Pondérations longitudinales et transversales dans les échantillons rotatifs

* * *

Application à l'enquête SILC

Pascal ARDILLY, INSEE

PLAN

1/ Vision longitudinale et vision transversale

2/ Un outil : le partage des poids

3/ La pondération longitudinale dans le cas de SILC

4/ La pondération transversale dans le cas de SILC

Vision longitudinale et vision transversale

Enquête répétée dans le temps $\rightarrow$ collecte de Y_i^t

Ω_t = population du champ de l'enquête l'année t

Quels sont les paramètres intéressants ?

Vision transversale :

$T_t = \sum_{i \in \Omega_t} Y_i^t$ (+ tous les "satellites")

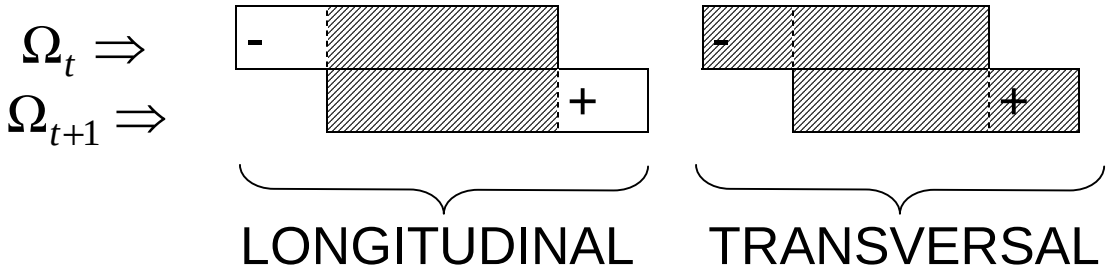
$\Delta^*_{t,t+1} = T_{t+1} - T_t = \sum_{i \in \Omega_{t+1}} Y_i^{t+1} - \sum_{i \in \Omega_t} Y_i^t$

Vision longitudinale :

$\Delta_{t,t+1} = \sum_{i \in \Omega_{t+1} \cap \Omega_t} Y_i^{t+1} - \sum_{i \in \Omega_{t+1} \cap \Omega_t} Y_i^t$

$\rightarrow$ élimine la part d'évolution expliquée par la démographie (interprétation complexe)

Deux notions de population d'inférence



Les "morts" : décès, émigration, passage de l'individu en communauté, sortie de champ,…

Les "naissances" : nouveau-nés, immigration, entrée dans le champ par le franchissement d'un seuil d'âge,…

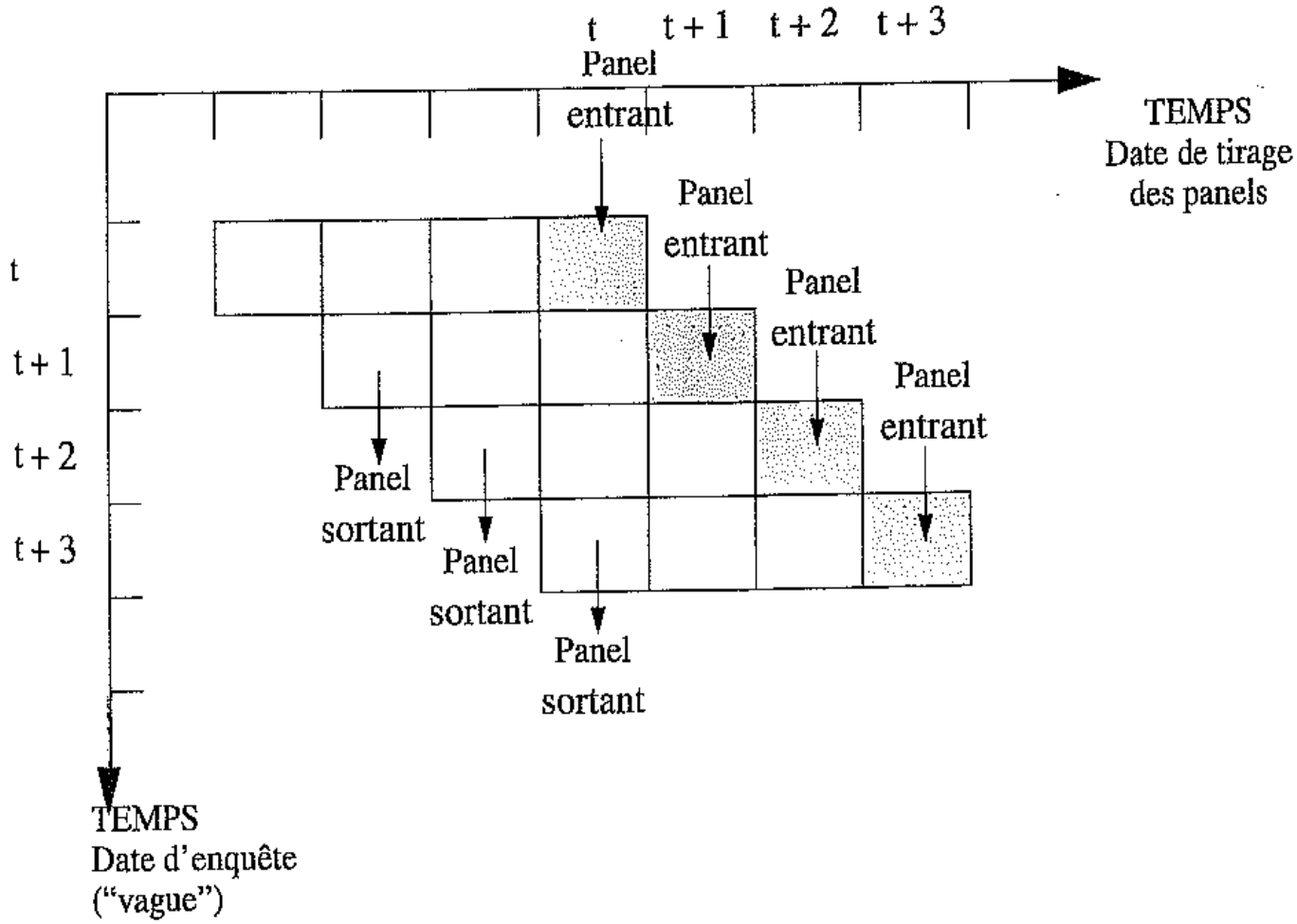
Stratégies d'échantillonnage envisageables

Objectif : produire **à la fois** des estimations longitudinales et des estimations transversales.

1. Un échantillonnage « indépendant » chaque année : largement perfectible en terme de précision des évolutions.

2. Un échantillonnage intégralement panélisé : problème de charge, irréaliste en pratique.

3. Un échantillon rotatif : bon compromis précision / charge d'enquête.



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

Avantages de l'échantillonnage rotatif

i. réduit l’erreur d’échantillonnage associée à la mesure des évolutions (bien que moins efficace en théorie que le panel « pur »).

ii. limite la charge des enquêtés par rapport au panel « pur ».

iii. permet de prendre en compte d’une manière très « naturelle » l’évolution de la population avec le temps.

iv. plus accessoirement, permet de réduire les erreurs d’observation (comme les panels).

Défauts de l'échantillonnage rotatif

i. nécessite un suivi des individus dans le temps, ce qui occasionne des coûts de dépistage et des non-réponses du fait des personnes non retrouvées (déménagements…).

ii. par nature, la longueur des séries individuelles est limitée, donc évidemment moins riche qu'un pur panel.

iii. la technique de pondération longitudinale / transversale n'est pas simple …

Un outil : le partage des poids

Estimation en contexte "simple" :

- 1 population Ω (base de sondage)
- 1 échantillon s (tiré dans Ω)

On connaît $P_i = \mathrm{Proba}(i \in s)$

$T = \sum_{i \in \Omega} Y_i$ estimé par $\hat{T} = \sum_{i \in s} \frac{Y_i}{P_i}$ (Horvitz-Thompson)

a) $E\hat{T} = T$ (estimateur sans biais)

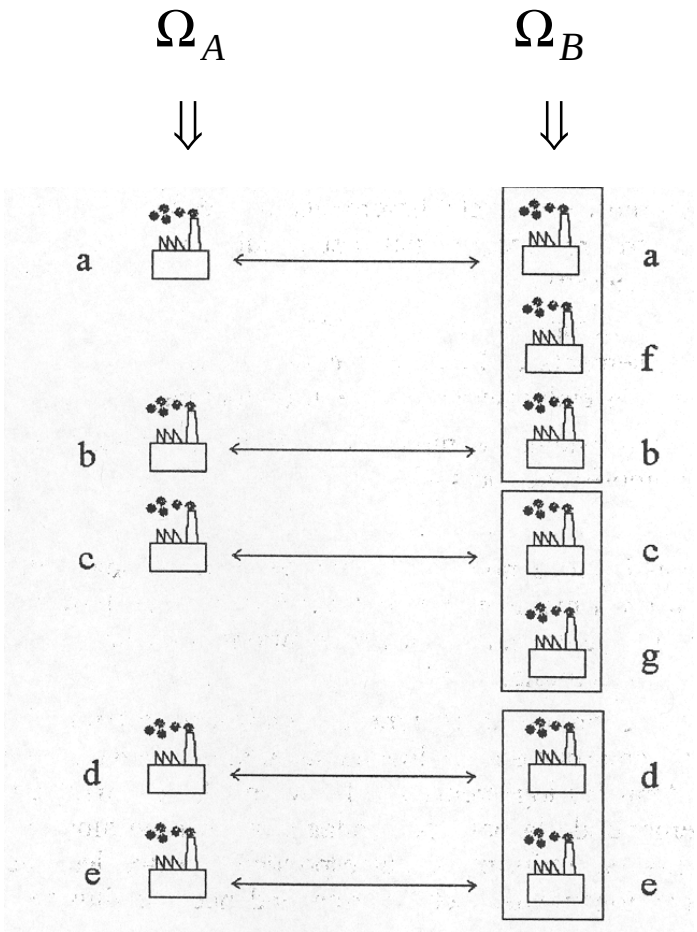
b) $V\hat{T} = \frac{1}{2} \sum_{\substack{i \neq j \\ (i,j) \in \Omega^2}} \sum \left(P_i . P_j - P_{i,j} \right) . \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$

Estimation dans le cadre de l'échantillonnage indirect

- Deux populations Ω_A et Ω_B, plus un système de LIENS permettant un passage de Ω_A vers Ω_B;

- Des GRAPPES "naturelles" dans Ω_B (ex : des ménages, des entreprises) ;

  $\rightarrow$ Grappe i, individu k $\rightarrow$ variable Y_{ik}



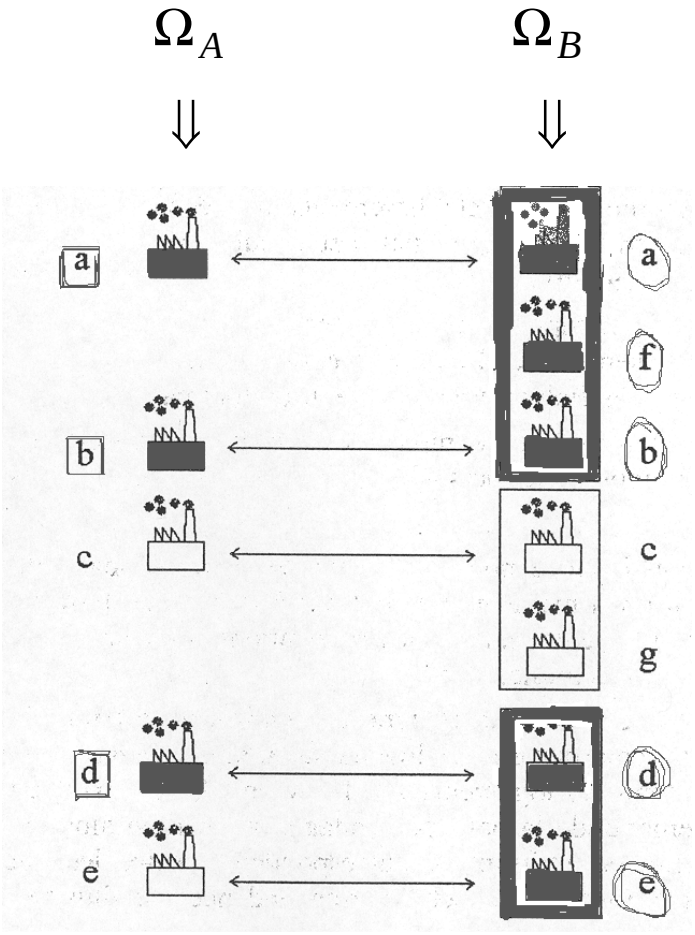
Principe d'échantillonnage en 2 temps :

a) échantillonnage s_A dans Ω_A (unités d'échantillonnage j) et prise en compte des liens $1_{j \to ik}$

$\Rightarrow$ échantillon intermédiaire dans Ω_B.

b) on décide d'enquêter l'intégralité des grappes recoupant cet échantillon

$\Rightarrow$ échantillon final de grappes s_B



Quel estimateur sans biais du total des $Y_{i,k}$ dans Ω_B ?

Solution 1 : estimateur classique de Horvitz-Thompson.

→ 2 problèmes :

1) proba $(i \in s_B)$ très compliquée

2) nécessité de connaître p $(j \in s_A)$ pour des individus j non tirés (ex : tirage en 2 phases)

Solution 2 : Méthode généralisée de partage des poids (MGPP)

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

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

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

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

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

Résumé opérationnel

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

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

Si on pose $L_{j,i} = \sum_{k \in i} \mathbf{1}_{j \to ik}$, on a le poids final

$$\boxed{w_{ik} = \sum_{j \in s_A} \frac{L_{j,i}}{L_i} . \frac{1}{P_j}}$$

AVANTAGES (déterminants !)

→ Pondération facile

→ Utilisation des P_j pour $j \in s_A$ seulement.

INCONVENIENTS

Risque d'erreur de mesure sur L_i

Si L_i SOUS-ESTIMÉ $\Rightarrow$ w_i devrait SUR ESTIMER la réalité !

$\Rightarrow$ une pseudo solution : un calage sur un total X^B

C'est "le" risque de la MGPP

PROPRIÉTÉS fondamentales

1) $E\hat{T} = T$ dès que $L_i > 0 \quad \forall i \in \Omega_B$

Attention : $L_i = 0 \quad \Rightarrow$ défaut de couverture

2) Dualité :

On pose $Z_j = \sum_{i \in \Omega_B} \frac{L_{j,i}}{L_i} . Y_i \qquad \left(L_{j,i} = \sum_{k \in i} \mathbf{1}_{j \to ik} \right)$

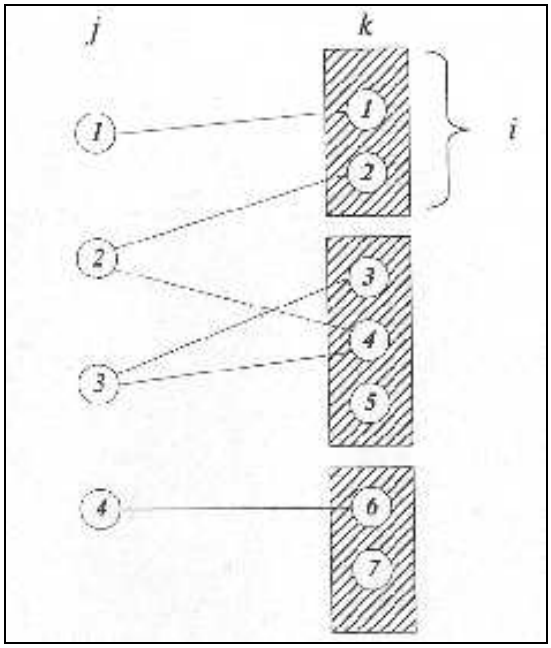
Alors $\hat{T} = \sum_{j \in s_A} \frac{1}{P_j} . Z_j = \hat{Z}$

$$\Rightarrow \qquad V(\hat{T}) = V(\hat{Z})$$

Pas de difficulté pour le calcul de précision.

A noter : la dualité permet aussi de traiter la non-réponse et de pratiquer des redressements de façon "propre".

Application numérique



$$s_A = \{1, 2\}$$
$$s_B = \{11, 12, 21, 22, 23\}$$

$$L_1 = 2 \qquad L_2 = 3$$

$$w'_{11} = \frac{1}{P_1} \qquad w'_{12} = \frac{1}{P_2}$$

$$w'_{21} = 0 \qquad w'_{22} = \frac{1}{P_2} \qquad w'_{23} = 0$$

$$w_1 = \frac{1}{2}\left[\frac{1}{P_1} + \frac{1}{P_2}\right] \quad \text{et} \quad w_2 = \frac{1}{3} \cdot \frac{1}{P_2}$$

$$\hat{T} = w_1 \left(Y_{11} + Y_{12}\right) + w_2 \left(Y_{21} + Y_{22} + Y_{23}\right)$$

Exemples concrets d'utilisation de la MGPP

1) **Absence de base de sondage** :

Enquête "Sans domicile fixe" INSEE-INED

Ω_A ={ centres d'accueil X prestations }

Ω_B = { personnes sans domicile fixe }

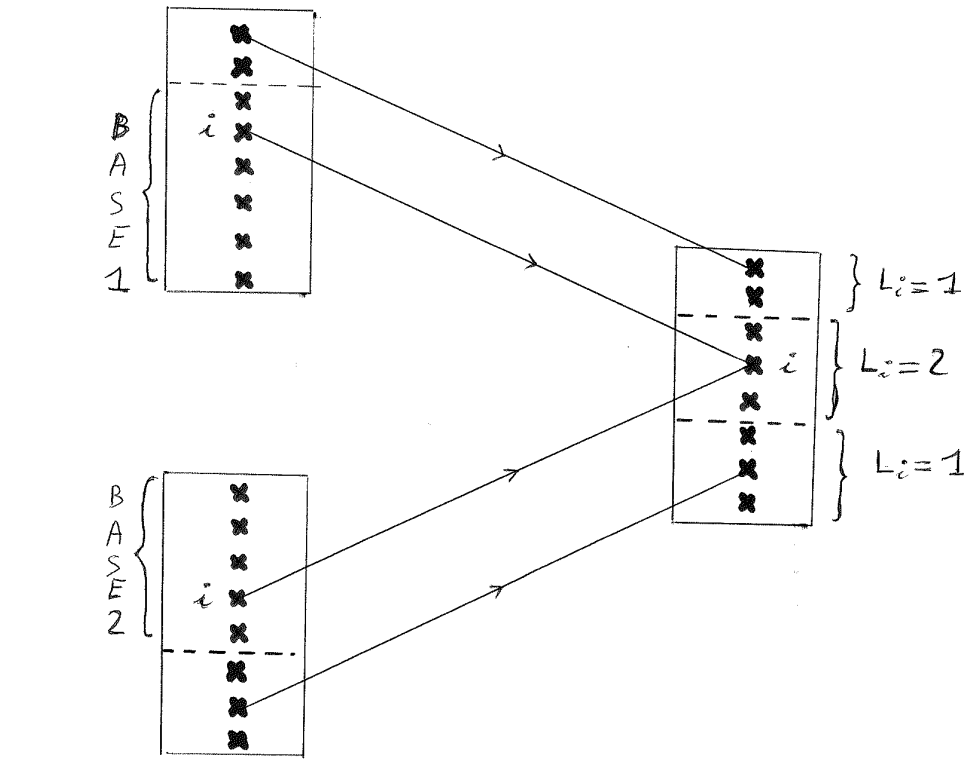
L_i = Nombre de prestations dans les centres fréquentés par i au cours de la collecte (collecté sur une semaine glissante + modèle pour passer au mois).

2) **Bases de sondage multiples**

L_i = Nombre de bases contenant i

(Attention : adapter le questionnaire !)

Exemple : enquête "Logement 2006"



3) **Approche par réseau** : étude d'une population rare.

"Citer un consommateur de drogue"

$\Omega_B \subset \Omega_A$: Taille de Ω_B $(Y_i = 1)$?

$L_i =$ Nombre de personnes qui citeront i
(problème de mesure !)

4) **Sondage adaptatif** : étude d'une population rare

$\Omega_A =$ { Logements }

$\Omega_B =$ { Ménages de 6 personnes et plus }

On constitue des grappes "naturelles" sur le terrain, de proche en proche.

$L_i =$ Taille de la grappe i (connue)

5) **Mesure d'un phénomène défini sur une période à partir d'un échantillonnage de jours.**

Ex : satisfaction d'une clientèle d'hôtel

→ Tirage d' 1 jour t dans le mois (probabilité $p(t)$), puis n clients parmi N_t.

→ La moyenne simple 'instantanée' va surestimer la satisfaction $\overline{Y}$

La pondération longitudinale dans le cas de SILC

SILC : échantillon rotatif constitué de 9 panels d'individus interrogés 9 années de suite;

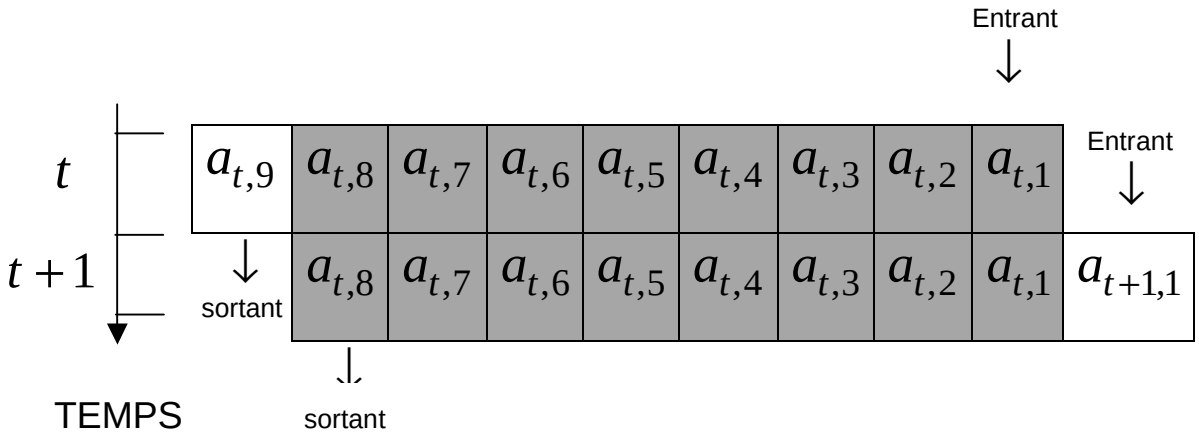
Les individus sont suivis au cours du temps;

Tirage annuel de logements dans l'échantillon-maitre Insee, tous les individus du ménage sont individus-panel (≈ 2000 logements par an);

Pour estimer l'évolution $t/t+1$ la population d'inférence est Ω_t;

Les "morts" sont simplement ignorés (pas de biais);

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



$a_{t,k}$ = sous-échantillon panel à enquêter l'année t en $k^{\text{ième}}$ interrogation ($a_{t+1,k+1} = a_{t,k} \quad (\forall t, \forall k \neq 9)$)

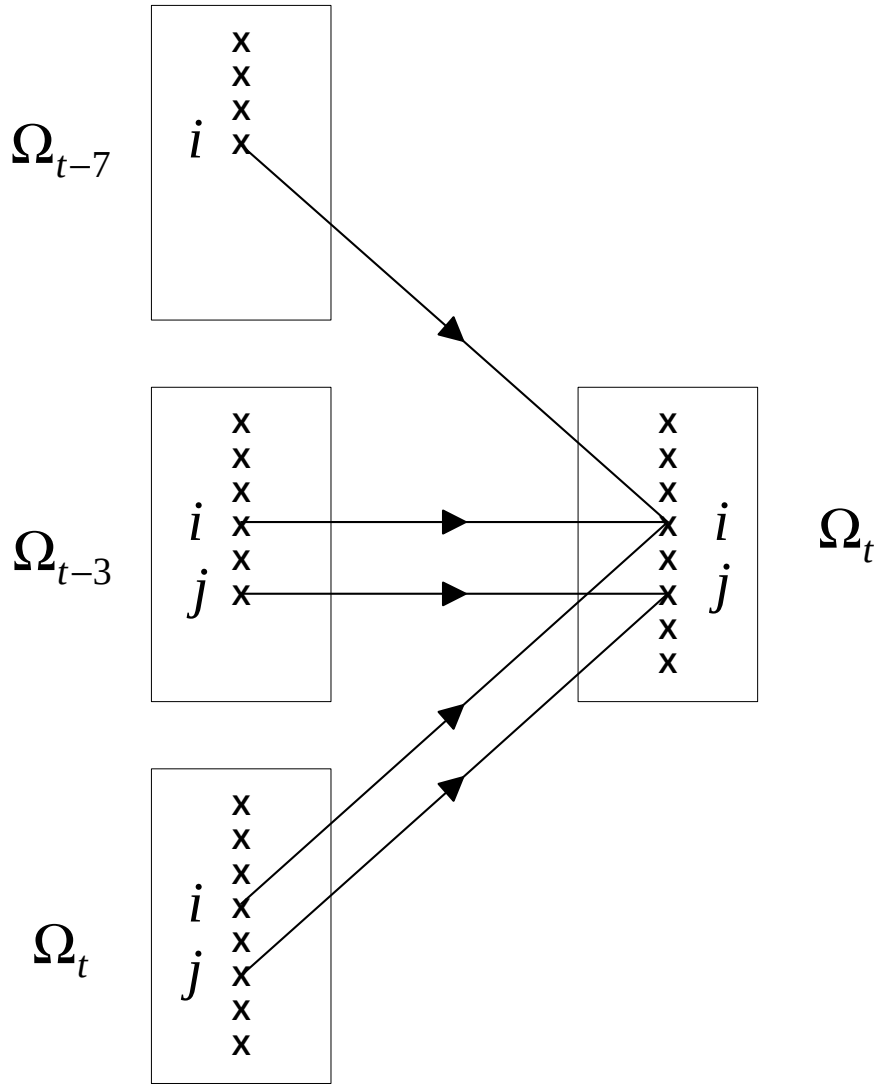
Partie grisée = éch. longitudinal = $s_{t,t+1} = \bigcup_{k=1}^{8} a_{t,k}$

Cas du panel "pur" : on conserve les poids d'origine - sans rien toucher.

Ici, il faut représenter Ω_t à partir de huit panels tirés à des dates différentes, **donc représentant des populations d'individus différentes**.

Intuitivement, un individu a *in fine* une probabilité de sélection à t fonction du nombre de panels dans lesquels il est susceptible d'être tiré.

Pondération à l'aide de la MGPP



1 individu de Ω_t = 1 grappe

$\forall i \in \Omega_t$, L_i = nombre d'années parmi $\{t-7, ..., t-1, t\}$ durant lesquelles i se trouvait dans le champ de l'enquête
($\Leftrightarrow$ tirable dans un panel « entrant »)

$$\rightarrow \text{on a } L_i \in \{1,2,3,\ldots,8\}.$$

On suppose que chaque année $t-k$, la base de sondage couvre exactement Ω_{t-k}.

Formalisation du poids longitudinal

Soit $K_i = \left\{ k \in 1,2,3,...8 \,\middle|\, i \in a_{t,k} \right\}$ = numéros des panels dans lesquels on retrouve i (sauf exception, K_i a un seul élément).

Si $i \in a_{t,k}$, $W_i(t,k)$ = poids du logement dans lequel se trouve i à la date de son tirage en tant qu'individu-panel, (tirage annuel dans Ω_{t-k+1}).

Nota : on a $\sum_{i \in a_{t,k}} W_i(t,k) \approx 60$ millions.

Le poids longitudinal de $i \in s_{t,t+1}$ est

$$W_i^{t,t+1} = \frac{1}{L_i} \sum_{k \in K_i} W_i(t,k)$$

⚠ Le questionnaire doit permettre d'obtenir L_i

(il est réaliste de supposer $\Omega_{t-7} \subset \Omega_{t-6} \subset \ldots \subset \Omega_{t-1} \subset \Omega_t$).

Simplifications éventuelles

Si on néglige les cas où un individu-panel peut être tiré deux fois ou plus (cas très rare), on a

$$W_i^{t,t+1} = \frac{W_i}{L_i}$$

où W_i est le poids de i relatif à l'unique sous-échantillon panel dans lequel il figure à la date t.

Si en plus (fictif) :

- la population n'évolue pas dans le temps,
- les panels sont tirés à proba égales ($W_i = W$),

alors

$$W_i^{t,t+1} = \frac{W}{8}$$

$\longrightarrow$ Résultat intuitif : tout se passe « comme si » n'importe quel individu de $s_{t,t+1}$ avait une probabilité de sélection égale à huit fois celle qui caractérise chaque sous-échantillon panel composant $s_{t,t+1}$.

Estimation finale

L'estimateur longitudinal de l'évolution est

$$\hat{\Delta}_{t,t+1} = \sum_{s_{t,t+1}} W_i^{t,t+1} . \left(Y_i^{t+1} - Y_i^{t} \right)$$

NB : a priori, les poids $W_i^{t,t+1}$ ne sont utilisés que dans le cadre d'une estimation d'évolution. Pour des estimations ponctuelles, ils apparaissent sans intérêt parce que la population d'inférence n'a pas grande signification à date donnée.

En pratique, la pondération sera soumise à des ajustements pour non-réponse et redressement.

La pondération transversale dans le cas de SILC

Rappel : inférence sur Ω_t, à la date courante *t*.

Considérons un sous-échantillon panel donné, tiré à t_0.

Difficulté principale : ce panel ne couvre correctement la population Ω_t que l'année de son tirage, soit t_0.

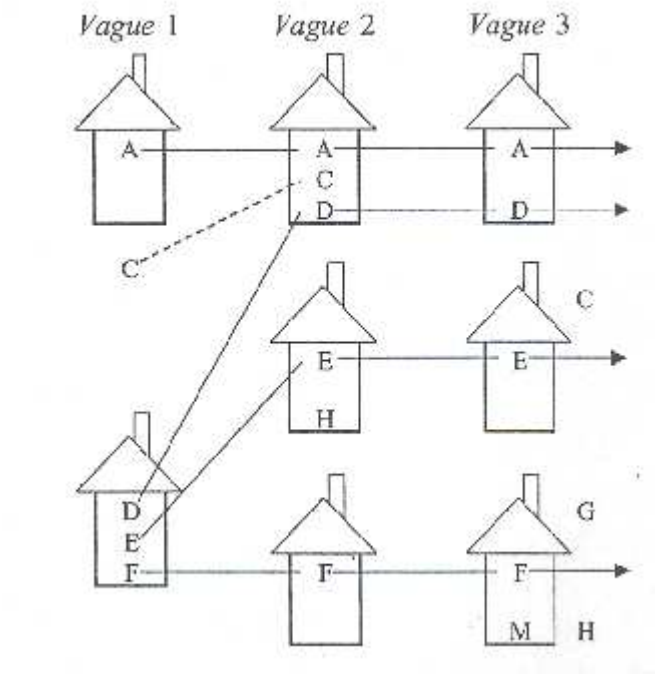
$\rightarrow$ comment représenter la population des « naissances » au-delà de t_0 ?

On ne peut pas échapper à un échantillon complémentaire au panel.

Pas de souci de biais avec les "morts" (problème dissymétrique).

Solution retenue :

On décide, pour chaque individu-panel enquêté lors du processus de suivi longitudinal, d'interroger l'ensemble des individus du ménage dans lequel se trouve l'individu-panel.



Ménage enquêté = { individu(s) panel + cohabitants }

$\Rightarrow$ sondage par grappes

⚠ Défaut : méthodologie ne permettant pas d'atteindre les ménages constitués seulement de « naissances » (ex : ménages 100% immigrants) au-delà de la date de tirage du panel.

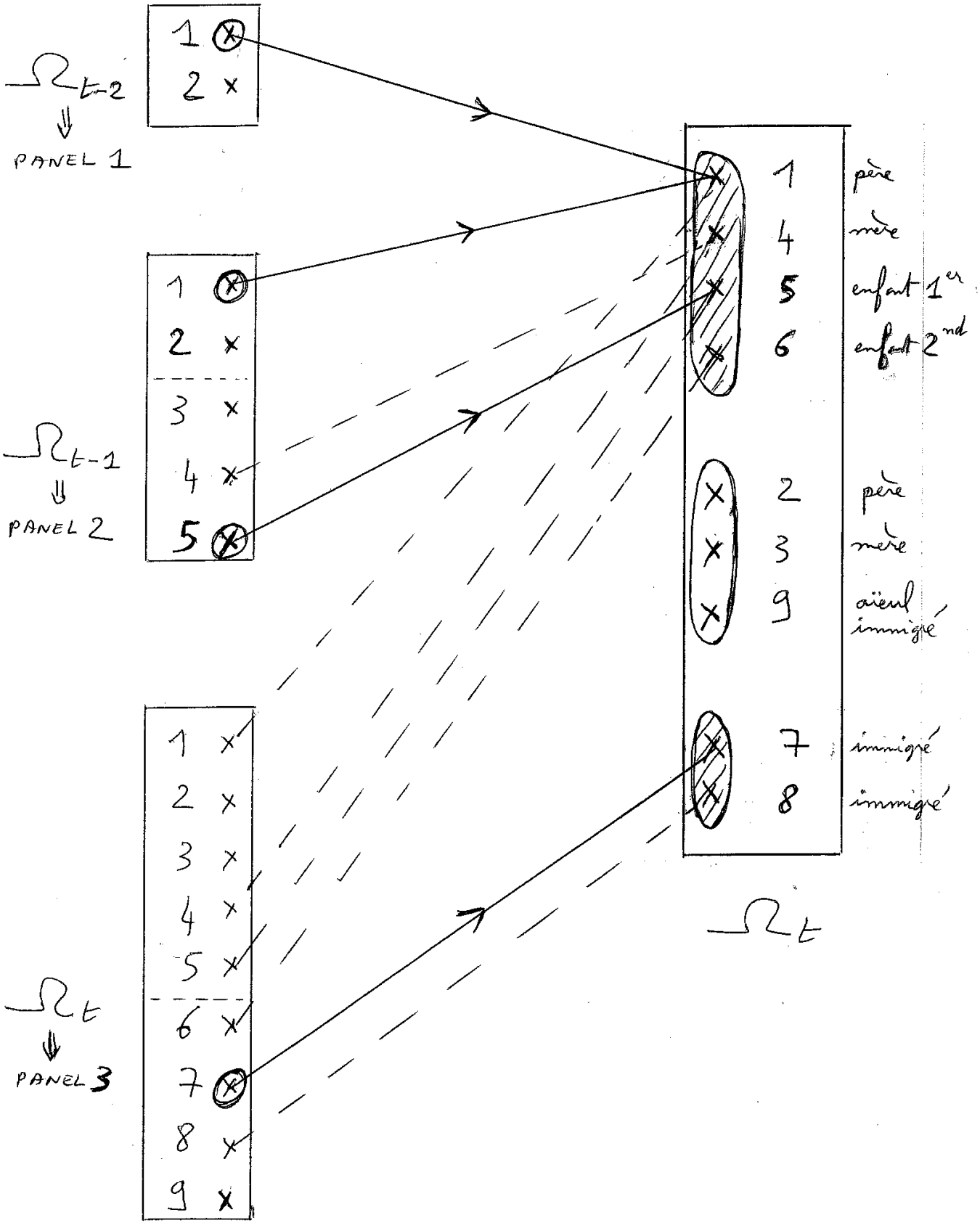
Solution à cet ultime problème : **l'échantillon rotatif** car certains panels entrant au-delà de t_0 (au moins celui tiré à t) permettent d'atteindre ces ménages.

L

Pondération à l'aide de la MGPP

L’approche la plus rigoureuse consiste à relier l’ensemble des neuf sous-échantillons panel $a_{t,k}$ (soit $\bigcup_{k=1}^{9} a_{t,k}$) à l’échantillon transversal $\tilde{u}_t$ de l’année t.

Définition du lien : un individu-panel quelconque de l’un des neuf sous-échantillons $a_{t,k}$ pointe sur lui-même en tant qu’individu de l’échantillon transversal à t .


Ω_{t-2}
PANEL 1
1
2
Ω_{t-1}
PANEL 2
1
2
3
4
5
Ω_t
PANEL 3
1
2
3
4
5
6
7
8
9
1 père
4 mère
5 enfant 1er
6 enfant 2nd
2 père
3 mère
9 aïeul immigré
7 immigré
8 immigré
Ω_t

W_i^t = poids transversal d'un individu i de $\tilde{u}_t$

m = ménage auquel appartient i.

En régime stationnaire,

$$W_i^t = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1}$$

où $W_j(t,k)$ = poids de j associé à l'échantillonnage $a_{t,k}$.

Le numérateur : somme de tous les poids « bruts » de tous les individus-panels du ménage,

Le dénominateur : pour chacune des neuf années $t-8$ à t considérées, nombre d'individus du ménage (individus-panels ou cohabitants) présents dans la base de sondage utilisée pour le tirage du panel entrant l'année en question.

Ce calcul nécessite évidemment la disponibilité de l'information via le questionnaire !!!

Propriétés, mises en garde

Tous les individus d'un même ménage ont le même poids.

La rotation assure que le nombre total de liens (dénominateur) est non nul pour chaque ménage i existant à $t \Rightarrow$ les poids transversaux sont "sans biais" …

…ou à peu près : la (probable) sous-estimation du nombre de liens est (en partie ?) corrigée par l'utilisation d'un redressement.

Un individu-panel n'apparaît *généralement* que dans un seul sous-échantillon (mais ce peut être davantage).

Si un individu n'a pas changé de logement, il ne peut pas apparaitre dans deux panels distincts.

Comme dans le cas longitudinal, le système informatique doit rattacher chaque individu-panel de $\tilde{u}_t$ à l'ensemble des échantillons panels $a_{t,k}$ dans lesquels il se trouve.

Penser au questionnaire : il faut la date de naissance des nouveau-nés, la date d'entrée des immigrants, etc.

Phase d'initialisation

La phase d'initialisation (2004 à 2011 compris) nécessite des adaptations.

2004 : pas de partage des poids (1er tirage)

2005 :

$$W_i^t = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 8 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}$$

2006 :

$$W_i^t = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2006}}} 1 \right) + \left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 7 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}$$

Estimation finale

Finalement, pour l'estimation de la différence

$$\Delta^*_{t,t+1} = Y_{t+1} - Y_t ,$$

on pourra calculer

$$\boxed{\hat{\Delta}^*_{t,t+1} = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t} Y_i^{t}}$$

Bibliographie :

P. Ardilly, *« Les techniques de sondage* », 2nde édition, Technip, 2006.

P. Ardilly et P. Lavallée, Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France, Techniques d'enquête, déc 2007, Vol. 33, N°2.

P. Lavallée, « *Le sondage indirect, ou la méthode généralisée du partage des poids* », Editions de l'Université de Bruxelles et Editions Ellipses, 2002.

* * * *