

**Echantillonnage et pondération dans  
les échantillons rotatifs : le cas de  
l'enquête européenne SILC  
sur le revenu et les conditions de vie**

Pascal Ardilly  
INSEE

et

Pierre Lavallée  
Statistique Canada

# Introduction

Cette enquête européenne sur le revenu et les conditions de vie (« *European Survey on Income and Living Conditions* » - *SILC* - ou encore « *Enquête sur les Revenus et les Conditions de Vie* » - *ERCV*) a remplacé le Panel communautaire à partir de 2004.

Elle produit en particulier des statistiques comparatives sur :

- la répartition des revenus
- le niveau et la nature de la pauvreté et de l'exclusion sociale

C'est une enquête concernant essentiellement les **individus physiques**.

# Principes généraux

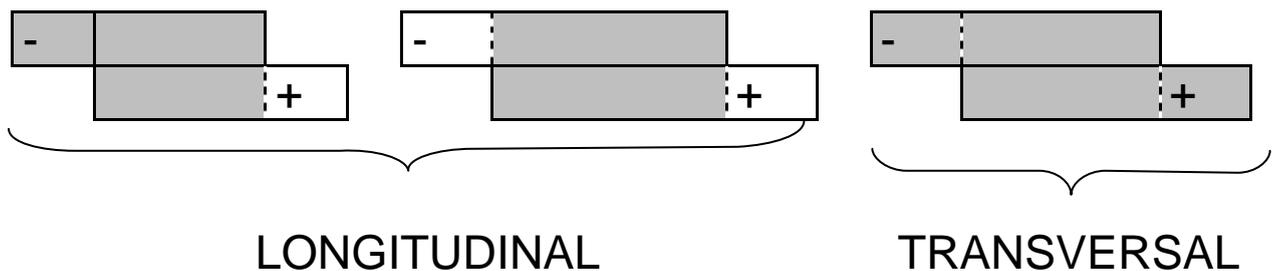
## 1) Approches longitudinale et transversale

- On veut mesurer un total une année  $t$  donnée ET des évolutions entre 2 années  $t-1$  et  $t$  ;
- On a 2 façons de considérer la population d'inférence :

1/ C'est une population fixe dans le temps  
⇒ approche **longitudinale**

**OU**

2/ C'est une population évolutive  
⇒ approche **transversale**



- = MORTS (décès, émigration, passage en collectivité,...)

+ = NAISSANCES (nouveau-nés, immigration, entrée dans le champ, ...)

L'approche longitudinale n'a pas de sens pour les ménages !

Problème : on veut une méthode d'échantillonnage qui permette des estimations selon ces 2 approches.

⇒ 3 stratégies principales

- *Scénario 1*: ECHANTILLON INDEPENDANT chaque année ;
- *Scénario 2* : échantillon tiré une année donnée  $t_0$ , et suivi dans le temps = PANEL ;
- *Scénario 3* : échantillon partiellement renouvelé chaque année  $t$  = ÉCHANTILLON ROTATIF.

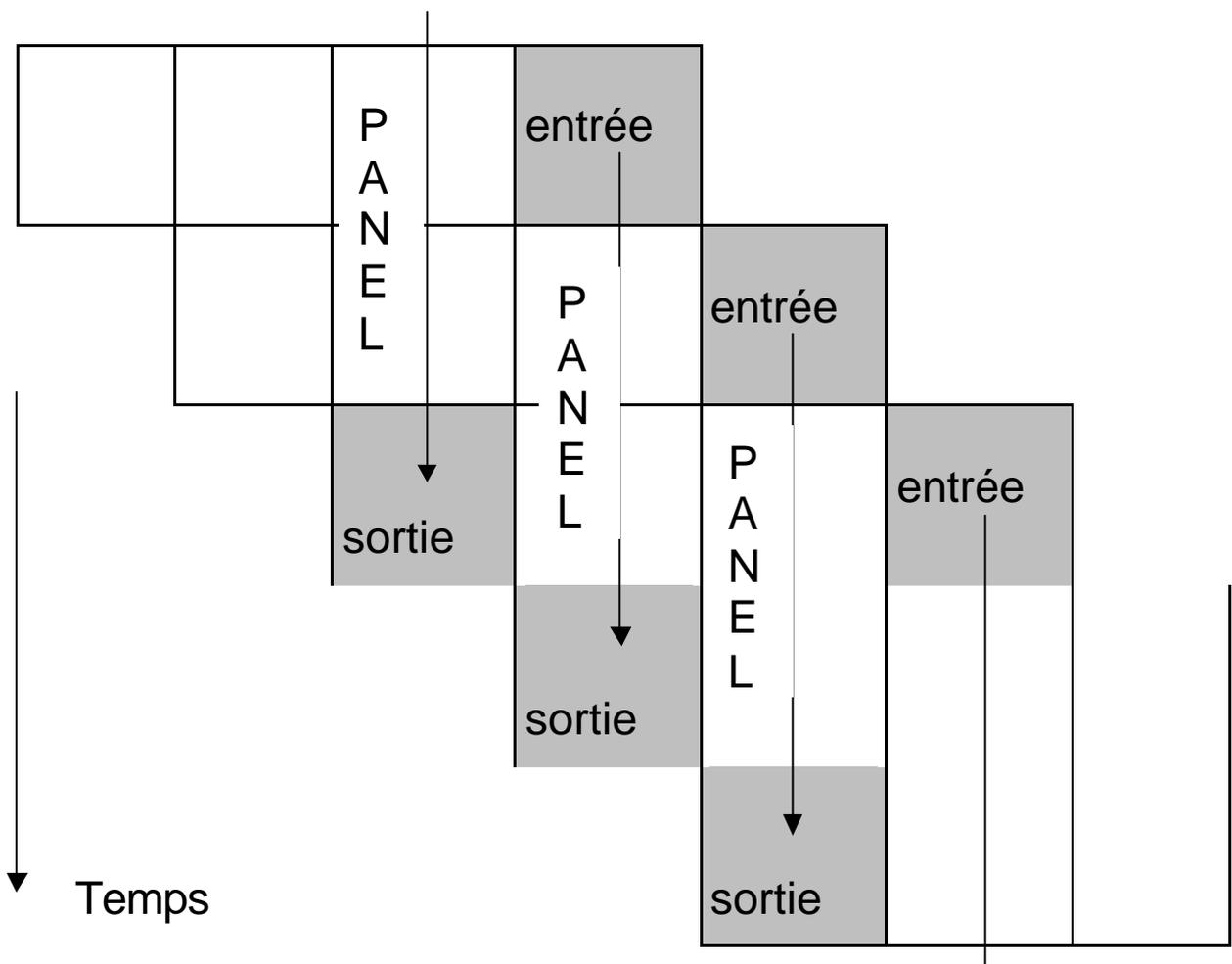
TYPE d'échantillon	Approche TRANSVERSALE	Approche LONGITUDINALE
Indépendant	NATUREL	POSSIBLE mais moins efficace
Panel	IMPOSSIBLE sans tirage complémentaire	NATUREL
Rotatif	POSSIBLE	POSSIBLE

## 2) Principe, avantages et inconvénients de l'échantillonnage rotatif

C'est une juxtaposition de PANELS d'individus :

- à durée limitée (4 ans)
- décalés dans le temps (1 an de décalage)
- logements tirés dans l'échantillon-maitre chaque année et dans les mêmes conditions, à la mise à jour de la base de logements près (BSLN)

**Dans chaque logement tiré, ON SUIT TOUS LES INDIVIDUS AU COURS DU TEMPS**



Cette technique :

- 1) a les avantages traditionnels du panel :  
réduction des erreurs d'observation et des erreurs d'échantillonnage pour estimer des évolutions;
- 2) présente la difficulté traditionnelle des panels d'individus, à savoir le pistage (**suivi au cours du temps**) ;
- 3) du fait de la rotation :
  - limite la charge des enquêtés par rapport à un panel (moins d'effets de lassitude) ;
  - permet une prise en compte « naturelle » dans l'échantillon de toutes les formes d'évolution de la population, aussi bien dans l'approche longitudinale que dans l'approche transversale
  - En contrepartie, limite la « longitudinalité » des données par rapport à un panel.

# Pondération longitudinale

Unité d'observation : exclusivement l'individu physique, bien que l'on tire des ménages, via leur logement.

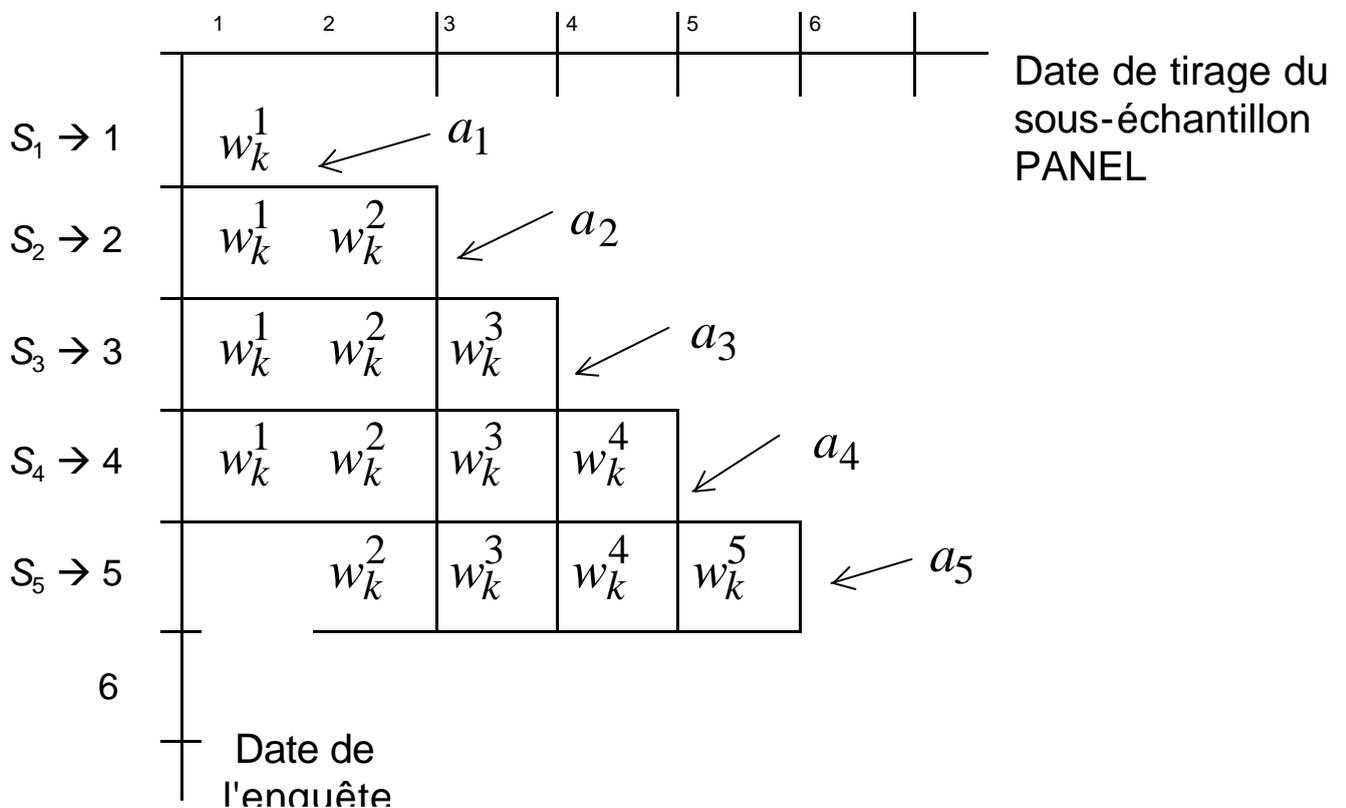
On supposera qu'il n'y a pas de non-réponse.

On note  $\Omega_t$  = population à la date t.

Objectif : estimer une évolution entre t et t+1

Optique longitudinale  $\Rightarrow$  la population d'inférence est  $\Omega_t$

## Configuration longitudinale en l'absence de non-réponse

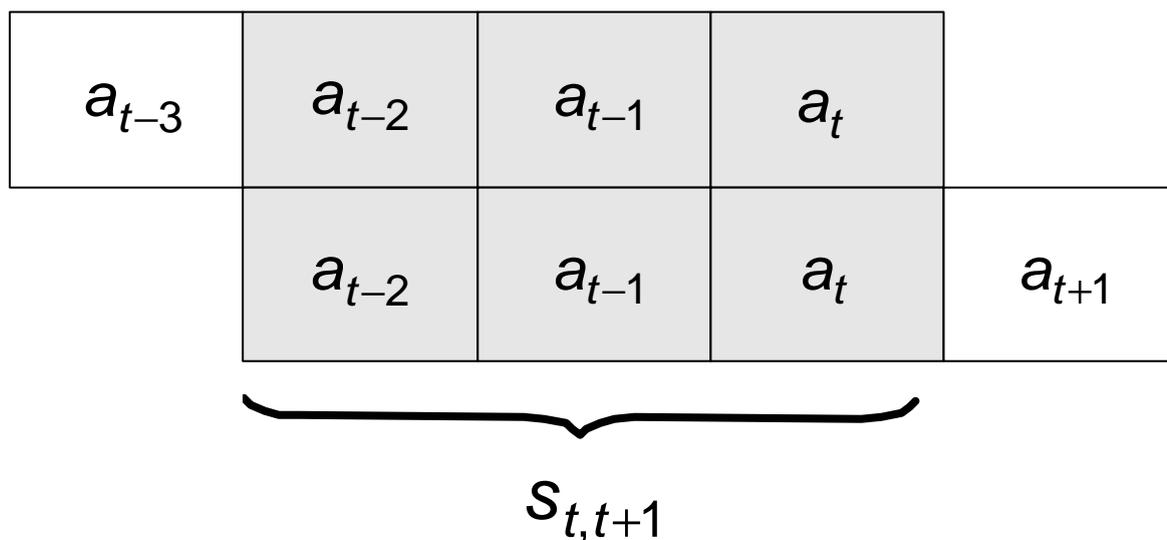


Poids **initiaux**  $w_k^a$  : déterminés, pour un sous-échantillon panel donné  $a_a$  pour représenter la population  $\Omega_a$  **à la date a du tirage** de ce sous-échantillon.

Entre les vagues  $t$  et  $t+1$ , l'échantillon opérationnel est

$$s_{t,t+1} = \bigcup_{a=t-2}^t a_a$$

*Rappel* :  $a_a$  représente  $\Omega_a$  et  $s_{t,t+1}$  représente  $\Omega_t$ .



A cause du schéma rotatif, un individu de  $\Omega_t$  **peut être tiré dans plusieurs panels**  $a_a$  : il faut donc en tenir compte dans la pondération.

Le **nombre** de panels dans lesquels il PEUT être tiré dépend de sa date d'entrée dans le champ de l'enquête.

## **Parenthèse sur la méthode de partage des poids**

- Deux populations  $\Omega_A$  et  $\Omega_B$  et un système de liens ;
- $\Omega_B$  est partitionnée en grappes "naturelles" (exemple : des ménages)  $\Rightarrow$  grappe  $i$ , individu  $k$  ;

- Principe en 2 temps :

- Echantillonnage  $s_A$  dans  $\Omega_A$  (unités d'échantillonnage  $j$ ) et prise en compte des liens  $\mathbf{1}_{j \rightarrow ik} \Rightarrow$  échantillon dans  $\Omega_B$ .

- On enquête l'intégralité des grappes recoupant cet échantillon  $\Rightarrow$  échantillon de grappes  $s_B$ .

**Question : comment pondérer SANS BIAIS les unités d'observation  $i,k$  enquêtées ?**

$$T = \sum_{ik \in \Omega_B} Y_{ik}$$

et 
$$\hat{T} = \sum_{i \in s_B} \sum_{k \in i} w_{ik} Y_{ik}$$

1) Calculer  $\forall i \in s_B, \forall k \in i$  : 
$$w'_{ik} = \sum_{\substack{j \in s_A \\ j \rightarrow ik}} w_j$$

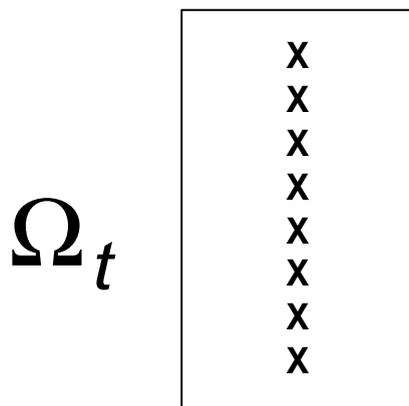
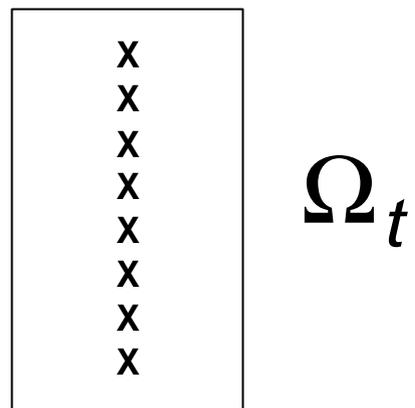
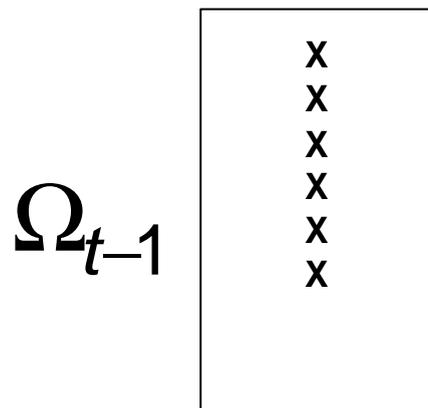
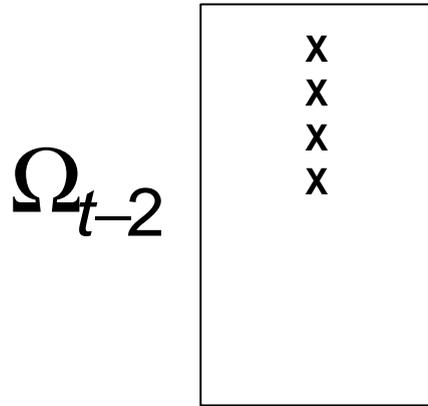
2) Calculer  $\forall i \in s_B$ , le nombre TOTAL de liens  $L_i$

$$L_i = \sum_{k \in i} \sum_{j \in \Omega_A} \mathbf{1}_{j \rightarrow ik}$$

3) Former 
$$w_i = \frac{1}{L_i} \sum_{k \in i} w'_{ik}$$

4) Attribuer à chaque  $k$  de  $i$  le poids  $w_{ik} = w_i$

On peut représenter ainsi la situation, et appliquer la méthode de partage des poids :



Nombre de liens de l'individu  $k$  (dans un ménage  $i$ ):

où  $L_k =$  nombre d'années parmi  $\{t-2, t-1, t\}$  au cours desquelles  $k$  était présent dans la population du champ de l'enquête. ( $L_k = 1, 2,$  ou  $3$ )

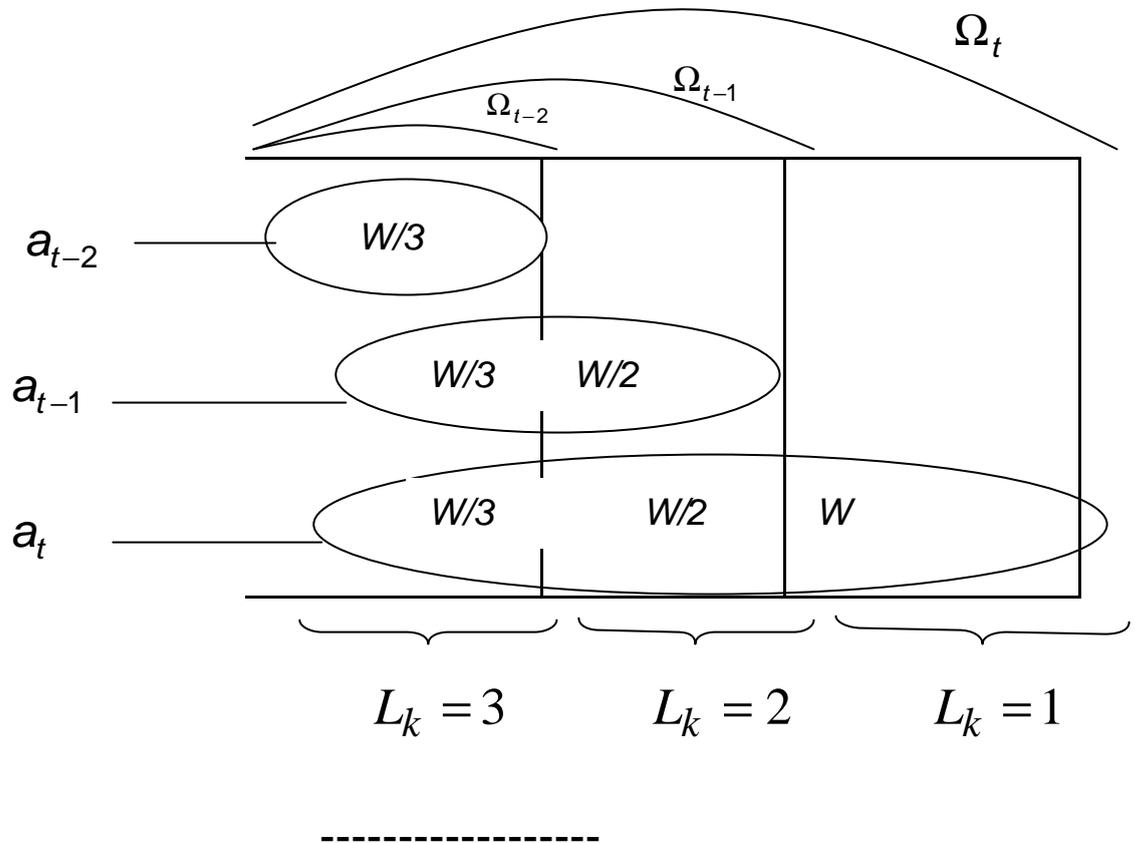
Si on néglige la probabilité de tirer  $k$  dans 2 échantillons  $a_a$  différents, alors **si  $k$  provient de  $a_a$**  :

$$w_k^{t,t+1} = \frac{w_k^a}{L_k}$$

Cas « idéal » : pas de naissance (en régime stationnaire) :

$$w_k^{t,t+1} = w/3, \quad \forall k$$

Cas plus réaliste : naissances, régime stationnaire



Finalemment si

$$\Delta = \sum_{\Omega_t} Y_k^{t+1} - \sum_{\Omega_t} Y_k^t$$

$$\Rightarrow \hat{\Delta}_{t,t+1} = \sum_{k \in S_{t,t+1}} w_k^{t,t+1} \times (Y_k^{t+1} - Y_k^t)$$

# Pondération transversale

L'extrapolation porte sur la population à la date courante. Pour un sous-échantillon donné, il manque :

- les nouveau-nés
- les immigrants (au sens large : tout ce qui n'est pas nouveau-né)



Il faut un échantillon COMPLEMENTAIRE  
à l'échantillon panel.

PRINCIPE : on va enquêter **tout** le ménage comprenant au moins un individu panel ⇒ cela forme naturellement l'échantillon complémentaire - mais il reste une faiblesse !

?

# 1) La solution offerte par le partage des poids à partir d'un panel

## a) Contexte :

- Population "initiale"  $\Omega_{t_0}$  d'unités d'échantillonnage à  $t_0$  (base de sondage = individus "panel" potentiels).
- Population d'inférence  $\Omega_t$  à  $t$  divisée en grappes (grappe = ménage).

## b) Principes :

- 1) On tire à  $t_0$  des individus panel  $j \Rightarrow$  poids  $w_j^0$
- 2) On enquête à  $t$  tout le ménage  $i$  si et seulement si il contient au moins un individu panel.

Unités « transversales » = individus panel + cohabitants enquêtés
---

Puis application directe du partage des poids :

On pose  $s_t = \bigcup_{a=t-3}^t a_a$  = ensemble des individus panel à t.

Pour chaque ménage enquêté  $i$ , on calcule un poids au niveau « ménage » :

$$w_i^t = \frac{\sum_{\substack{k \in i \\ k \in \Omega_{t0} \\ k \in s_t}} w_{ik}^0}{\sum_{\substack{k \in i \\ k \in \Omega_{t0}}} 1}$$

Affecter in fine cette valeur à chaque individu  $k$  du ménage  $i$ , soit

$$w_{TR, ik}^t = w_i^t.$$

### c) Où est la faiblesse ?

Les ménages sans AUCUN lien ne sont pas représentés.

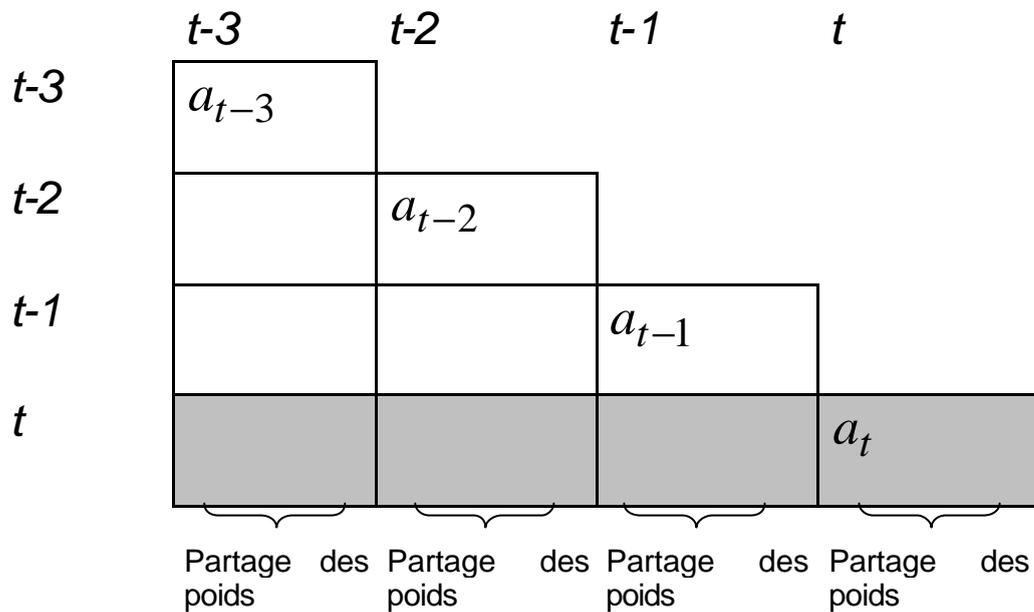
C'est le problème des ménages composés seulement d'immigrants  $\Rightarrow$  BIAIS.

## 2) Adaptation au schéma rotatif

Le schéma rotatif va lever la faiblesse :

Echantillon transversal rafraîchi chaque année  
⇒ on peut représenter les ménages  
composés seulement d'immigrants.

Solution *la plus simple* : appliquer la méthode du partage des poids sous-échantillon  $a_a$  par sous-échantillon.



- Population d'inférence pour chaque  $a_a$  : population complète  $\Omega_t$  à la date  $t$  de l'enquête.
- Chaque ménage enquêté contient au moins un individu panel (provenant de  $a_a$ ,  $a = t-3, \dots, t$ ), et il peut aussi contenir des cohabitants .
- Chaque ménage enquêté à  $t$  a **au moins un lien** avec la population  $\Omega_a$  - sauf s'il est constitué uniquement d'immigrants entre les dates  $a + 1$  et  $t$ .

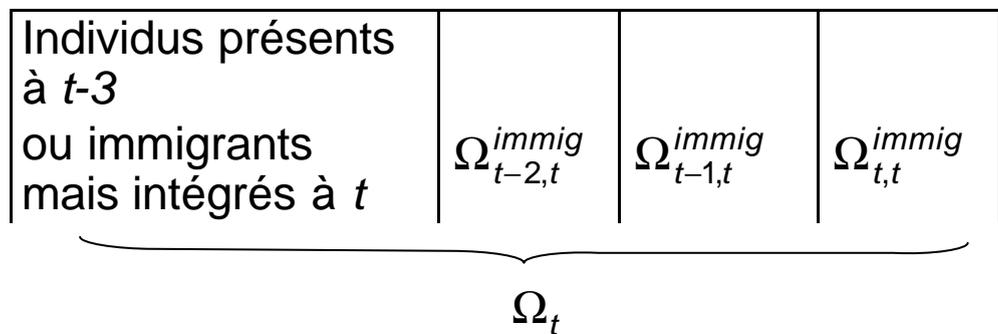
L'enquête a lieu l'année  $t$ . On note :

$\Omega_{a,t}^{immig}$  : Population d'immigrants à la date  $a$ , présents à  $t$  dans un ménage ne comprenant QUE des immigrants arrivés à une date égale ou postérieure à  $a$  ( $t-3 \leq a \leq t$ ).

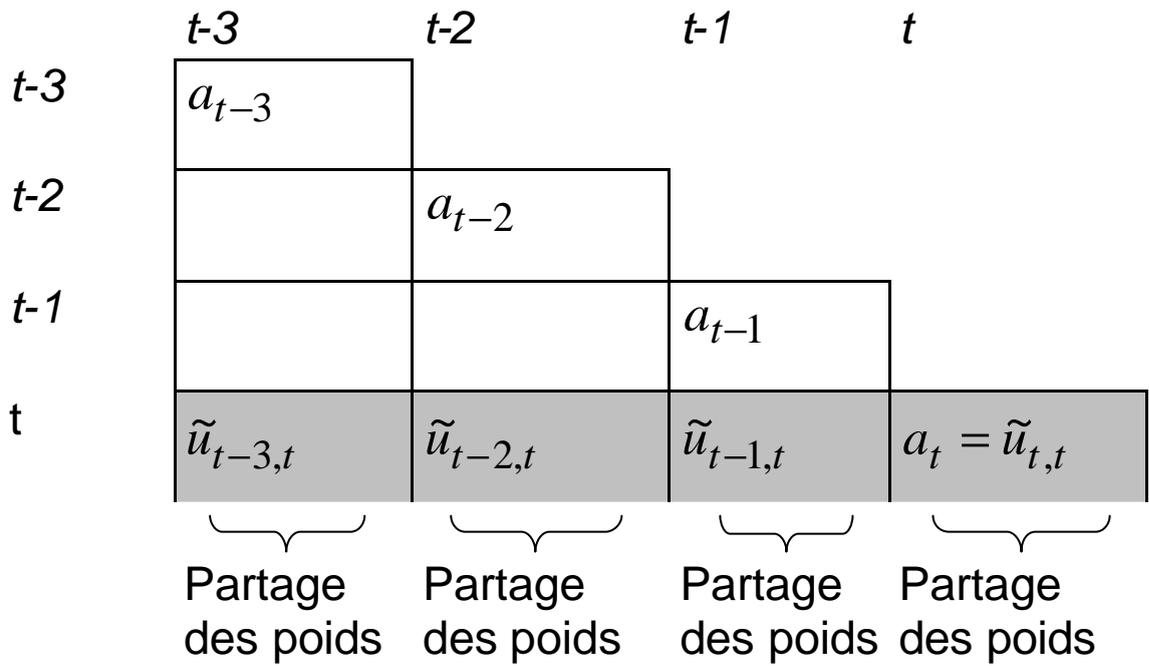
$Y^t$  : Vrai total des  $Y_k^t$  sur  $\Omega_t$ .

$Y_{a,t}^{immig}$  : Vrai total des  $Y_k^t$  sur  $\Omega_{a,t}^{immig}$ .

Un individu  $k$  immigrant en  $a$  est "intégré" à  $t$  s'il n'est pas dans  $\Omega_{a,t}^{immig}$ .



$\tilde{u}_{a,t}$  : Échantillon transversal à la date  $t$  associé à  $a_a$ .



$$\begin{aligned}
 \sum_{\tilde{u}_{t-3,t}} w_{TR,k}^{t-3,t} \cdot Y_k^t &\xrightarrow{\text{estime}} Y^t - Y_{t-2,t}^{immig} - Y_{t-1,t}^{immig} - Y_{t,t}^{immig} \\
 \sum_{\tilde{u}_{t-2,t}} w_{TR,k}^{t-2,t} \cdot Y_k^t &\xrightarrow{\text{estime}} Y^t - Y_{t-1,t}^{immig} - Y_{t,t}^{immig} \\
 \sum_{\tilde{u}_{t-1,t}} w_{TR,k}^{t-1,t} \cdot Y_k^t &\xrightarrow{\text{estime}} Y^t - Y_{t,t}^{immig} \\
 \sum_{u_t} w_{TR,k}^{t,t} \cdot Y_k^t &\xrightarrow{\text{estime}} Y^t
 \end{aligned}$$

a) Scénario\_1 (simple mais approximatif) :

On "néglige" les  $\Omega_{a,t}^{immig}$ ,  $a = t-2, t-1, t$  devant  $\Omega_t$

$$\Rightarrow \sum_{\tilde{u}_t} \frac{w_{TR,k}^t}{4} \times Y_k^t \text{ estime "à peu près" sans biais } Y^t$$

où  $\tilde{u}_t = \tilde{u}_{t-3,t} \cup \tilde{u}_{t-2,t} \cup \tilde{u}_{t-1,t} \cup \tilde{u}_{t,t}$   
= échantillon transversal

b) Scénario 2 (plus compliqué, mais plus rigoureux)

Prendre en compte les immigrants : on peut montrer que le jeu de poids suivant convient pour chaque unité tirée  $k$  :

$$\left\{ \begin{array}{ll} w_{TR,k}^t & \text{si } k \text{ est dans } \Omega_{t,t}^{immig} \\ w_{TR,k}^t / 2 & \text{si } k \text{ est dans } \Omega_{t-1,t}^{immig} \\ w_{TR,k}^t / 3 & \text{si } k \text{ est dans } \Omega_{t-2,t}^{immig} \\ w_{TR,k}^t / 4 & \text{dans tous les autres cas} \end{array} \right.$$

où  $w_{TR,k}^t = w_{TR,k}^{a,t}$  pour  $k \in \bar{u}_{a,t}$ .

Pour repérer l'appartenance d'un individu aux  $\Omega_{a,t}^{immig}$ , il faut prévoir une question individuelle sur la première année de présence dans un logement tirable.

-----

Finalement, on utilise

$$\hat{Y}_t = \sum_{\substack{k \text{ enquêtés} \\ \text{à } t}} w_{TR,k}^t \cdot Y_k^t$$

et

$$\hat{\Delta}_{t-1,t} = \sum_{\substack{k \text{ enquêtés} \\ \text{à } t}} w_{TR,k}^t \cdot Y_k^t - \sum_{\substack{k \text{ enquêtés} \\ \text{à } t-1}} w_{TR,k}^{t-1} \cdot Y_k^{t-1}$$

# Conclusion

- Il est possible, grâce au **schéma rotatif**, de satisfaire à la fois les attentes longitudinales et transversales ;
- La méthode de partage des poids est un outil central ;
- L'ensemble du traitement s'avère néanmoins assez compliqué ;
- Il faut adapter ce schéma en modifiant les poids du fait :
  - du traitement de la non-réponse totale ;
  - des redressements ;

ce que l'on sait faire (ça complique sensiblement !)

- Les calculs de variance deviennent vraiment difficiles.

\*\*\*\*\*