PONDÉRATIONS LONGITUDINALES DANS L'ENQUÊTE EMPLOI DE L'INSEE

Pascal Ardilly

Insee, Département des méthodes statistiques

Contexte et objectifs

Source
Enquête Emploi trimestrielle en France

Objectif
Sur une période donnée, estimer les transitions entre les 3 états d'activité BIT (matrice 3 X 3) :

actif occupé / chômeur / inactif

Nécessité d'un échantillon de personnes physiques dit '*longitudinal*' - c'est-à-dire interrogé aux 2 dates délimitant la période choisie.

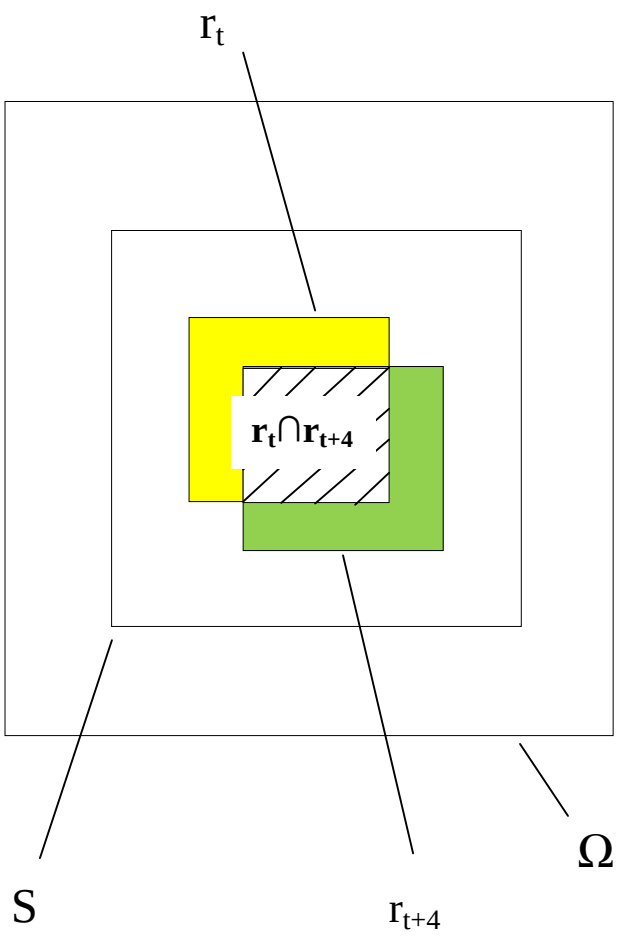
Deux types d'exigence :

- *technique :* poids sans biais et faible variance (autant que possible).

- *communication :* cohérence avec certaines statistiques transversales de début et de fin de période (au minimum les effectifs selon l'activité)

Eléments de méthode

Période retenue = 1 an (par exemple).



$r_{t+4}^{t} = r_t \cap r_{t+4}$ (hachuré) : échantillon *longitudinal* d'individus.

r_t n'est pas identique à r_{t+4} car :

- changements de ménage (déménagements)
- modifications de périmètre des ménages
- non-réponse due à l'érosion (lassitude)
- plus quelques non-réponses 'de circonstance'.

Approche *transversale* à la date t :

$\rightarrow$ poids de sondage de l'individu i $= w_i^{(1)}$

$\rightarrow$ poids après correction de non-réponse à t $= w_i^{(2)}$

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

Estimations des évolutions de statut entre t et $t+4$: variables individuelles $\Delta_i^{t/t+4}$ (indicatrices d'évolution).

$\rightarrow$ poids « *longitudinal* » individuel w_i^L

$\rightarrow$ estimateur *longitudinal* : $\sum_{i \in r^t_{t+4}} w_i^L \cdot \Delta_i^{t/t+4}$

Contraintes associées aux exigences

Respecter les estimations *transversales* nationales $\hat{Z}^{t}_{TRANS}$ de chacun des effectifs des 3 états d'activité pour chacune des 2 dates limitant la période, soit

$$\sum_{i \in r^{t}_{t+4}} w^{L}_{i} \cdot Z^{t}_{i} = \hat{Z}^{t}_{TRANS} \quad \text{et} \quad \sum_{i \in r^{t}_{t+4}} w^{L}_{i} \cdot Z^{t+4}_{i} = \hat{Z}^{t+4}_{TRANS}$$

En option : étendre ces contraintes à un « certain nombre » d'autres variables *collectées* X^{t}_{i}

$$\sum_{i \in r^{t}_{t+4}} w^{L}_{i} \cdot X^{t}_{i} = \hat{X}^{t}_{TRANS} \quad \text{et / ou} \quad \sum_{i \in r^{t}_{t+4}} w^{L}_{i} \cdot X^{t+4}_{i} = \hat{X}^{t+4}_{TRANS}$$

Très grand choix de variables à ce niveau (toutes les variables de collecte…).

→ le respect des contraintes va imposer la **mise en œuvre d'un calage** (avec %calmar).

Normalisations préalables au calage

Impératif : adapter au préalable certains effectifs sur lesquels on cale.

$\sum_{i \in r_{t+4}^{t}} w_i^L$ = taille de la population d'inférence

= effectif national estimé pour toute variable qualitative participant aux marges

Mise en cohérence des marges : par convention on adapte seulement les marges $\tilde{N}_{t+4}^{k}$ de **fin** de période.

- conservation des rapports entre les effectifs relatifs aux différentes modalités ;

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

et

- taille de population globale en *fin* de période = taille de population globale en *début* de période.

<u>Solution</u> :

'Règle de trois' affectant toute variable qualitative :

$$\tilde{N}_{t+4}^{k} = \hat{N}_{t+4}^{k} \cdot \frac{\hat{N}_t}{\hat{N}_{t+4}} = \text{nouvelle marge à } t+4$$

Mise en œuvre du calage

L'opération de calage fait intervenir trois éléments :

i) l'échantillon longitudinal r_{t+4}^{t} ;

ii) les variables de calage $\left(Z_i^t, X_i^t\right)$ et $\left(Z_i^{t+4}, X_i^{t+4}\right)$;

iii) les poids initiaux.

Rappel : $r_{t+4}^{t} \subset r_t$

Plusieurs scénarios pour obtenir w_i^L - mais 2 privilégiés :

* *Scénario 1*

- partir des poids transversaux $w_i^{(2)}$ pondérant r_t;
- **calage en 1 étape,** à partir de r_{t+4}^{t}.

$\rightarrow$ corrige la non-réponse à $t+4$ sachant qu'on a répondu à t.

* *Scénario 2*

- correction préalable des poids $w_i^{(2)}$ par une probabilité de réponse estimée

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

- **calage standard**.

- *A quel niveau s'effectue le calage ?*

En approche *transversale*, calage au niveau *ménage*.

En approche *longitudinale*, au contraire optique fondamentalement *individuelle* :

- un ménage est instable dans le temps
- la non-réponse due au décalage $t / t+4$ a une forte composante individuelle

$\rightarrow$ *in fine*, chaque individu physique a un poids qui lui est propre.

- *Quelle stratégie retenir ?*

Scénario 2 théoriquement plus satisfaisant mais ce n'est pas la méthode de repondération de l'enquête Emploi.

Scénario 1 apparaît néanmoins *a priori* comme un bon compromis entre simplicité, communication et efficacité.

Dans les 2 scénarios, le choix des variables explicatives des probabilités de réponse individuelles est essentiel car les individus qui quittent leur logement ont des transitions spécifiques : ***a priori*** **contexte de non-réponse non-ignorable**.

Deux questions difficiles

1) Tout individu non-répondant est-il dans le champ de l'enquête (perturbe le scénario 2) ?

2) Quelle est la population d'inférence ?

$\rightarrow$ c'est bien confus (à cause du calage) : il vaut donc mieux s'en tenir à traiter des périodes courtes …

Application à la période annuelle
T1 2014 - T1 2015

Filtrage préalable des individus de 15 à 74 ans.

* **Avec le scénario 1**, différentes options de choix des variables de calage :

a) structures d'activité aux dates respectives de *début* **et** de *fin* de période (= version minimale !);

b) *ajout* de 2 variables caractérisant le parcours de l'individu d'après son activité mensuelle sur une année complète (selon sa déclaration 'spontanée');

c) *nouvel ajout* de variables sociodémographiques :
- sexe
- âge (en tranches)
- diplôme
- nationalité
- catégorie sociale
- type de ménage
- tranche d'unité urbaine
- appartenance à une ZUS
- deux variables complexes *ad hoc* liées à l'activité, construites à partir de la nature de l'employeur, du type de contrat et de l'ancienneté au chômage.

Certaines variables sont exploitées à la fois aux dates de début et de fin de période.

* **Avec le scénario 2 :**

d) estimation préalable des probabilités de réponse par une régression logistique, avec sélection *stepwise* des régresseurs ;

régresseurs du c) mais (évidemment) uniquement à la date t de début de période ;

une version pondérée / une version non pondérée.

Marges de calage : exactement comme au c)

Eléments de volumétrie :

- Avant filtrage sur l'âge :

Taille de r_t = 32 700 individus
Taille de r_{t+4}^t = 27 200 individus

- Après filtrage sur l'âge :

Taille de r_{t+4}^t = 23 807 personnes.

Ventilation de la population filtrée entre les 9 types de transitions (en %)

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

Référence = utilisation du poids transversal $w_i^{(2)}$.

S*tabilité* = diagonale de la matrice des transitions.

Ventilation de la population filtrée selon le degré de stabilité de l'activité (en %)

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

Conclusion

- l’ajout de marges augmente significativement la proportion d’individus dont l’activité BIT change ;

- l’échantillon longitudinal contient trop d’individus ‘stables’ géographiquement (la non-réponse agit en ce sens !) et ***donc*** ‘stables’ dans leur activité;

- « décrochage » surtout au niveau de la méthode b) ;

- mais la dispersion des poids augmente avec le nombre de marges.

Nota : la méthode finalement appliquée par l’Insee relève d’options un peu différentes.