LE SONDAGE INDIRECT

Un survol

Pierre Lavallée

IASS

26 février 2021

CONTENU

1.INTRODUCTION

Sondage en grappes couramment utilisé en pratique.

Sélection de **grappes** (ensembles d'unités) et enquête auprès de **toutes** les unités des grappes choisies.

Enquêtes sociales :

Ménages : grappes d'individus

Enquêtes économiques :

Entreprises : grappes d'établissements, ou d'unités locales

Considérer toutes les unités appartenant à la même grappe :

1) Réduction des coûts de collecte
2) Production d'estimations sur les grappes

Situation classique :

Pour une enquête donnée,
on dispose d'une liste (base de sondage) contenant les unités de collecte désirées.

On tire un échantillon de la liste et on effectue l'enquête.

Problème étudié :

Pour une enquête donnée,
pas de liste contenant les unités de collecte désirées.

Mais plutôt,
on dispose d'une **autre liste** d'unités **reliée** cependant à la liste des unités de collecte.

Un peu de théorie :

Deux populations U^A et U^B **reliées entre elles.**
On désire produire une estimation pour U^B (population cible).
Base de sondage disponible pour U^A seulement.

Solution :

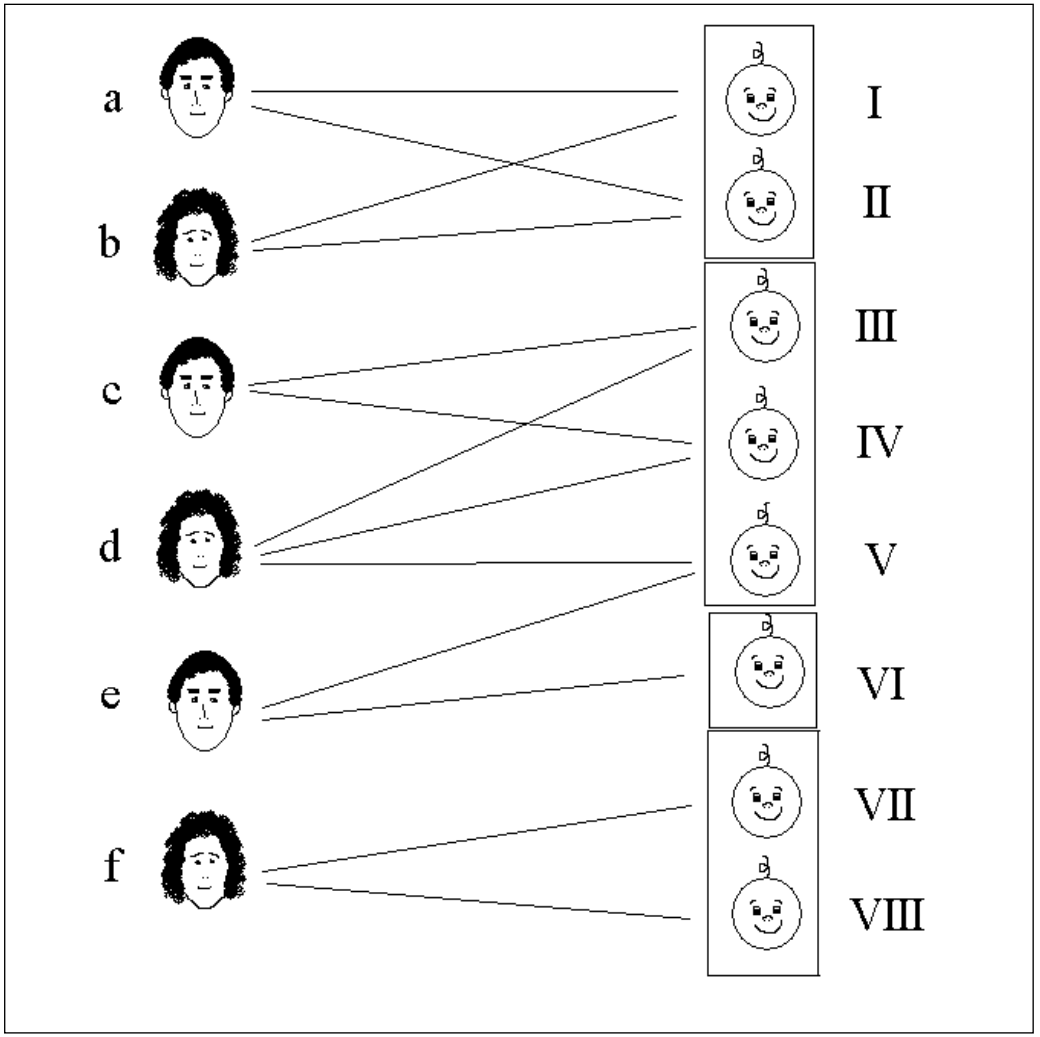
Tirage d'un échantillon de U^A
afin de produire une estimation pour U^B
en se servant de la correspondance existante entre les deux populations.

$\Rightarrow$ ***Sondage indirect***

Problème du statisticien : Obtenir la pondération pour une estimation sans biais.

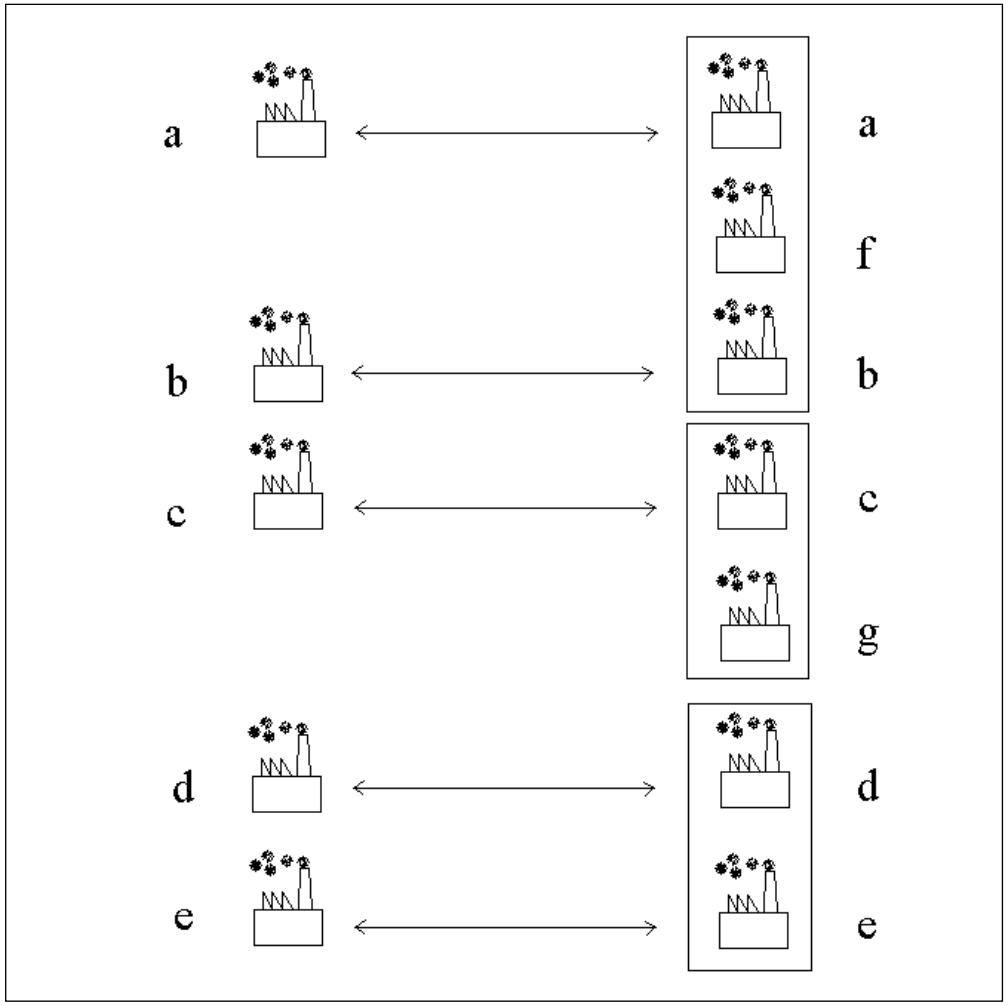
Exemple social :

Figure 1.1



Exemple économique :

Figure 1.2



Estimation du total Y^B en se servant de s^A tiré de U^A :

Défi de taille si les liens entre les unités de U^A et U^B ne sont pas bijectifs.

Difficulté d'associer une probabilité de sélection, ou un poids d'estimation, aux unités enquêtées dans U^B.

Solution :

Méthode généralisée du partage des poids (MGPP).

Permet d'obtenir un poids d'estimation pour chaque unité enquêtée de la population cible U^B.

2. LA MGPP
2.1 Description

Échantillon s^A contenant m^A unités tiré de U^A contenant M^A unités.

$\pi_j^A > 0$: Probabilité de sélection de l'unité j.

Population cible U^B contient M^B unités.

U^B divisée en N^B grappes, où la grappe i contient M_i^B unités.

Lien (ou correspondance) entre les unités *j* de U^A et les unités *k* des grappes *i* de U^B.

Lien est identifié par $l_{i,jk}$, où $l_{j,ik} = 1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité $ik \in U^B$, et 0 sinon.

Possible d'avoir : $L_j^A = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik} = 0$

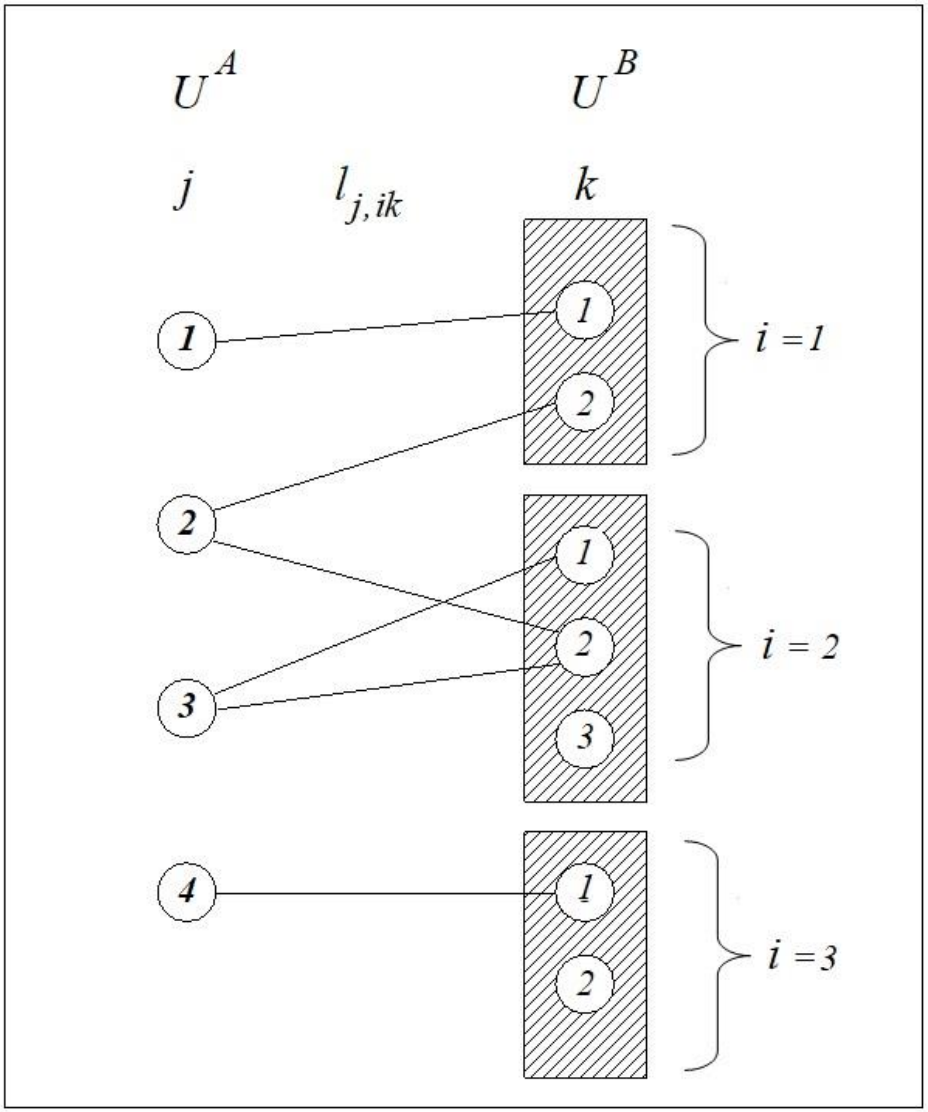
$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik} = 0$ $\quad$ $L_{ik}^B = 1$ $\quad$ $L_{ik}^B > 1$

Contrainte :

Chaque grappe i de U^B doit posséder au moins un lien avec une unité j de U^A, c'est-à-dire $L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} > 0$.

Nécessaire afin d'assurer l'aspect sans biais de la MGPP.

Exemple de liens :



Processus d'enquête du sondage indirect :

1. Pour chaque unité j de s^A, on identifie les unités ik de U^B qui ont $l_{j,ik} = 1$.

2. Pour chaque unité ik identifiée, on établit la liste des M_i^B unités de la grappe i contenant cette unité.

   s^B : Ensemble des n^B grappes identifiées par les unités $j \in s^A$.

3. On enquête auprès de <u>toutes</u> les unités k des grappes $i \in s^B$ pour mesurer la variable d'intérêt y.

On cherche à estimer le total Y^B :

$$Y^B = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik}$$

Solution classique (Horvitz-Thompson, 1952) :

$$\hat{Y}^{HT,B} = \sum_{i=1}^{n^B} \sum_{k=1}^{M_i^B} \frac{1}{\pi_{ik}^B} y_{ik}$$

Poids d'estimation : $1/\pi_{ik}^B$

Permet la production d'estimations sans biais.

Demande de connaître π_{ik}^B pour $i \in s^B$ …

MGPP :

Attribue un poids d'estimation w_{ik} à chaque unité k d'une grappe enquêtée i.

Estimateur de Y^B :

$$\hat{Y}^B = \sum_{i=1}^{n^B} \sum_{k=1}^{M_i^B} w_{ik} y_{ik}$$

Étapes de la MGPP :

Étape 1 : Pour chaque unité k des grappes i de s^B, on calcule le poids initial $w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$,

où $t_j = 1$ si $j \in s^A$, et 0 sinon.

Étape 2 : Pour chaque unité k des grappes i de s^B, on obtient le nombre total de liens $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$.

Étape 3 : On calcule le poids final $w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$.

Étape 4 : Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in U_i^B$.

Résultat :

$$
\begin{aligned}
w_{ik} &= \sum_{k=1}^{M_i^B} \frac{w'_{ik}}{L_i^B} \\
&= \sum_{k=1}^{M_i^B} \frac{1}{L_i^B} \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik} \\
&= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{1}{L_i^B} \sum_{k=1}^{M_i^B} l_{j,ik} \\
&= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{L_{j,i}}{L_i^B}
\end{aligned}
$$

2.2 Utilité de la MGPP

Solution simple à des problèmes de sondage et de pondération complexes.

En général, donne les mêmes résultats que la théorie classique pour les problèmes simples.

Solution intéressante,

même si la MGPP n'est pas toujours la plus précise (variance minimale) par rapport à une autre méthode d'estimation plus complexe.

1) Sondage indirect visant des populations rares

Populations rares difficiles à cerner aux fins d'enquête.

Pas de base de sondage adéquate.

Populations rares se retrouvent en grappes.

Réductions des coûts reliés à l'identification de ces populations.

2) Pondération utilisant seules les probabilités de sélection des unités sélectionnées

MGPP a besoin des probabilités de sélections π_j^A seulement pour les unités *j* sélectionnées dans l'échantillon s^A.

Simplification majeure par rapport à d'autres méthodes de pondération.

On peut considérer une pondération basée sur le calcul exact des probabilités de sélection de unités de s^B.

En pratique, peut être très difficile, voire même impossible, à effectuer.

3) Pondération de populations liées par des liens complexes

Les liens entre la population U^A et la population cible U^B sont souvent complexes (« plusieurs à plusieurs »).

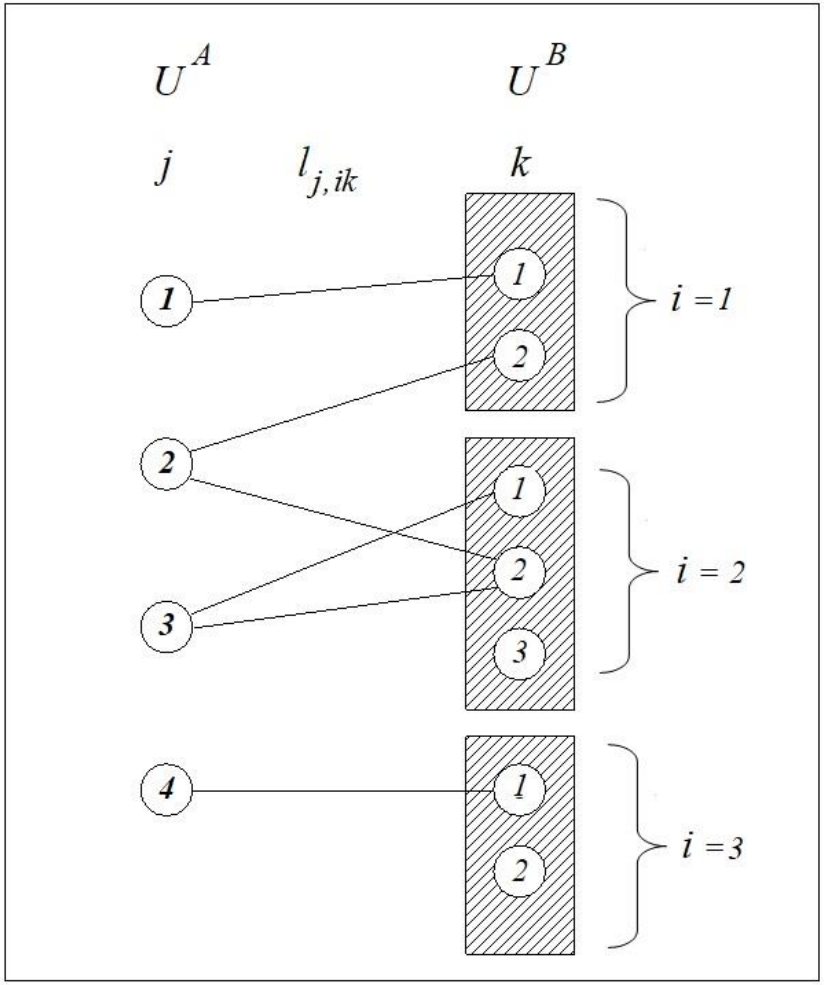
Pour obtenir un poids d'estimation pour chaque unité de s^B, la MGPP s'avère alors très utile.

4) Pondération d'unités non liées

Situation illustrée par l'unité 3 de la grappe 2 et l'unité 2 de la grappe 3 de U^B.

Ex : Nouveau-nés et immigrants dans les enquêtes longitudinales auprès d'individus appartenant à des ménages.

Problème pour obtenir un poids d'estimation pour ces unités.



3. PROPRIÉTÉS

3.1 Biais et variance

Théorème : Dualité de la forme de $\hat{Y}^B$ par rapport à U^A et U^B

Soit $z_{ik} = Y_i / L_i^B$ *où* $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$ *et* $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$ *pour tous les* $k \in U_i^B$. *L'estimateur* $\hat{Y}^B$ *peut alors également s'écrire sous la forme*

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j .$$

où

$$Z_j = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik}$$

$\hat{Y}^B$ peut donc s'écrire en fonction des unités *ik* de U^B, ou en fonction des unités *j* de U^A.

$\hat{Y}^B$: Estimateur d'Horvitz-Thompson avec la variable d'intérêt Z_j.

Corollaire 1 : Biais de $\hat{Y}^B$

L'estimateur $\hat{Y}^B$ est sans biais pour l'estimation de Y^B, par rapport au plan de sondage.

Corollaire 2 : Variance de $\hat{Y}^B$

La formule de la variance de l'estimateur $\hat{Y}^B$, par rapport au plan de sondage, est donnée par

$$Var(\hat{Y}^B) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'}$$

ou, de façon équivalente, par

$$Var(\hat{Y}^B) = -\frac{1}{2} \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} (\pi_{jj'}^A - \pi_j^A \pi_{j'}^A) \left(\frac{Z_j}{\pi_j^A} - \frac{Z_{j'}}{\pi_{j'}^A} \right)^2$$

où $\pi_{jj'}^A$: Probabilité conjointe de sélection de j et j'.

Estimateur de la variance $Var(\hat{Y}^B)$:

$$\hat{Var}(\hat{Y}^B) = \sum_{j=1}^{m^A} \sum_{j'=1}^{m^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} Z_j Z_{j'}$$

Estimateur de la variance $Var(\hat{Y}^B)$ de la forme Yates-Grundy :

$$\hat{Var}(\hat{Y}^B) = -\frac{1}{2} \sum_{j=1}^{m^A} \sum_{j'=1}^{m^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A} \left(\frac{Z_j}{\pi_j^A} - \frac{Z_{j'}}{\pi_{j'}^A} \right)^2$$

Autres estimateurs de la variance proposés dans la littérature :

Jackknife, Bootstrap...

Corollaire 3 : Forme alternative de l'estimateur $\hat{Y}^B$

L'estimateur $\hat{Y}^B$ *peut aussi s'écrire sous la forme*

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N^B} Y_i \frac{L_{j,i}}{L_i^B}$$

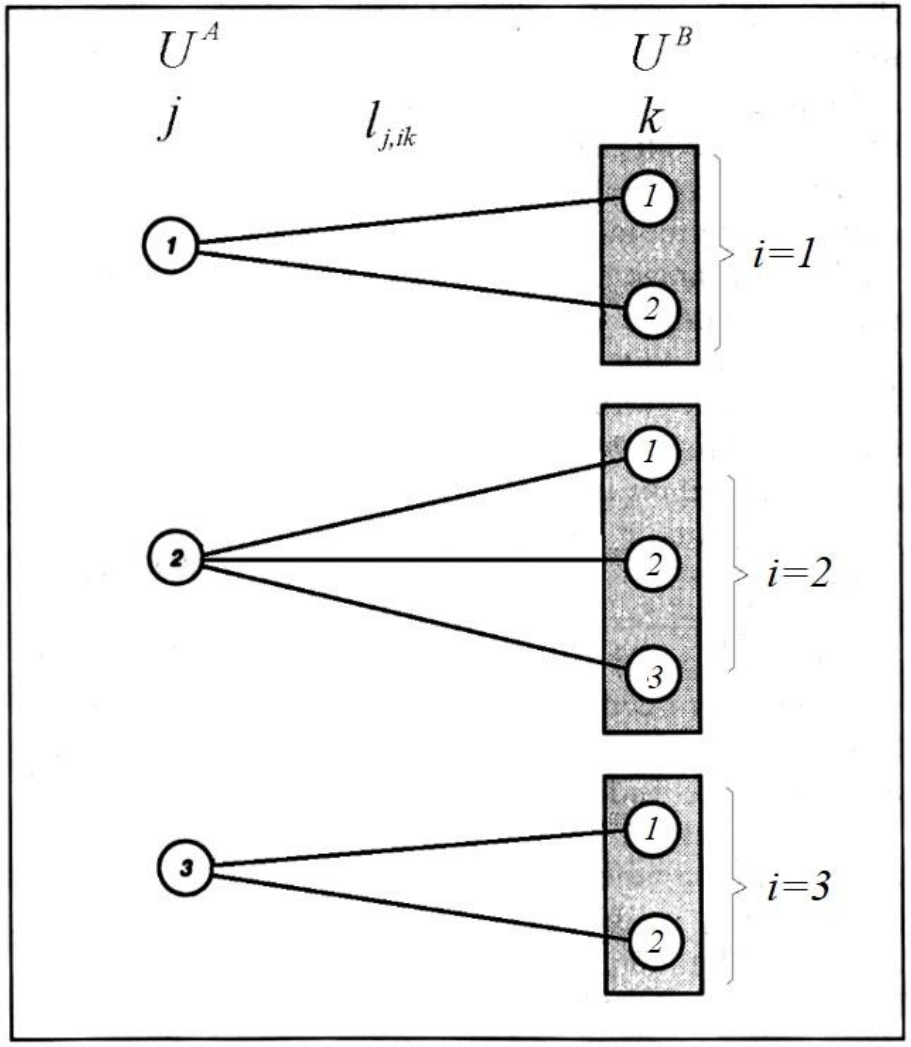
où $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$

3.2 Cas particulier 1 : Le sondage en grappes

Dans le cadre d'un sondage en grappes conventionnel.

La MGPP donne les mêmes résultats que la théorie classique.

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Y_j = \sum_{j=1}^{m^A} \frac{Y_j}{\pi_j^A} = \hat{Y}^{CLUS,B}$$



3.3 Cas particulier 2 : Recensement de la population U^A

En effectuant un recensement de U^A, on estime le total Y^B avec certitude.

3.4 Cas particulier 3 : Recensement de la population U^B

Si le nombre de grappes enquêtées n^B correspond au nombre total N^B de grappes de U^B, la variance de $\hat{Y}^B$ n'est pas nécessairement nulle.

4. AUTRES GÉNÉRALISATIONS

4.1 Utilisation de liens pondérés

Variable indicatrice $l_{j,ik}$:

- Indique s'il y a un lien ou non entre les unités *j* et *ik* des populations U^A et U^B.
- N'indique pas l'importance relative que pourraient avoir certains liens par rapport à d'autres.

Possible de remplacer $l_{j,ik}$ par une variable quantitative $\theta_{j,ik}$ représentant l'importance qu'on veut donner au lien $l_{j,ik}$.

$\theta_{j,ik}$ définie sur $[0,+\infty$.

$\theta_{j,ik}$ équivaut à $l_{j,ik}=0$.

Si le processus d'assignation des valeurs de $\theta_{j,ik}$ est indépendant du tirage de s^A, la MGPP reste sans biais.

Affecte cependant la précision de $\hat{Y}^B$.

4.2 Sondage indirect à deux degrés

1er degré :

- Sélection de s^A à partir de U^A.
- Pour chaque unité j sélectionnée dans s^A, on identifie les unités ik de U^B qui ont $l_{j,ik} = 1$.
- Pour chaque unité ik identifiée, on établit la liste des M_i^B unités de la grappe i contenant cette unité.
- s^B : Ensemble des n^B grappes $i$dentifiées par les unités $j \in s^A$.

2ème degré :

- De chaque $i \in s^B$, on sélectionne un sous-échantillon s_i^B contenant m_i^B unités parmi les M_i^B unités de la grappe.

4.3 Aspect arbitraire de la formation des grappes

En pratique, les grappes de U^B sont, la plupart du temps, formées de façon naturelle.

Enquêtes sociales : Ménages ou familles
Enquêtes économiques : Entreprises

Formation des grappes peut être effectuée de façon arbitraire.

Si le processus de formation des grappes est indépendant du tirage de s^A, la MGPP reste sans biais.

Affecte cependant la précision de $\hat{Y}^B$.

Raisons opérationnelles pour ne pas former des grappes de grande taille :

1. Difficulté d'établir l'inventaire des liens pour les grappes sélectionnées.

2. Instabilité des coûts de collecte.

    Causes : liens complexes, disparité entre les tailles des grappes

5. MGPP ET CALAGE SUR MARGES

Jusqu'à présent, la MGPP n'a pas utilisé d'information auxiliaire pour l'obtention des poids d'estimation.

Deux sources possibles d'information auxiliaire :

- Population U^A d'où est tiré l'échantillon.
- Population cible U^B.

On désire utiliser le calage sur marges avec la MGPP :

Correction des poids de la MGPP pour que les estimations produites correspondent à des totaux connus (information auxiliaire).

5.1 Rappel sur le calage sur marges

Développé par Deville et Särndal (1992).

Consiste à ajuster les poids de sondage de sorte que les estimations soient calées sur des totaux connus.

Total $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$ supposé connu.

Problème :

Obtenir des poids w_k^{CAL} **le plus près possible** des poids de sondage $d_k = 1/\pi_k$ de sorte que les totaux $\mathbf{X}$ soient respectés.

(Minimisation de la distorsion faite aux d_k)

5.2 La MGPP avec calage sur marges

Deux sources possibles d'information auxiliaire.

U^A :

- Vecteur colonne $\mathbf{x}_j^A$ de dimension p^A pour $j \in U^A$.
- Total $\mathbf{X}^A = \sum_{j=1}^{M^A} \mathbf{x}_j^A$ supposé connu.

U^B :

- Vecteur colonne $\mathbf{x}_{ik}^B$ de dimension p^B pour $ik \in U^B$.
- Total $\mathbf{X}^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \mathbf{x}_{ik}^B$ supposé connu.

Après une transformation simple, on peut utiliser directement la formulation de Deville et Särndal (1992) pour utiliser le calage sur marge avec la MGPP.

6. NON-RÉPONSE

6.1 Types de non-réponse totale

1. Non-réponse au sein des unités de s^A.

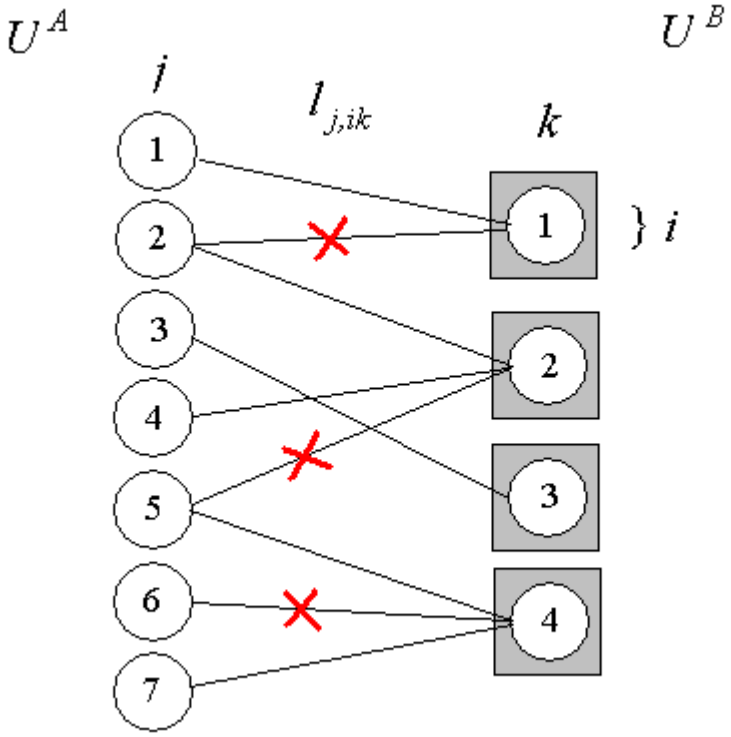
2. Non-réponse totale au sein des unités identifiées pour être enquêtées au sein de U^B.

   Deux cas :
   - Non-réponse de grappes
   - Non-réponse d'unités

3. Erreurs dans l'identification des liens (non-réponse de liens).

6.2 Non-réponse de liens

Impossible d'établir si l'unité *ik* de U^B est liée à l'unité *j* de U^A.



Problème sérieux pour la MGPP.

Produit des estimations biaisées.

Solutions possibles :

- Obtenir les liens $l_{j,ik}$ (ou $L_{j,i}$) par couplage d'enregistrements entre U^A et U^B.
- Estimer la probabilité $\phi_{j,ik}$ d'un lien entre les unités *j* et *ik* avec un modèle logistique.
- Estimer le nombre total de liens L_i^B avec un modèle log-linéaire.
- Corriger $\hat{Y}^{B*}$ avec un calage sur marges.
- Autres solutions… (Xu et Lavallée, 2009)

7. CONCLUSION

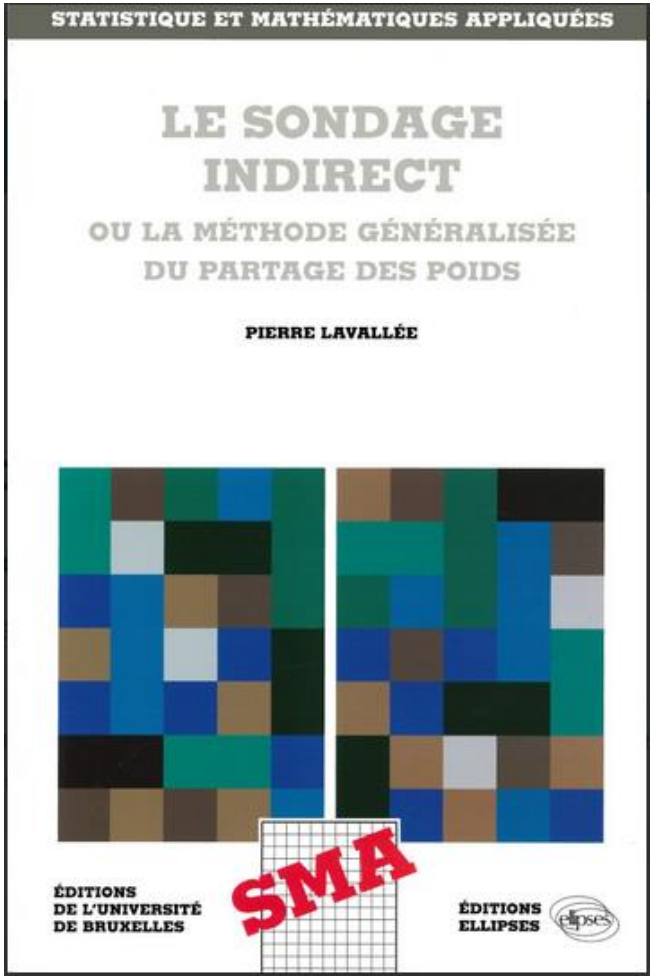
Sondage indirect :

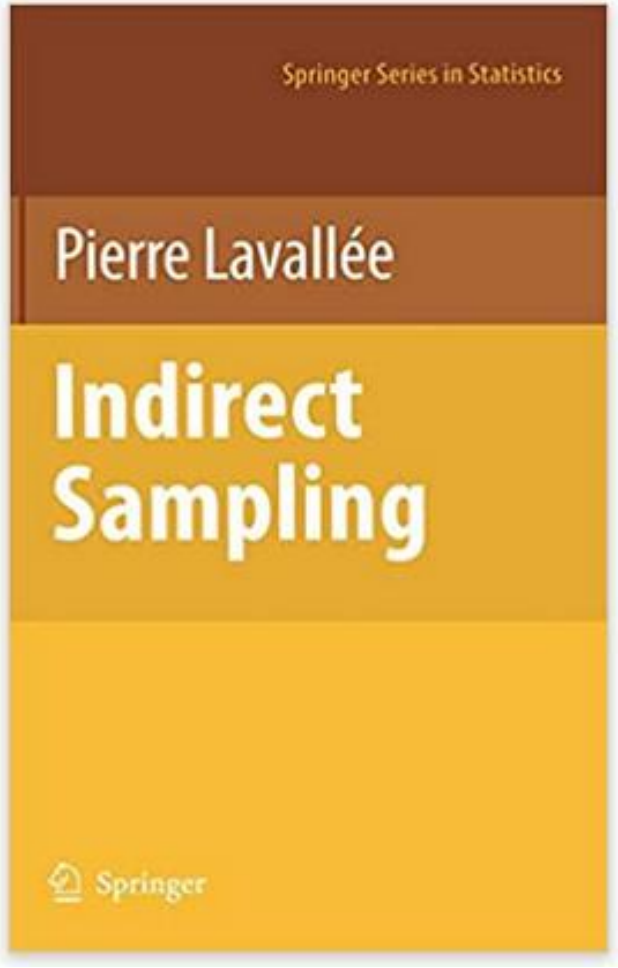
Situation courante où ne dispose pas de base de sondage, ou que celle-ci n'est pas à jour.

MGPP :

Solution viable pour des problèmes de sondage indirect de grappes.

Applications multiples… même au-delà de ce que l'on pensait au départ!





« Le bon sens est la chose du monde la mieux partagée. »
René Descartes