raisonnement porte sur l'échantillon d'individus, et non de logements.

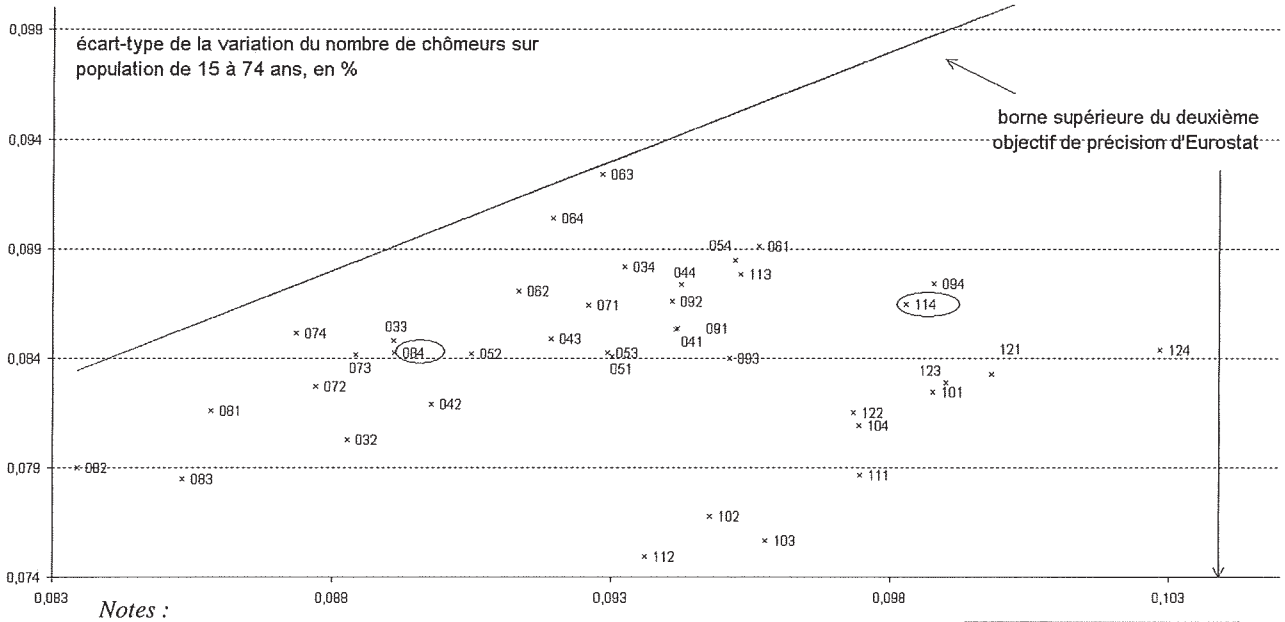
II.5 résultats obtenus pour les critères d'Eurostat

– La population de 15 à 74 ans au dénominateur des objectifs de précision est ici celle du recensement de 2008 (extraite des données locales du site Insee.fr).

– Pour mémoire, l'échantillon est exclusivement issu de la TH à partir du quatrième trimestre 2011 [VL]. Il était tiré uniquement 'dans le RP' jusqu'à fin 2008.

→ La deuxième contrainte d'Eurostat (celle sur $\sigma(\Delta\widehat{U}) \,/\, P$ au niveau national) est plus largement respectée depuis le troisième trimestre 2009 (Graphique 11). Ainsi, le bénéfice de l'augmentation du nombre de logements répondants ressort clairement au regard de la précision relative des variations trimestrielles nationales.

Graphique 11 – précision de la variation trimestrielle du nombre de chômeurs rapportée à la population en âge de travailler



Notes :

– *La borne supérieure du deuxième objectif d'Eurostat est calculée par la formule* $\sqrt{(U \,/\, P)\,(1 - U \,/\, P)\,\dfrac{1}{47500}}$, *où* P *est la population de référence de 15 à 74 ans.*

– *L'abscisse de ce graphique représente cette borne supérieure de la précision. Ainsi, chaque date* t *est représentée par le point de coordonnées* $\left(\sqrt{\widehat{U}_t \,/\, P_t\left(1 - \widehat{U}_t \,/\, P_t\right)/47500}, \widehat{\sigma}(\Delta\widehat{U}_t) \,/\, P_t\right)$. *La droite tracée est la bissectrice du graphique. Donc les points au dessus de cette droite ne respectent pas l'objectif de précision.*

– *Les trimestres sont identifiés sous le format* `aat`.

• Quatre régions ne satisfont pas le premier objectif de précision d'Eurostat, celui sur $\sigma(\widehat{\overline{U}}) \Big/ P$, au quatrième trimestre 2012 (Graphique 12). C'était le cas d'une seule région au dernier trimestre 2008.

L'objectif de précision d'Eurostat sur les moyennes régionales se traduit par l'inégalité (**e II.4.13**) :

$$\frac{\sigma\left(\widehat{\overline{U}}_t\right)}{\overline{P}_t} \leqslant \sqrt{\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\left(1-\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\right)/2969}$$

Or, à taille de population P approximativement constante dans le temps, plusieurs facteurs jouent, dans des sens différents et selon des ampleurs difficiles à apprécier *ex ante*, sur le respect de cette inégalité :

→ Pour une région donnée, l'écart type de la moyenne annuelle du nombre de chômeurs augmente avec le taux de chômage ; parallèlement, la borne $\sqrt{\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\left(1-\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\right)/2969}$ augmente également avec le taux de chômage, sans qu'il soit toutefois possible de déterminer *ex ante* si cette hausse est plus ou moins marquée que celle de $\frac{\sigma\left(\widehat{\overline{U}}_t\right)}{\overline{P}_t}$. Ainsi, le niveau du chômage régional affecte la position relative de $\frac{\sigma\left(\widehat{\overline{U}}_t\right)}{\overline{P}_t}$ par rapport à la borne $\sqrt{\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\left(1-\frac{\widehat{\overline{U}}_t}{\overline{P}_t}\right)/2969}$, ce qui impacte le respect du premier objectif de précision d'Eurostat dans des proportions impossibles à quantifier.

→ La hausse de la taille de l'échantillon opérée entre fin 2008 et fin 2011 joue bien évidemment à la baisse sur l'écart type de la moyenne annuelle du nombre de chômeurs, et donc en faveur d'un meilleur respect du premier objectif de précision d'Eurostat.

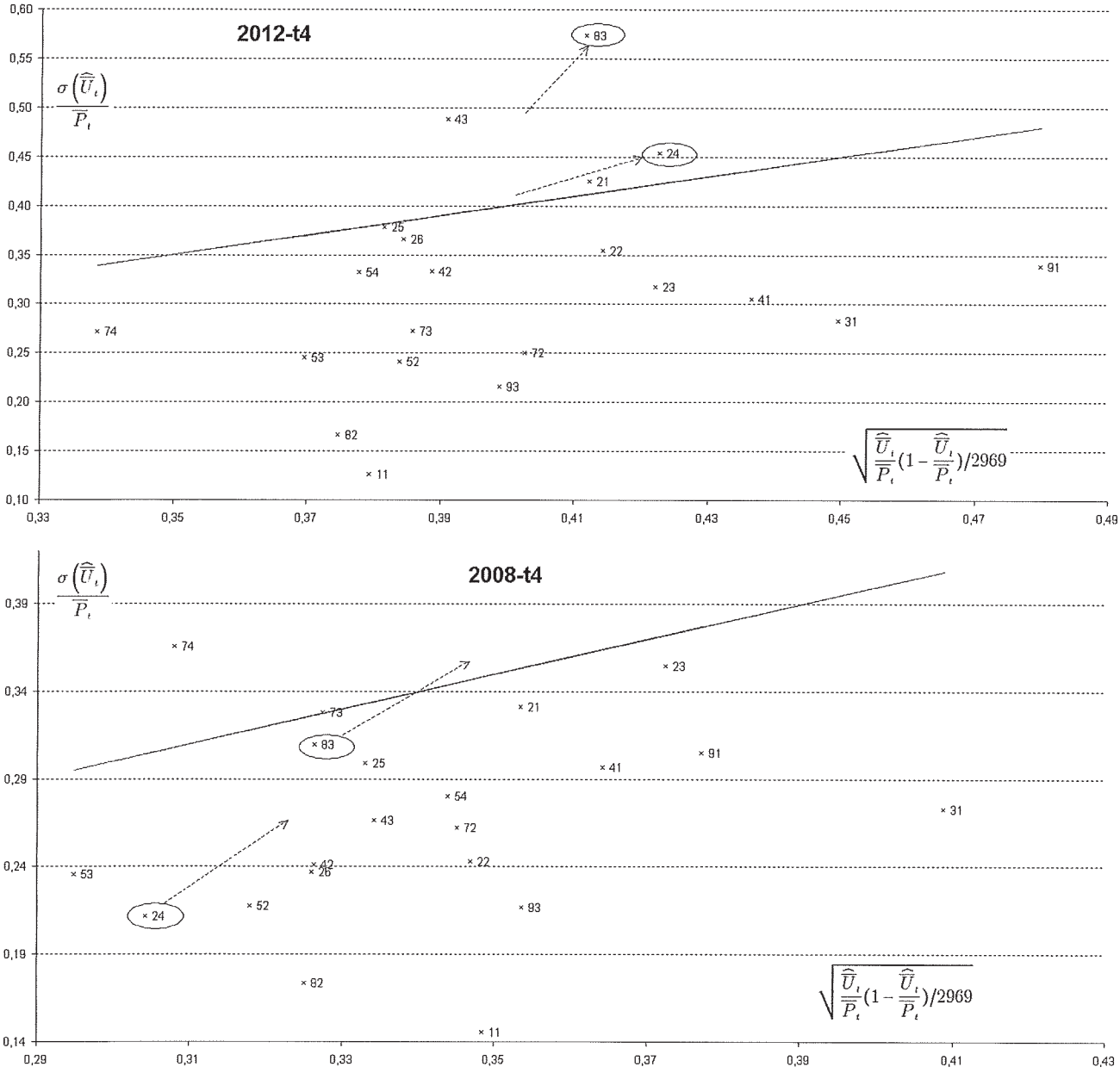
→ La taille de la région considérée joue également sur la précision de la moyenne annuelle régionale du nombre de chômeurs : en effet, la taille de l'échantillon par région étant approximativement proportionnelle à la taille de la région, plus la région est petite, plus la variance de la moyenne annuelle régionale du nombre de chômeurs est importante, et donc plus il est difficile de respecter le premier objectif de précision d'Eurostat.

→ Enfin, les taux de réponse régionaux varient d'un trimestre sur l'autre et d'une région à l'autre, ce qui n'est pas sans affecter la précision de la moyenne annuelle régionale du nombre de chômeurs.

Ainsi, la conjonction de ces différents facteurs explique les résultats décevants et *a priori* paradoxaux observés quant au premier objectif de précision d'Eurostat sur les moyennes régionales – à savoir un moindre respect au 4ème trimestre 2012 par rapport au 4ème trimestre 2008 de cet objectif malgré l'augmentation de 50% de la taille de l'échantillon –, comme le montrent les analyses détaillées présentées dans la suite de cette partie.

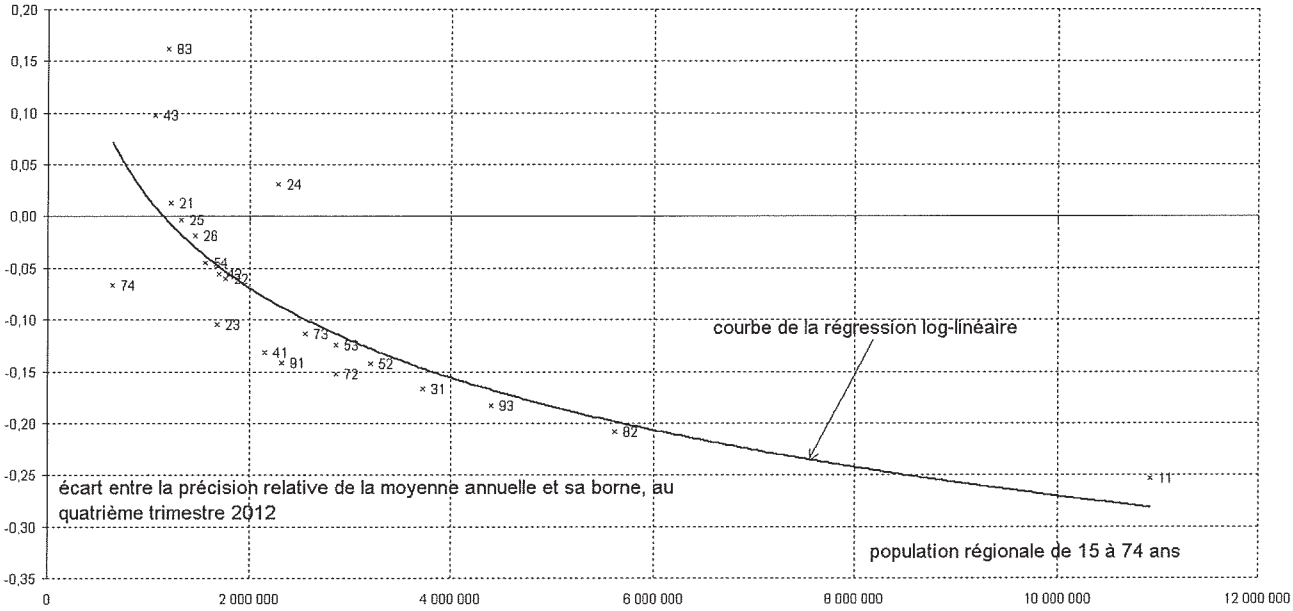
Le Limousin (74) est la plus petite région de cette étude. Pour les régions Auvergne (83) et Centre (24), le déplacement au dessus de la borne est concomittant avec la hausse de celle-ci. Ces cas illustrent l'effet de la conjoncture sur le positionnement relativement à la borne.

Graphique 12 – précision régionale de la moyenne annuelle du nombre de chômeurs rapportée à la population en âge de travailler, au quatrième trimestre 2012 et au dernier trimestre 2008



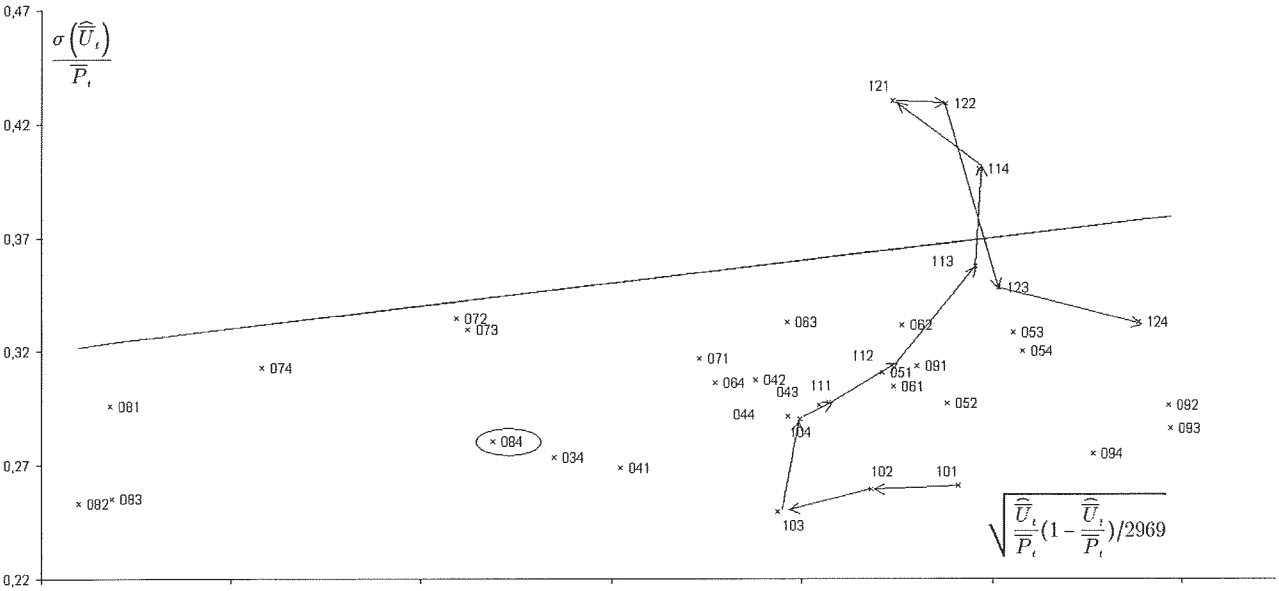
Il apparait clairement une relation linéaire décroissante entre l'écart de $\sigma(\widehat{\overline{U}}) \Big/ P$ à sa borne et le logarithme de la population utilisée au dénominateur de ce ratio (Graphique 13, $corr = -0.78$).

Graphique 13 – écart entre la précision relative de la moyenne annuelle du taux de chômage et sa borne pour Eurostat, au quatrième trimestre 2012



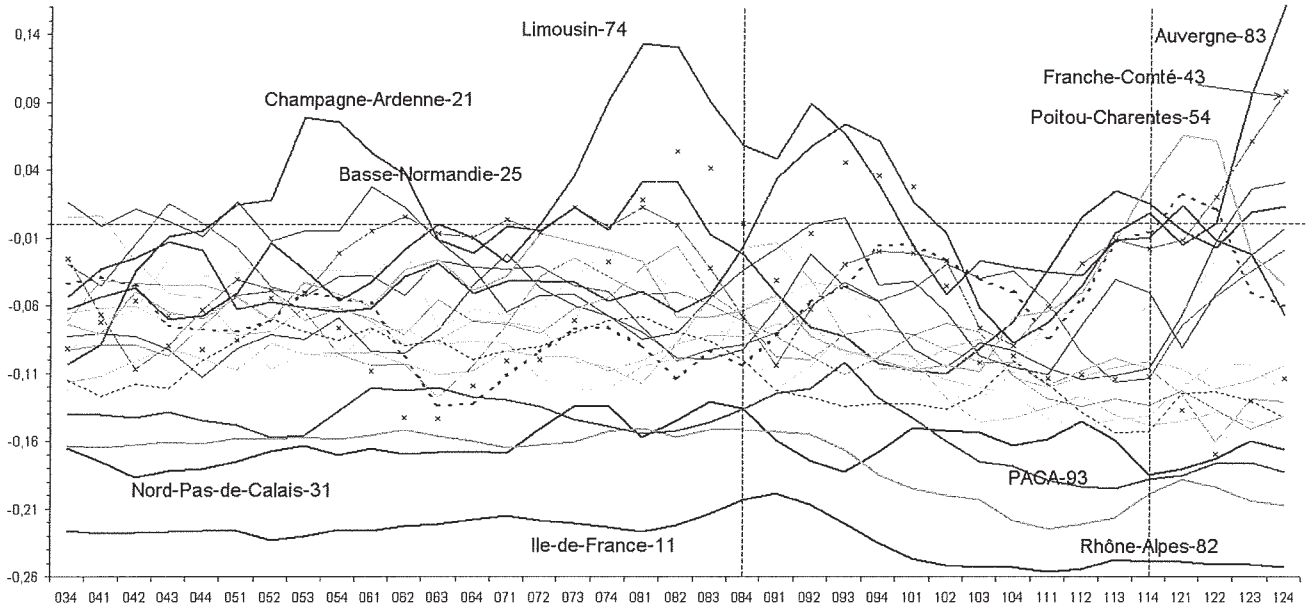
Pour le Poitou-Charentes (54), la précision de la moyenne annuelle ne dépasse sa borne que sur la période 2011/t4-2012/t2, alors que la précision de 2009/t2 à 2010/t3 est meilleure que précédemment (Graphique 14). Ceci suggère que ces points correspondent à un choc exceptionnel.

Graphique 14 – Précision de la moyenne annuelle du nombre de chômeurs pour le Poitou-Charentes



Pour plusieurs autres régions, l'écart entre la précision relative de la moyenne annuelle régionale et la borne proposée par Eurostat fluctue largement au cours du temps (Graphique 15). Il ne se dégage pas de tendance évidente depuis 2010. La précision relative ne s'améliore entre 2008/t4 et 2012/t4 que pour 7 régions. Par contre, la différence entre l'écart-type relatif de la moyenne annuelle et sa borne baisse pour 12 régions.

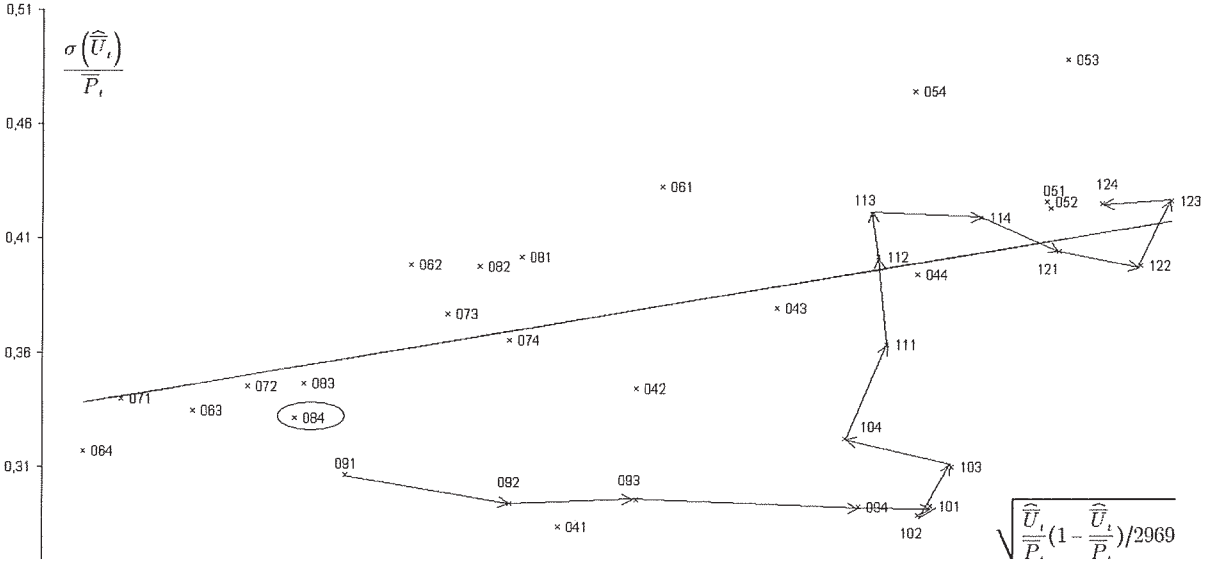
Graphique 15 – évolution régionale de la différence entre la précision relative de la moyenne annuelle du nombre de chômeurs et sa borne $\left(\sigma\left(\widehat{\overline{U}}\right)/P - \text{borne Eurostat}\right)$



Il ressort de cette analyse que l'évolution de la conjoncture de l'emploi peut masquer celle de la précision de l'estimation du nombre de chômeurs, surtout au niveau régional. En 2012, sept régions dépassent la borne de précision Eurostat au moins un trimestre. Cependant, depuis début 2009, aucune région n'est observée durablement au dessus de la borne Eurostat.

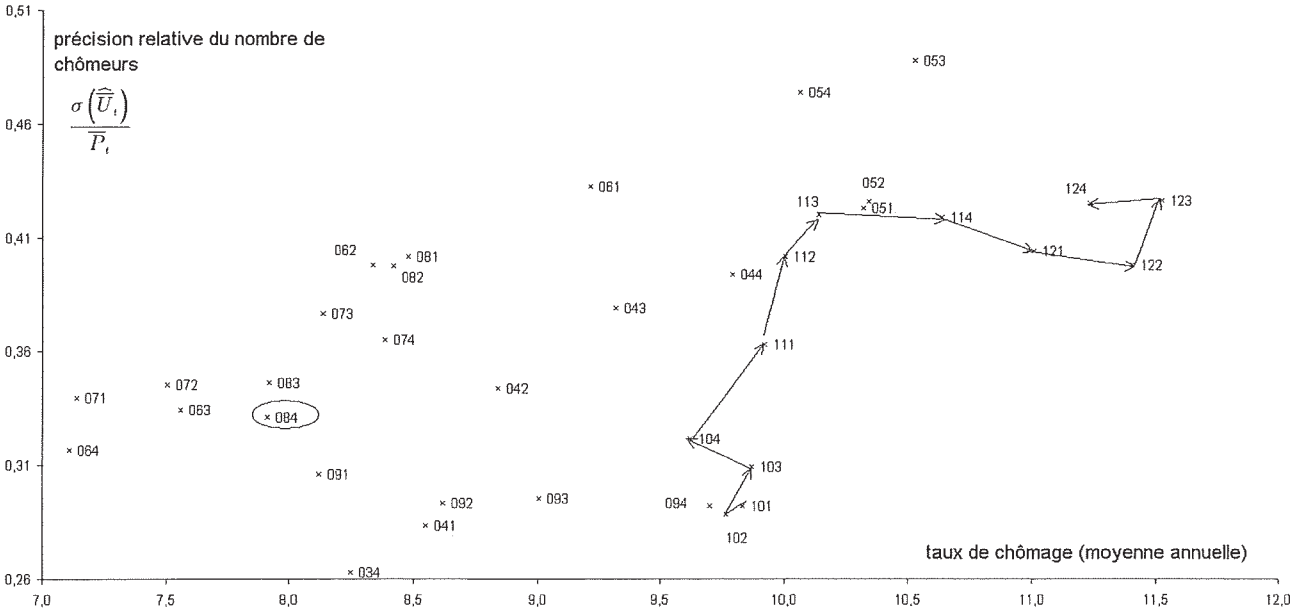
Comme le Poitou-Charentes, la Champagne-Ardenne a bénéficié d'une amélioration de la précision de la moyenne annuelle relativement à sa borne entre 2009/t1 et 2010/t4, mais la détérioration observée sur la période récente semble durable (Graphique 16).

Graphique 16 – précision de la moyenne annuelle du nombre de chômeurs - Champagne-Ardenne



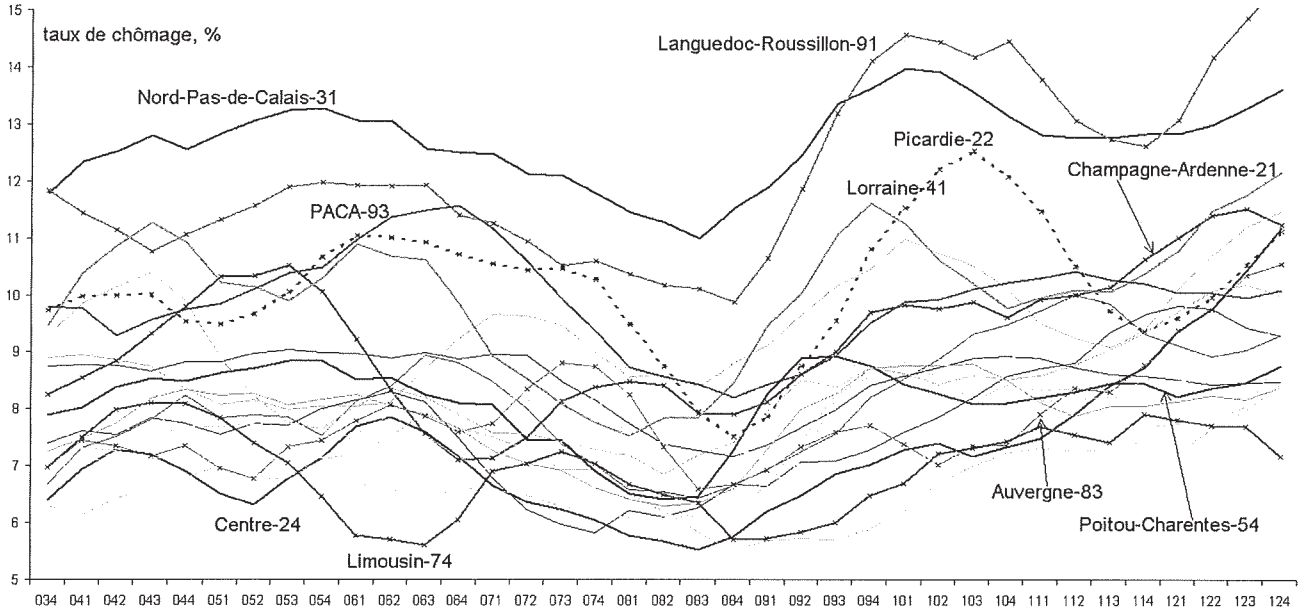
En fait, pour cette région, la précision de la moyenne annuelle du nombre de chômeurs en 2012 apparait relativement bonne compte tenu du taux de chômage particulièrement élevé (Graphique 17). C'est un des éléments qui suggèrent que la prise en compte de l'effet du taux de chômage par la borne d'Eurostat n'est pas toujours complète.

Graphique 17 – précision en fonction du taux de chômage de la Champagne-Ardenne



Pour la Champagne-Ardenne, le pic en 2005/t3 du coefficient de variation correspond à celui du taux de chômage (Graphique 18). En revanche, la stabilité de l'écart-type relatif entre le 2011/t3 et le 2012/t2 a lieu alors qu'une hausse du taux de chômage se poursuit. Dans ce cas, l'évolution de la précision pourrait s'expliquer par celle des répondants. En effet, le nombre de logements répondants a baissé de 14% entre 2010/t1 et 2011/t3, avant de se redresser légèrement. Pour le Limousin, le pic de l'écart-type relatif en 2008/t1 ne s'explique ni par l'évolution du taux de chômage (en baisse) ni par celle du nombre de logements répondants (stable).

Graphique 18 – Taux de chômage par région, en moyenne annuelle

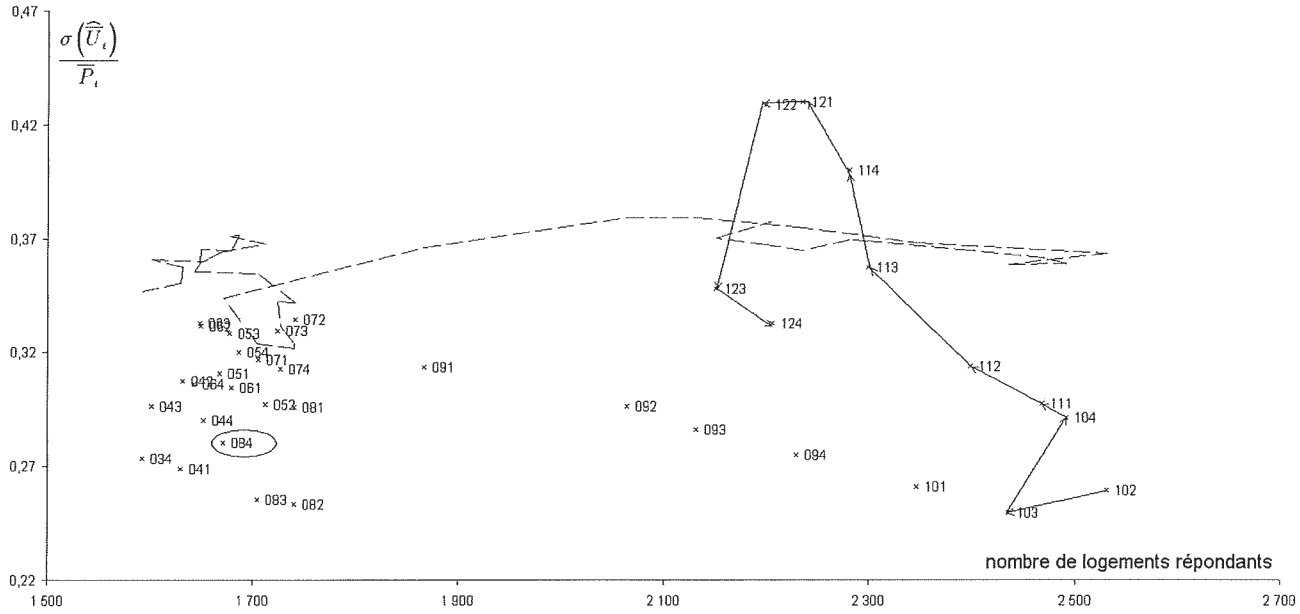


Note : Comme pour le graphique précédent, le taux de chômage est le rapport entre les moyennes annuelles des nombres de chômeurs et d'actifs, en pourcentage. Pour mémoire, les taux de chômage régionaux présentés ici constituent des résultats bruts issus d'estimations directes réalisées de manière ad hoc pour les calculs de variance, et ne correspondent donc pas aux taux de chômage localisés par régions diffusés par l'Insee.

Pour l'Auvergne et le Centre, le passage au dessus de la borne entre 2008/t4 et 2012/t4 coïncide avec une très forte augmentation du chômage (+58% et +94%).

Pour le Poitou-Charentes, la dégradation de la précision de la moyenne annuelle du nombre de chômeurs fin 2011-début 2012 ne correspond pas à une évolution notable du taux de chômage. En revanche, le nombre de logements répondants – qui joue à la baisse de l'écart-type de la moyenne annuelle en diminuant les variances transversales, via la variance due à la non-réponse – a baissé de 15% entre 2010/t2 et 2012/t3 dans cette région (Graphique 19), ce qui peut expliquer la dégradation de précision observée.

Graphique 19 – Précision relative et nombre de répondants - Poitou-Charentes



Notes :

– *Le nombre de répondants est celui du dernier trimestre de la moyenne annuelle. Il intègre les logements hors champ.*

– *La courbe en tirets représente la borne de précision Eurostat.*

Références

[AMYT] Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size, A.Matei, Y.Tillé, Journal of Official Statistics, vol.24,n° 4,2005

[BC] Improved variance estimation for balanced samples drawn via the Cube method, FJ Breidt, G Chauvet, Journal of Statistical Planning and Inference, 2010

[DT] Variance approximation under balanced sampling, JC Deville, Y Tillé, Journal of statistical planning and inference 2005

[GC] Échantillon équilibré stratifié, G Chauvet, Insee-Crest n° 2008-08

[JR] Calcul de la variance longitudinale des résultats de l'enquête Emploi, J.Richet, rapport de stage Insee, juillet 2011

[VL] La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation, V Loonis, document Insee JMS 2009

Annexe A estimation d'une forme quadratique de totaux par strate

– En général, l'approche analytique du calcul de variance estime la variance de l'estimateur $\widehat{Y}$ du total d'une variable d'intérêt y par une forme quadratique des observations sur l'échantillon sans remise s. C'est le cas en particulier pour :

– l'estimateur d'Horvitz-Thompson : $\widehat{Var\left(\widehat{Y}\right)}^{HT} = \sum_{i,j\in s} \frac{\pi(i,j)-\pi(i)\pi(j)}{\pi(i,j)} \frac{y}{\pi}(i)\frac{y}{\pi}(j)$

(où π est la probabilité d'inclusion, supposée telle que : $\forall i,j, \pi(i,j) > 0$)

– l'estimateur de Yates-Grundy : $\widehat{Var\left(\widehat{Y}\right)}^{YG} = -\frac{1}{2}\sum_{i,j\in s} \frac{\pi(i,j)-\pi(i)\pi(j)}{\pi(i,j)} \left(\frac{y}{\pi}(i) - \frac{y}{\pi}(j)\right)^2$

(non biaisé si la taille d'échantillon est fixe)

– l'estimateur de Deville-Tillé pour un sondage équilibré sur un vecteur x :

$$\widehat{Var\left(\widehat{Y}\right)}^{DT} = \sum_{i\in s} c_i \left(\frac{y}{\pi}(i) - \frac{x}{\pi}(i)'\left[\sum_s c\frac{xx'}{\pi^2}\right]^- \sum_s c\frac{yx}{\pi^2}\right)^2$$

(A^- désigne l'inverse généralisée de la matrice A. La pondération usuelle est : $c_i = (1-\pi_i)\frac{|s|}{|s|-|x|}$, où $|s|$ est la taille d'échantillon et $|x|$ est la dimension d'équilibrage.)

– Pour un sondage à plusieurs degrés, la variance de premier degré est une forme quadratique des totaux de premier degré. Pour estimer la variance d'un estimateur de second degré, il se pose la question d'estimer cette composante de la variance sur l'échantillon de deuxième degré.

A.1 principe général

Soit H un ensemble de strates (d'un univers U). Pour une variable d'intérêt y, y_H désigne le vecteur dont la composante associée à $h \in H$, y_h, vaut le total de y sur la strate h. Q est une forme quadratique de totaux par strate, c'est-à-dire une fonction du type :

$$Q(y_H) = \sum_{h_1,h_2\in H} q_{h_1,h_2} y_{h_1} y_{h_2}$$

où q_{h_1,h_2} est une constante réelle

→ La propriété fondamentale utilisée pour estimer $Q(y_H)$ est que si s_H est une famille d'échantillons indépendants dans les strates H et $\widehat{y}_H$ est un estimateur sans biais de y_H en fonction de s_H [15] alors la nullité des covariances entre estimateurs de strates différentes implique (**eA.17**).

$$E\{Q[\widehat{y}_H]\} = Q(y_H) + \sum_{h\in H} q_h Var[\widehat{y}_h] \qquad \textbf{(eA.17)}$$

où $q_h = q_{h,h}$ décrit les coefficients diagonaux de la matrice associée à Q

▫ <u>une démonstration plus détaillée :</u>

· La trace d'une matrice carrée $(n \times n)$ est la somme de sa diagonale : $tr(A) = \sum_i A_{i,i}$.

· C'est une fonction linéaire sur l'espace vectoriel réel des matrices carrées de dimension donnée.

15. Plus précisément : pour tout $h \in H$, $\widehat{y}_h$ est fonction de s_h.

· Si A est une matrice $n \times m$ et B une matrice $m \times n$ alors $tr(AB) = tr(BA)$.

· L'utilisation de la trace pour exprimer la forme quadratique permet de calculer son espérance ainsi :

$$\begin{aligned} Q[\widehat{y}_H] &= tr\{\widehat{y}_H' Q \widehat{y}_H\} = tr\{Q\widehat{y}_H \widehat{y}_H'\} \\ \Longrightarrow E\{Q[\widehat{y}_H]\} &= E[tr\{Q\widehat{y}_H \widehat{y}_H'\}] \\ &= tr[QE\{\widehat{y}_H \widehat{y}_H'\}] \qquad (\text{car } A \longmapsto tr(QA) \text{ est linéaire}) \\ &= tr\{Q(Var[\widehat{y}_H] + E[\widehat{y}_H] E[\widehat{y}_H]')\} \\ &= tr\{QVar[\widehat{y}_H]\} + tr\{E[\widehat{y}_H]' Q E[\widehat{y}_H]\} \\ &= E[\widehat{y}_H]' Q E[\widehat{y}_H] + tr\{QVar[\widehat{y}_H]\} = Q\{E[\widehat{y}_H]\} + tr\{QVar[\widehat{y}_H]\} \end{aligned}$$

remarque : Si $Q \geqslant 0$ alors $tr\{QVar[\widehat{y}_H]\} \geqslant 0$ et alors l'estimateur 'plug-in' $Q[\widehat{y}_H]$ surestime $Q(y_H)$.

· La dernière étape de la démonstration repose sur le fait que si les échantillons s_H sont indépendants alors $Var[\widehat{y}_H]$ est diagonale, de coefficient diagonal égal à $Var[\widehat{y}_h]$ pour l'indice h.

Par suite, (**eA.17**) permet de construire un estimateur sans biais de la forme quadratique de totaux en fonction d'estimateurs sans biais des variances intra-strate selon la formule suivante :

$$\widehat{Q(y_H)} = Q[\widehat{y}_H] - \sum_{h \in H} q_h \widehat{Var}[\widehat{y}_h]$$

– Plus généralement, à un estimateur (éventuellement biaisé) de la somme $\sum_{h \in H} q_h Var[\widehat{y}_h]$ correspond un estimateur de la forme quadratique défini par (**eA.18**).

$$\widehat{Q(y_H)} = Q[\widehat{y}_H] - \sum_{h \in H} q_h \widehat{Var}[\widehat{y}_H] \qquad \textbf{(eA.18)}$$

• De manière équivalente, un estimateur d'une forme quadratique de moyennes par strate s'obtient ainsi [16] :

$$\widehat{Q[\overline{y}_H]} = Q\left(\widehat{\overline{y}}_H\right) - \sum_{h \in H} q_h \widehat{Var}\left(\widehat{\overline{y}}_h\right) \qquad \textbf{(eA.19)}$$

remarque : Si δ_h désigne l'indicatrice sur H de la strate h, le coefficient diagonal q_h peut se caractériser par $q_h = Q(\delta_h)$.

16. Ceci peut être vu en appliquant (**eA.18**) à la variable qui vaut $z = \dfrac{y}{N_h}$ sur la strate h, où N_h est le nombre d'uniés de la strate h, lorsqu'il est connu.

A.2 application au 'collapse' d'unités primaires

- En particulier, si $|H|$ désigne le cardinal de l'ensemble H, le coefficient diagonal de :

$$\mathcal{S}^2(y_H) = \frac{1}{|H|-1}\sum_{h\in H}(y_h - \overline{y_H})^2 = \frac{1}{|H|-1}\sum_{h\in H}(y_h)^2 - \frac{1}{|H|(|H|-1)}\left[\sum_H y\right]^2 \qquad 17$$

vaut $\frac{1}{|H|-1} - \frac{1}{|H|(|H|-1)} = \frac{1}{|H|}$.

- Par suite, l'équation (**eA.17**) s'applique à cette forme quadratique ainsi :

$$E\left[\mathcal{S}^2\left(\widehat{\overline{y}}_H\right)\right] = \mathcal{S}^2(\overline{y}_H) + \sum_{h\in H}\frac{1}{|H|}Var\left(\widehat{\overline{y}}_h\right)$$

$\Longrightarrow$ Une somme par strate de variances d'estimateurs sans biais de moyennes par strate peut être surestimée par :

$$\widehat{\sum_{h\in H} Var\left(\widehat{\overline{y}}_h\right)} = \mathcal{S}^2\left(\widehat{\overline{y}}_H\right)|H| = \mathcal{S}^2\left\{\sqrt{|H|}\widehat{\overline{y}}_H\right\}$$

avec un biais positif de $\mathcal{S}^2(\overline{y}_H)|H|$, d'autant plus faible que les moyennes par strate sont moins dispersées. Cette approche est un cas particulier de la méthode de Hansen, Hurvitz et Madow (cf 'Introduction to variance estimation', KM Wolter, Springer 2007, pages 52-54).

Cette technique est utile lorsque la variance intra-strate ne peut pas être estimée. C'est le cas lorsqu'une seule unité est échantillonnée par strate. Plus précisément, le collapse consiste à regrouper les unités échantillonnées par groupes de deux ou trois unités, et à estimer pour un groupe H $\sum_{h\in H} Var_h(y)$ [18] par $\mathcal{S}^2_{s_H}\left\{\sqrt{|H|}y\right\}\left(=\mathcal{S}^2\left[\sqrt{|H|}y_{s_H}\right]\right)$. Cette méthode se généralise aisément à l'estimation d'une somme pondérée de variances intra-strate, pourvu que la pondération soit positive.

---

17. $\overline{y_H} = \frac{\sum_{h\in H} y_h}{|H|}$ est distingué de $\overline{y}_H$, qui désigne le vecteur dont la composante indicée par $h \in H$ vaut $\overline{y}_h$.

18. notation : Pour un ensemble A et une variable y définie sur A, $\overline{y_A} = \frac{\sum_A y}{|A|}$ et $Var_A(y) = \frac{1}{|A|}\sum_A (y - \overline{y_A})^2$.

Annexe B résultats détaillés de l'estimation transversale

Tableau B.1 – Série des estimations de la variance transversale

date	$\widehat{N}$	nb.log. répondants	nb.mén. répondants	nombre chômeurs	écart-type	CV	Deff	taux chômage	CV	Deff
031	30 814 181	46 970	37 168	2 316 809	45 632	2,0	1,4	8,6	2,0	1,4
032	30 780 290	46 281	36 558	2 184 050	44 090	2,0	1,4	8,1	2,0	1,4
033	30 898 296	45 381	35 462	2 226 935	44 677	2,0	1,4	8,2	2,0	1,4
034	30 986 342	46 857	37 173	2 449 053	46 863	1,9	1,4	9,1	1,9	1,4
041	31 147 289	47 593	37 888	2 499 626	47 290	1,9	1,4	9,2	1,9	1,5
042	31 184 876	47 329	37 504	2 262 949	45 013	2,0	1,4	8,4	2,0	1,5
043	31 235 712	46 296	36 395	2 377 418	46 319	1,9	1,4	8,7	1,9	1,4
044	31 484 122	47 699	37 699	2 505 363	47 796	1,9	1,4	9,2	1,9	1,5
051	31 651 292	47 880	37 808	2 436 400	46 171	1,9	1,4	9,0	1,9	1,4
052	31 723 524	47 355	37 154	2 301 082	46 581	2,0	1,5	8,4	2,0	1,5
053	31 929 946	46 391	35 905	2 431 868	47 269	1,9	1,4	8,8	1,9	1,4
054	32 177 978	47 116	36 598	2 559 043	48 915	1,9	1,4	9,3	1,9	1,5
061	32 331 835	47 863	37 321	2 583 464	49 333	1,9	1,4	9,4	1,9	1,5
062	32 367 648	47 401	36 779	2 346 710	46 954	2,0	1,4	8,6	2,0	1,5
063	32 495 258	46 542	35 601	2 426 470	49 342	2,0	1,5	8,7	2,0	1,5
064	32 576 692	47 886	37 080	2 379 162	46 159	1,9	1,4	8,6	1,9	1,4
071	32 660 285	49 128	38 431	2 414 036	48 883	2,0	1,5	8,7	2,0	1,5
072	32 725 866	49 257	38 455	2 154 800	43 325	2,0	1,4	7,8	2,0	1,4
073	32 888 141	48 237	37 033	2 191 521	46 231	2,1	1,5	7,8	2,1	1,5
074	33 069 955	49 365	38 189	2 137 171	43 570	2,0	1,4	7,7	2,0	1,4
081	33 084 463	50 101	39 065	2 060 334	44 561	2,2	1,5	7,4	2,2	1,5
082	33 433 673	49 752	38 334	1 943 486	41 582	2,1	1,4	7,0	2,1	1,4
083	33 403 431	48 928	37 286	2 035 652	45 114	2,2	1,5	7,2	2,2	1,5
084	33 429 091	49 641	38 228	2 228 128	45 796	2,1	1,4	7,9	2,1	1,5
091	33 572 441	53 851	41 682	2 501 145	48 094	1,9	1,5	8,9	1,9	1,6
092	33 734 204	57 906	44 665	2 495 402	48 587	1,9	1,7	8,8	1,9	1,7
093	33 873 986	60 845	46 466	2 553 290	47 010	1,8	1,6	9,0	1,8	1,7
094	33 944 720	65 698	50 689	2 765 330	46 933	1,7	1,6	9,8	1,7	1,7
101	34 265 858	70 045	53 921	2 764 454	45 608	1,6	1,6	9,8	1,6	1,6
102	34 391 554	73 574	56 163	2 533 571	41 232	1,6	1,5	9,0	1,6	1,5
103	34 481 103	72 238	54 738	2 590 632	42 685	1,6	1,5	9,1	1,6	1,5
104	34 625 344	73 430	56 302	2 685 974	43 851	1,6	1,6	9,5	1,6	1,6
111	34 745 436	74 009	57 316	2 687 000	43 209	1,6	1,6	9,5	1,6	1,6
112	34 871 483	73 760	57 137	2 465 530	41 710	1,7	1,5	8,7	1,7	1,6
113	34 717 726	72 300	55 997	2 564 649	43 886	1,7	1,6	9,0	1,7	1,6
114	34 944 226	73 965	57 982	2 736 033	46 445	1,7	1,7	9,6	1,7	1,7
121	34 637 926	71 435	56 536	2 825 728	46 759	1,7	1,7	9,9	1,6	1,7
122	34 737 768	72 259	57 132	2 680 107	45 330	1,7	1,6	9,4	1,7	1,6
123	34 998 975	71 624	55 935	2 777 467	45 128	1,6	1,6	9,7	1,6	1,6
124	35 335 937	73 714	57 727	3 009 547	48 001	1,6	1,7	10,5	1,6	1,7

Notes :

- *La pondération est celle de Hajek, y compris pour l'échantillonnage RP.*
- *Le collapse des unités primaires est effectué avec la stratification de première phase et tri par* $\sqrt{(1-\mu_1)^+}\frac{1}{\pi_2}\alpha$.
- *Les coefficients de variation (CV) et le taux de chômage sont en pourcentage.*

Tableau B.2 – Distribution du taux d'erreur sur la racine carrée de l'écart quadratique moyen de 5 000 simulations du tirage de deuxième phase, selon la méthode d'estimation

variable		moy	moyvar	min	p5	p10	q1	median	q3	p90	p95	max
foy_rev	v1a	-12,6	-2,1	-59,1	-48,7	-45,2	-38,7	-27,3	-3,7	38,7	82,8	269,5
foy_rev	v1b	-9,4	1,1	-56,6	-46,1	-42,5	-36,2	-24,7	-1,1	43,5	88,5	271,1
foy_rev	v1c	-13,4	-3,0	-59,6	-49,3	-45,8	-39,3	-28,1	-4,6	37,6	81,2	266,6
foy_rev	v2a	-7,1	3,9	-55,2	-44,9	-41,4	-34,8	-23,1	1,3	47,2	97,2	278,6
foy_rev	v2b	-4,5	6,2	-52,1	-41,7	-38,3	-32,0	-20,2	3,3	49,7	97,7	276,3
foy_rev	v2c	-7,9	3,0	-55,7	-45,5	-42,0	-35,4	-23,8	0,5	46,1	95,8	275,7
foy_rev	rena	-8,2	2,4	-56,1	-45,5	-41,9	-35,4	-23,8	0,6	45,9	90,6	271,7
foy_rev	renb	-7,0	3,5	-54,6	-44,0	-40,6	-34,2	-22,5	1,3	47,4	91,2	271,9
foy_rev	renc	-9,1	1,5	-56,6	-46,0	-42,5	-36,0	-24,5	-0,4	44,8	88,9	268,8
foy_rev	cena	-7,6	3,0	-55,7	-45,0	-41,4	-35,0	-23,3	0,9	46,6	90,9	272,1
foy_rev	cenb	-6,6	4,0	-54,1	-43,6	-40,2	-33,9	-22,1	1,6	47,7	91,3	271,9
foy_rev	cenc	-8,5	2,1	-56,2	-45,5	-42,0	-35,6	-24,1	0,0	45,4	89,2	269,2
foy_rev	vha	-5,7	5,5	-54,4	-43,9	-40,4	-33,8	-21,9	2,8	50,0	99,7	281,2
foy_rev	vhb	-3,6	7,3	-51,6	-41,3	-37,9	-31,5	-19,6	4,2	52,1	99,8	278,8
foy_rev	vhc	-6,5	4,6	-54,9	-44,5	-41,0	-34,4	-22,6	1,9	48,9	98,1	278,4
nbpi2	v1a	-5,7	-5,7	-12,7	-9,3	-8,5	-7,2	-5,7	-4,2	-2,9	-2,0	3,1
nbpi2	v1b	3,2	3,2	-4,5	-0,5	0,3	1,6	3,2	4,7	6,1	7,0	13,0
nbpi2	v1c	-6,9	-6,9	-13,9	-10,5	-9,8	-8,5	-7,0	-5,5	-4,1	-3,2	1,9
nbpi2	v2a	6,3	6,3	-1,3	2,2	3,0	4,6	6,2	7,9	9,6	10,6	16,7
nbpi2	v2b	13,1	13,1	4,9	8,9	9,8	11,3	13,1	14,8	16,4	17,5	25,4
nbpi2	v2c	5,1	5,2	-2,4	1,1	1,9	3,5	5,1	6,8	8,4	9,4	15,6
nbpi2	rena	5,2	5,2	-3,4	1,1	2,0	3,5	5,1	6,9	8,4	9,6	17,2
nbpi2	renb	8,2	8,2	-0,3	4,0	4,9	6,4	8,2	9,9	11,4	12,6	20,9
nbpi2	renc	4,1	4,1	-4,5	0,0	0,8	2,3	4,0	5,7	7,3	8,5	16,1
nbpi2	cena	6,8	6,8	-1,6	2,6	3,5	5,0	6,7	8,5	10,1	11,2	19,1
nbpi2	cenb	9,2	9,2	0,5	5,0	5,9	7,4	9,2	10,9	12,5	13,7	22,3
nbpi2	cenc	5,7	5,7	-2,8	1,5	2,4	3,9	5,6	7,4	8,9	10,1	18,0
nbpi2	vha	9,4	9,4	1,8	5,1	6,0	7,6	9,3	11,0	12,8	13,8	20,4
nbpi2	vhb	14,8	14,9	6,5	10,5	11,5	13,0	14,8	16,5	18,2	19,4	26,9
nbpi2	vhc	8,3	8,3	0,6	4,0	5,0	6,5	8,3	10,0	11,7	12,7	19,3

Notes :

- *La structure du codage des méthodes d'estimation est la suivante :*
  - *Le suffixe (a, b, c) correspond au paramètre μ_1 substitué à $q_1\pi_1^2$ dans la formule d'estimation* (**e I.1.3**) *:*
    - $a : 1 - \pi_1 .(„ 1)$
    - $b : \delta(c_1)$
    - $c : 1$
  - *Le début du code représente le choix de la pondération des résidus d'équilibrage (c_1) :*
    - *v1 :* $c_1 = 1 - \pi_1$
    - *v2 :* $c_1 = (1 - \pi_1) \sum_s 1 - \pi_1 \Big/ \sum_s \delta(1 - \pi_1)$
    - *ren : pondération c qui minimise* $\|\delta(c) - (1 - \pi_1)\|$
    - *cen : calage du poids de ren sur la trace de l'estimateur d'Horvitz-Thompson*
    - *vh : pondération de Hajek* $c_h = (1 - \pi_1)\,|s| \,/\, (|s| - |x|)$
- `moy` *désigne la moyenne des écarts-types estimés, alors que* `moyvar` *désigne la moyenne des variances estimées.*

Dans les deux cas, le taux d'erreur est mesuré par rapport à l'écart-type mesuré sur les simulations.

variable		moy	moyvar	min	p5	p10	q1	median	q3	p90	p95	max
mar	v1a	-3,8	-3,8	-10,1	-6,6	-6,0	-4,9	-3,8	-2,6	-1,6	-0,9	3,5
mar	v1b	15,9	15,9	9,4	13,1	13,7	14,8	15,8	17,0	18,1	18,7	23,3
mar	v1c	-5,7	-5,7	-12,2	-8,6	-7,9	-6,9	-5,7	-4,6	-3,5	-2,8	1,6
mar	v2a	11,2	11,2	4,0	7,8	8,6	9,8	11,1	12,6	13,9	14,7	19,5
mar	v2b	44,6	44,7	34,8	41,0	41,8	43,1	44,6	46,2	47,6	48,5	53,0
mar	v2c	9,5	9,6	2,3	6,1	6,9	8,1	9,5	10,9	12,3	13,1	17,8
mar	rena	10,6	10,6	2,9	7,0	7,7	9,1	10,5	12,0	13,5	14,3	18,3
mar	renb	25,9	25,9	16,0	22,1	22,9	24,3	25,9	27,5	29,0	30,0	34,4
mar	renc	8,9	8,9	1,1	5,2	6,0	7,4	8,8	10,4	11,8	12,7	16,8
mar	cena	12,7	12,7	5,0	9,0	9,7	11,2	12,6	14,2	15,7	16,5	20,8
mar	cenb	28,8	28,8	17,9	24,7	25,5	27,0	28,7	30,5	32,1	33,0	38,0
mar	cenc	11,0	11,1	3,3	7,3	8,0	9,5	10,9	12,5	14,1	14,9	19,2
mar	vha	15,2	15,2	7,5	11,5	12,4	13,7	15,1	16,6	18,1	19,0	23,7
mar	vhb	44,3	44,3	34,2	40,5	41,3	42,7	44,3	45,8	47,4	48,3	54,0
mar	vhc	13,6	13,6	5,8	9,9	10,7	12,1	13,5	15,0	16,5	17,4	22,1
isf	v1a	1,7	1,9	-16,1	-7,4	-5,6	-2,3	1,6	5,6	9,2	11,6	26,4
isf	v1b	6,8	7,0	-11,3	-2,2	-0,4	2,8	6,7	10,6	14,3	16,6	32,0
isf	v1c	0,6	0,7	-17,1	-8,5	-6,6	-3,4	0,4	4,4	8,0	10,3	24,9
isf	v2a	9,8	9,9	-8,7	0,5	2,4	5,7	9,6	13,6	17,5	19,8	35,2
isf	v2b	13,0	13,1	-6,2	3,7	5,6	8,9	12,9	16,8	20,6	22,9	38,9
isf	v2c	8,7	8,9	-9,7	-0,5	1,4	4,7	8,5	12,5	16,3	18,6	33,9
isf	rena	9,1	9,2	-10,5	-0,3	1,7	5,0	8,9	12,9	16,7	19,0	34,5
isf	renb	10,6	10,7	-9,1	1,4	3,2	6,5	10,5	14,4	18,1	20,4	36,8
isf	renc	8,0	8,1	-11,4	-1,3	0,7	3,9	7,8	11,8	15,5	17,8	33,2
isf	cena	10,0	10,2	-9,7	0,7	2,6	5,9	9,9	13,8	17,6	19,9	35,4
isf	cenb	11,1	11,2	-8,6	1,8	3,7	7,0	10,9	14,9	18,6	20,9	37,2
isf	cenc	8,9	9,1	-10,5	-0,3	1,6	4,9	8,8	12,7	16,5	18,7	34,2
isf	vha	11,9	12,0	-6,9	2,5	4,5	7,8	11,7	15,6	19,5	21,9	37,2
isf	vhb	14,3	14,5	-5,0	5,1	6,9	10,2	14,2	18,1	22,0	24,3	40,5
isf	vhc	10,8	11,0	-7,7	1,5	3,5	6,7	10,6	14,6	18,5	20,8	36,1

Note : Description rapide des simulations effectuées :

– Le tirage de première phase (grappes de secteurs) a été effectué sous le plan de sondage effectif. Il s'agit d'un tirage stratifié par région, à probabilités d'inclusion inégales et équilibré sur 32 variables, y compris celles pour équilibrer l'échantillon complémentaire. Le programme repose sur la macro `exe_cube` *de Guillaume Chauvet.*

– La répartition des grappes de secteurs par groupe et sous-groupe de rotation est également effectuée par les programmes de tirage effectifs. Il s'agit d'un sondage stratifié par région avec atterrissage national, selon le principe décrit par [GC]. La macro principale s'appelle `echantstrat2`*, rédigée par G. Chauvet. Un seul tirage de cette partition est effectué pour chacun des tirages de première phase.*

– La simulation du tirage de deuxième phase diffère du tirage effectif uniquement par l'absence de prise en compte de l'emboîtement des tirages des 6 grappes de logements par secteur. Pour évaluer l'estimateur de la variance transversale, les 5 000 *simulations exploitées dans le tableau ci-dessus sont limitées au tirage de la première grappe de logements, donc celui effectué avec la liste complète des variables d'équilibrage de deuxième phase. Un seul tirage de deuxième phase est effectué pour chaque simulation du tirage de première phase (ainsi que la répartition associée en 12 sous-groupes de rotation).*

Annexe C estimation transversale sur la période de l'échantillonnage RP1999

C.1 variance de première phase sur l'échantillon RP

Le tirage des unités primaires de l'échantillon RP est considéré comme un sondage de taille fixe à probabilités inégales. L'approximation de Deville est un cas particulier de Deville-Tillé avec pondération de Hajek, $c_k = (1-\pi)\frac{|s|}{|s|-1}$, et $x = \pi$. Cet estimateur de variance se formule donc ainsi :

$$\widehat{V}^D = \sum_s (1-\pi)\frac{|s|}{|s|-1}\left\{\frac{y}{\pi} - \frac{\sum_s (1-\pi)\frac{y}{\pi}}{\sum_s (1-\pi)}\right\}^2 = \frac{|s|}{|s|-1}\left\{\sum_s (1-\pi)\left(\frac{y}{\pi}\right)^2 - \frac{\left\{\sum_s (1-\pi)\frac{y}{\pi}\right\}^2}{\sum_s 1-\pi}\right\} \quad \textbf{(eC.20)}$$

Cette formule définit la forme quadratique utilisée pour la variance de première phase de l'échantillonnage RP 1999 dans l'estimateur (**e I.1.3**).

NB : Pour que les calculs de (**e I.2.4**) et (**eC.20**) soient corrects, toutes les unités primaires échantillonnées doivent être prises en compte, même lorsque leur groupe de rotation [19] n'est pas enquêté pour le trimestre considéré.

C.2 partage des poids RP-TH

Durant les six premiers trimestres de la phase de transition RP-TH 2009T1-2011T3, l'échantillon du groupe de rotation RP entrant à chaque trimestre est complété par un échantillon tiré dans la TH et de taille égale à la moitié du précédent. Ceci peut expliquer qu'un coefficient de partage de poids de 2/3 est appliqué au premier (sous-)échantillon, et de 1/3 pour le second. Pour être cohérent avec ce choix, le poids de tirage de l'échantillon TH correspond à une extrapolation au groupe de rotation, et donc environ trois fois plus élevé qu'en régime de croisière.

A cause de l'indépendance des échantillons RP et TH, la variance de l'estimateur peut se décomposer ainsi :

$$\begin{aligned} Var\left[\widehat{Y}\right] &= Var\left\{\widehat{\sum \alpha_{RP} y} + \widehat{\sum \alpha_{TH} y}\right\} \\ &= Var\left\{\widehat{\sum \alpha_{RP} y}\right\} + Var\left\{\widehat{\sum \alpha_{TH} y}\right\} \end{aligned}$$

Il suffit donc d'appliquer les formes quadratiques RP et TH définies par (**e I.1.3**) aux variables $\alpha_{RP} y$ et $\alpha_{TH} y$ (nulles si pas d'échantillon correspondant), et d'additionner les deux.

remarque : Le partage des poids est appliqué après le calcul des résidus du calage, mais avant celui des résidus d'équilibrage. Pour la variance tenant compte du calage, l'estimateur (**e I.2.4**) est calculé en remplaçant la variable d'intérêt par le résidu de calage multiplié par le coefficient de partage des poids.

- adaptation du collapse des unités primaires au partage des poids :

19. Le groupe de rotation est l'ensemble des unités primaires de même rang d'interrogation.

Les secteurs sont fusionnés dans l'ordre du tri par la variable $\sqrt{(1-\mu_1)^+}\frac{1}{\pi_2}\alpha$, où :

▷ $\mu_1^{RP} = \left(1 - \pi_1 - \frac{(1-\pi_1)^2}{\sum_{s_1} 1 - \pi_1}\right)\frac{|s_1|}{|s_1|-1}$.

▷ $\frac{1}{\pi_2}$ est le poids de tirage de la grappe de logements calculé par `redecha` pour l'échantillon RP, où `redecha` est lu dans la table `secteur07v2`

▷ α est le coefficient de partage des poids [20]. Sa prise en compte parait judicieuse du fait que la variance entre une unité de $\alpha = 0$ et une unité de $\alpha \neq 0$ doit être relativement élevée. En fait, son incidence s'avère négligeable.

Le collapse est effectué sur l'ensemble des secteurs échantillonnés dans chacune des sources, y compris sur la période de transition. Néanmoins, du fait de la prise en compte du coefficient de partage des poids dans la construction des collapses, ces regroupements peuvent changer entre les trimestres.

Annexe D une interprétation du facteur de proportionnalité pour la variance de la moyenne annuelle

Si • $\overline{s} = \frac{\sum_{l=0}^{3} s_{t-l}}{4}$ est considéré comme un sondage 'généralisé' de taille n [21]

• $u_{t-l} = u$ pour $l \in [0,3]_{\mathbb{N}}$ (constance du chômage sur la période de la moyenne)

et si le Deff du sondage généralisé $\overline{s}$ est considéré comme constant [22] alors :

$$Var\left(\widehat{\overline{U}}_t\right) = Var\left[\left(\frac{u}{\pi}\right)'\overline{s}\right]$$
$$\cong \text{Deff}\frac{N^2}{n}Var\,(u)$$

20. Un point délicat pour le programme d'estimation de la variance est d'assurer la cohérence entre ce coefficient et les poids de tirage sur la période de montée en régime de l'échantillonnage TH .

21. Si $s \in [0,1]^{\mathscr{P}}$ et pour $i \in \mathscr{P}$, $\pi(i) = E(s_i) > 0$ alors $\left(\frac{y}{\pi}\right)'s$ est un estimateur sans biais du total de y.

22. s est un sondage de Deff constant ssi $Var\left(\frac{s}{\pi}\right) = \text{Deff}\frac{N^2\left(1-\frac{n}{N}\right)}{n}\left(\text{id} - \frac{\mathbb{1}\mathbb{1}'}{N}\right) = \text{Deff}\frac{N^2\left(1-\frac{n}{N}\right)}{n}\text{proj}_{\mathbb{1}}^{\perp}$. Cette condition est très contraignante. Par exemple, elle exclut l'équilibrage sur un vecteur non constant, comme $x'\text{proj}_{\mathbb{1}}^{\perp}x \neq 0$.

Annexe E évaluation d'une prise en compte partielle des covariances intra-unité primaire

Cette annexe évalue l'importance du deuxième terme de (**e II.2.9**), omis dans l'estimateur de covariance choisi. Pour cela, un estimateur altenatif est construit, qui tient compte des covariances intrasectorielles, approximativement et uniquement pour les secteurs dont l'échantillon de deuxième phase (soit la grappe de logements) n'est pas renouvelé entre $t-l$ et t.

L'objectif est restreint à estimer :

$$\sum_{s_1^{nr}} \left(\frac{1}{\pi_1^2} - q_1 \right) Cov\left(\widehat{y}_{1,t-l}, \widehat{y}_{1,t} | s_1\right) \qquad \textbf{(eE.21)}$$

où s_1^{nr} est l'échantillon des unités primaires dont l'échantillon de deuxième phase n'est pas renouvelé entre $t-l$ et t.

– une technique de collapse pour estimer une somme de covariances :

On se place dans le cadre simplifié de sondages SAS (1) indépendants dans l'ensemble de strates H. Dans l'équation suivante, $\mathcal{S}^2$ désigne à la fois la forme quadratique donnant la variance empirique et sa matrice. $\overline{y}_H$ est le vecteur dont la h-ième composante vaut $\overline{y}_h$, la moyenne de y sur la strate h. s_H est la famille d'échantillon indicée par les strates ; $|s_h| = 1$.

$$E\left[\mathcal{S}^2_{s_H}(x,y)\right] = E\left[x'_{s_H}\mathcal{S}^2 y_{s_H}\right]$$
$$= \overline{x}'_H \mathcal{S}^2 \overline{y}_H + \sum_{h\in H} \frac{1}{|H|} Cov_h(x,y)$$

(vu l'indépendance inter-strates de s_H et le fait que le coefficient diagonal de $\mathcal{S}^2_{s_H}$ vaut $1/|H|$)

Comme $\overline{x}'_H \mathcal{S}^2 \overline{y}_H = \sum_{h\in H} \left(\overline{x}_h - \overline{\overline{x}}_H\right)\left(\overline{y}_h - \overline{\overline{y}}_H\right)/(|H|-1) = \mathcal{S}^2\left(\overline{x}_H, \overline{y}_H\right)$, il s'ensuit que $|H|\mathcal{S}^2_{s_H}(x,y)$ estime $\sum_{h\in H} Cov_h(x,y)$ avec un biais de $|H|\mathcal{S}^2\left(\overline{x}_H, \overline{y}_H\right)$. Ce défaut d'estimation est petit si les strates de H sont homogènes pour les moyennes de la variable (x,y) (homogénéité inter-strates pour ces deux variables).

→ En notant $\widetilde{i}_1$ l'élément générique d'une partition de s_1^{nr} en unités de collapse de deux ou trois secteurs, le terme des covariances 'intra' (**eE.21**) peut être estimé sur l'échantillon de deuxième phase par :

$$\sum_{i_1\in s_1^{nr}} \left(\frac{1}{\pi_1^2} - q_1 \right) \widehat{\frac{1}{\pi_{2|1}{}^2} Cov_{i_1}\left(y_{2,t-l}, y_{2,t}\right)} = \sum_{\widetilde{i}_1} |\widetilde{i}_1| \mathcal{S}^2_{\widetilde{i}_1} \left\{ \sqrt{|1-q_1\pi_1^2|}\frac{y_{2,t-l}}{\pi_2}, \text{signe}\left(1-q_1\pi_1^2\right)\sqrt{|1-q_1\pi_1^2|}\frac{y_{2,t}}{\pi_2} \right\}$$

→ Il en découle l'estimateur de covariance approximatif sur l'échantillon de deuxième phase (**eE.22**), en notant $\mu_1 = q_1\pi_1^2$.

$$\widehat{C}^3 = \widehat{y_{1,t-l}}^{3\prime} Q_1 \widehat{y_{1,t}}^3 + \sum_{\widetilde{i}_1 \subset s_1^{nr}} |\widetilde{i}_1| \mathcal{S}^2_{\widetilde{i}_1} \left\{ \sqrt{|1-\mu_1|}\frac{\widehat{y_{2,t-l}}^3}{\pi_2}, \text{signe}\left(1-\mu_1\right)\sqrt{|1-\mu_1|}\frac{\widehat{y_{2,t}}^3}{\pi_2} \right\} \qquad \textbf{(eE.22)}$$

• Il s'avère que l'estimateur (**eE.22**) donne des résultats qui ne diffèrent que marginalement de ceux de l'estimateur retenu, plus simple.

▷ Pour l'incertitude relative sur la variation nationale du nombre de chômeurs, qui est exprimée en pourcentage, l'effet de cette prise en compte des covariances est manifestement négligeable

(moins de 0.005% de la population retenue, ou 6% de l'écart-type de la variation).

▷ Pour la moyenne annuelle du nombre de chômeurs, l'effet sur l'écart-type de la prise en compte (partielle) des covariances intra-UP varie entre -13.7% à +10.3% de l'écart-type. Ceci ne joue que sur la deuxième décimale du ratio $\sigma(\widehat{\overline{U}}) \Big/ P$ exprimé en pourcentage, soit entre -0.04% et +0.05%. Sur la période 2003/t4-2008/t4, le terme additionnel des covariances n'est négatif que dans un seul cas (région X trimestre). En revanche entre 2011/t4 et 2012/t4, il est plus souvent négatif (65 cas) que positif (40 cas). Ceci correspond sans doute à une modification du signe du coefficient $1 - q_1\pi_1^2$, qui est positif pour l'estimateur de la période 'RP' [23].

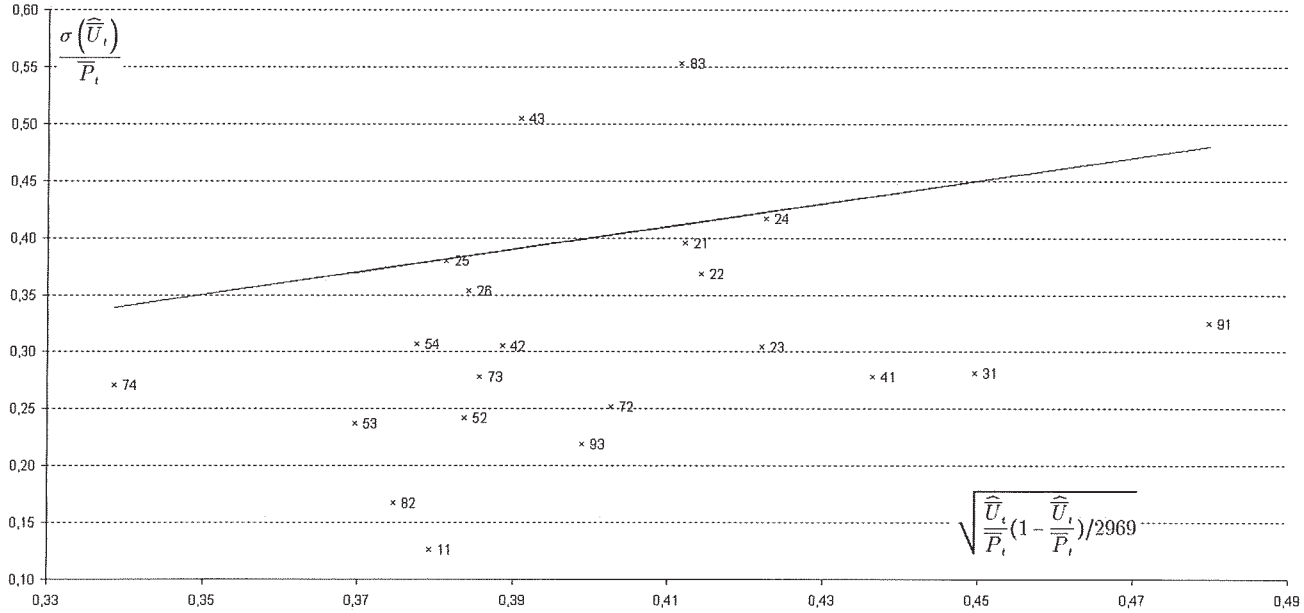
---

23. Pour cet estimateur, le coefficient diagonal vaut $q_1 = \dfrac{1-\pi_1}{\pi_1^2}\dfrac{|s_1|}{|s_1|-1}\left(1 - \dfrac{1-\pi_1}{\sum_{s_1}(1-\pi_1)}\right)$, qui est inférieur à $\dfrac{1}{\pi_1^2}$. En effet, pour $0 \leqslant \theta_s \leqslant 1$, $\theta_i \dfrac{|s|}{|s|-1}\left[1 - \dfrac{\theta_i}{\sum\theta_s}\right] = \theta_i \dfrac{\sum_{s\neq i}\theta \Big/ |s_{\neq i}|}{\sum_s \theta \Big/ |s|}$. Ce ratio est du type :

$\dfrac{xy}{\alpha x + (1-\alpha)y} \leqslant \dfrac{xy}{\inf(x,y)} \leqslant \sup(x,y)$ (avec $\alpha \in [0,1]$ et $0 < x, y \leqslant 1$).

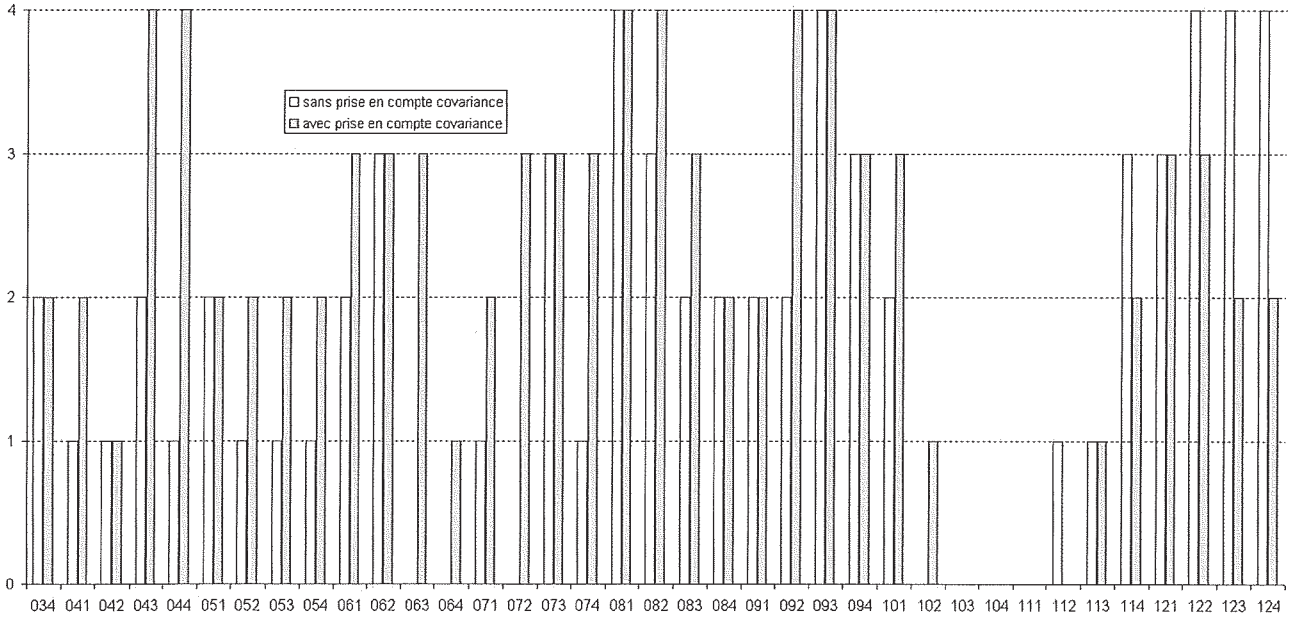
→ Au quatrième trimestre 2012, la prise en compte approximative des covariances limite à deux le nombre de régions excédant le premier objectif de précision Eurostat (Graphique 20).

Graphique 20 – estimation alternative de la précision de la moyenne annuelle du nombre de chômeurs au quatrième trimestre 2012



Inversement, le nombre de régions excédant l'objectif de précision est plus élevé sur la période 'RP' avec le terme additionnel (Graphique 21).

Graphique 21 – Nombre de régions qui excédent l'objectif de précision régionale



– Il est vraisemblable que pour l'échantillonnage dans la TH, le collapse tende à sous-estimer le terme additionnel des covariances. Ce serait donc une option qui mènerait à sous-estimer la variance des moyennes annuelles estimées.

Annexe F covariance temporelle entre estimateurs du chômage

- justification simplifiée :

Si • $s_{t-l} = s_t$ (pas de renouvellement de l'échantillon)
• $u_{t-l} = u_t$ (pas d'évolution du chômage entre les deux dates)

alors $Cov\left(\widehat{U}_{t-l}, \widehat{U}_t\right) = Var\left(\widehat{U}_t\right) \cong \text{Deff}\frac{N^2}{n} Var\,(u)$. Ceci découle de la définition du Deff. Mais pour être utile, il faut supposer de plus que le Deff ne varie pas en fonction du chômage.

- justification avec un renouvellement simple : Soit $s_{t-l,t}$ l'échantillon des individus pérennes entre les deux dates. Si les échantillons entrant, pérenne et sortant sont considérés comme indépendants alors $Cov\left(\widehat{U}_{t-l}, \widehat{U}_t\right) = Cov\left\{\sum_{s_{t-l,t}} \frac{u_{t-l}}{\pi_{t-l}}, \sum_{s_{t-l,t}} \frac{u_t}{\pi_t}\right\}$. La probabilité d'inclusion dans l'échantillon pérenne est approximée par $\pi_{t-l,l} = \pi_{t-l} o\,(t-l,t) = \pi_t o\,(t-l,t)$, avec $o\,(t-l,t) = |s_{t-l,t}| \,/\, |s_t|$. Alors l'utilisation du Deff du sondage $s_{t-l,t}$ pour la variable $u_{t-l} = u_t$ donne :

$$Cov\left\{\sum_{s_{t-l,t}} \frac{u_{t-l}}{\pi_{t-l}}, \sum_{s_{t-l,t}} \frac{u_t}{\pi_t}\right\} = Cov\left\{\sum_{s_{t-l,t}} \frac{u_{t-l}}{\pi_{t-l,t}} o\,(t-l,t), \sum_{s_{t-l,t}} \frac{u_t}{\pi_{t-1,t}} o\,(t-l,t)\right\}$$

$$\cong \text{Deff}\frac{N^2}{no\,(t-l,t)} o\,(t-l,t)^2\, Cov\,(u_{t-l}, u_t) = \text{Deff}\frac{N^2}{n} o\,(t-l,t)\, Cov\,(u_{t-l}, u_t)$$

- justification avec un renouvellement plus réaliste :

En notant $\Delta s_{t-l/t}$, $s_{t-l,t}$ et $\Delta s_{t/t-l}$ les échantillons sortant, pérenne et entrant entre les dates $t-l$ et t, les tailles des échantillons renouvelés valent $no\,(t-l,t)$ et la taille de l'échantillon pérenne vaut $n(1-o\,(t-l,t))$. On suppose que $\Delta s_{t-l/t}$ est tiré par un SAS, $s_{t-l,t}$ est tiré par un SAS dans le complémentaire de $\Delta s_{t-l/t}$ et $\Delta s_{t/t-l}$ est tiré par un SAS dans le complémentaire de s_{t-l}.

→ La covariance peut toujours se formuler à l'aide des probabilités d'inclusion simple et double dans les échantillons s_{t-l} et s_t selon (**eF.23**).

$$Cov\left(\widehat{U}_{t-l}, \widehat{U}_t\right) = \sum_{i,j} \frac{u_{t-l}}{\pi}(i) \frac{u_t}{\pi}(j)\left(\pi_{t-l,t}\,(i,j) - \pi\,(i)\,\pi\,(j)\right) \qquad \textbf{(eF.23)}$$

→ si $i = j$ alors $\pi_{t-l,t}\,(i,j) = \pi\,(i)\,o\,(t-l,t)$ (la probabilité de présence dans l'échantillon pérenne)

→ si $i \neq j$, comme les trois tirages sont des SAS conditionnels, $\pi_{t-l,t}(i,j)$ est constant lorsque la population est pérenne entre les deux dates [24]. Sous cette hypothèse supplémentaire, on obtient :

$$\sum_{i,j} \pi_{t-l,t}(i,j) = E\left(\sum s_{t-l} \sum s_t\right) = n^2$$

$$= N(N-1)\pi_{t-l,t}(i,j) + N\frac{n}{N} o(t-l,t)$$

$$\Longrightarrow \pi_{t-l,t}(i,j) = \frac{n^2 - no(t-l,t)}{N(N-1)}$$

$$= \frac{n}{N}\frac{n-o(t-l,t)}{N-1}$$

$\Longrightarrow$ Le calcul de la covariance entre les estimateurs de totaux s'en déduit :

$$Cov\left(\widehat{U}_{t-l}, \widehat{U}_t\right) = \sum_i \frac{u_{t-l}u_t}{\pi^2}(\pi o(t-l,t) - \pi^2) + \sum_{i\neq j} \frac{u_{t-l}}{\pi}(i)\frac{u_t}{\pi}(j)\left[\frac{n}{N}\frac{n-o(t-l,t)}{N-1} - \left(\frac{n}{N}\right)^2\right]$$

$$= \sum_i u_{t-l}u_t(i)\left[\frac{N}{n}o(t-l,t) - 1\right] + \sum_{i\neq j} u_{t-l}(i)u_t(j)\left[\frac{1}{N-1} - \frac{N}{n}\frac{1}{N-1}o(t-l,t)\right]$$

$$= \left[\frac{N}{n}o(t-l,t) - 1\right]\left\{\sum_i u_{t-l}u_t(i) - \frac{1}{N-1}\sum_{i\neq j} u_{t-l}(i)u_t(j)\right\}$$

$$= \left[\frac{N}{n}o(t-l,t) - 1\right]\left\{\sum_i u_{t-l}u_t(i)\frac{N}{N-1} - \frac{1}{N-1}\sum_{i,j} u_{t-l}(i)u_t(j)\right\}$$

$$= \left[\frac{N}{n}o(t-l,t) - 1\right]\frac{N^2}{N-1}Cov(u_{t-l}, u_t)$$

$\Longrightarrow$ Dans ce cadre, la covariance vérifie l'équation (**e II.4.15**).

Annexe G estimation longitudinale sur la période de transition entre les échantillonnages RP et TH

- L'expression (**e II.2.7**) reste valable lorsque les échantillons RP et TH sont simultanément présents pour l'une des deux dates, pourvu que la forme quadratique Q_1 en tienne compte, c'est-à-dire qu'elle soit définie sur $s_1^{RP} \cup s_1^{TH}$. L'indépendance entre les échantillons RP et TH implique que Q_1 est bloc-diagonale, ce qui correspond à sommer les covariances calculées sur chacune des sources :

$$\widehat{Y}_t^1 = \sum_{s_{1,t}^{RP}} \alpha_{RP}\frac{y_{1,t}}{\pi_1^{RP}} + \sum_{s_{1,t}^{TH}} \alpha_{TH}\frac{y_{1,t}}{\pi_1^{TH}} = \widehat{Y}_t^1\left(s_{1,t}^{RP}\right) + \widehat{Y}_t^1\left(s_{1,t}^{TH}\right)$$

$$Cov\left(\widehat{Y_{t-l}^1}, \widehat{Y}_t^1\right) = Cov\left\{\widehat{Y_{t-l}^1}\left(s_{1,t-l}^{RP}\right), \widehat{Y}_t^1\left(s_{1,t}^{RP}\right)\right\} + Cov\left\{\widehat{Y_{t-l}^1}\left(s_{1,t-l}^{TH}\right), \widehat{Y}_t^1\left(s_{1,t}^{TH}\right)\right\}$$

- Pour la période de transition, l'entrée et sortie de secteurs RP et TH est gérée en intégrant tous les secteurs échantillonnés des deux sources et en annulant la variable d'intérêt pour les observations manquantes ainsi générées. Ainsi, $\widehat{Y}_t^1\left(s_{1,t}^{RP}\right) = \sum_{s_{1,t}^{RP}} \alpha_{RP}\frac{y_{1,t}}{\pi_1^{RP}}$ peut être considéré comme l'estimateur du total d'une variable nulle sur les groupes de rotation RP sortants. La définition de ceux-ci est

24. Une permutation sur l'univers ne modifie pas la probabilité de tirage des trois sous-échantillons.

considérée ici ne pas dépendre de l'échantillon s_1^{RP} [25].

note : Pour l'échantillonnage RP, la formule de calcul de la covariance (**e II.2.7**) se transcrit (**eG.24**).

$$y_{t-l}{}' Q_1^{RP} y_t = \frac{n_1}{n_1 - 1} \left\{ \sum_{s_1} \frac{y_{t-l}}{\pi_1} \frac{y_t}{\pi_1} (1 - \pi_1) - \frac{\sum_{s_1} \frac{y_{t-l}}{\pi_1} (1 - \pi_1) \sum_{s_1} \frac{y_t}{\pi_1} (1 - \pi_1)}{\sum_{s_1} (1 - \pi_1)} \right\} \qquad \textbf{(eG.24)}$$

25. Ce n'est pas rigoureusement vrai, puisque la répartition en 6 groupes de rotation a été effectuée conditionnellement à l'échantillon de secteurs.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS – K.MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS – K.MOUSSALAM