ERREUR D'OBSERVATION DES LIENS DANS LE SONDAGE INDIRECT : PISTES DE SOLUTION

Pierre Lavallée
et
Xiaojian Xu

Neuchâtel

Octobre 2005

CONTENU

1. INTRODUCTION

Situation classique en sondage:

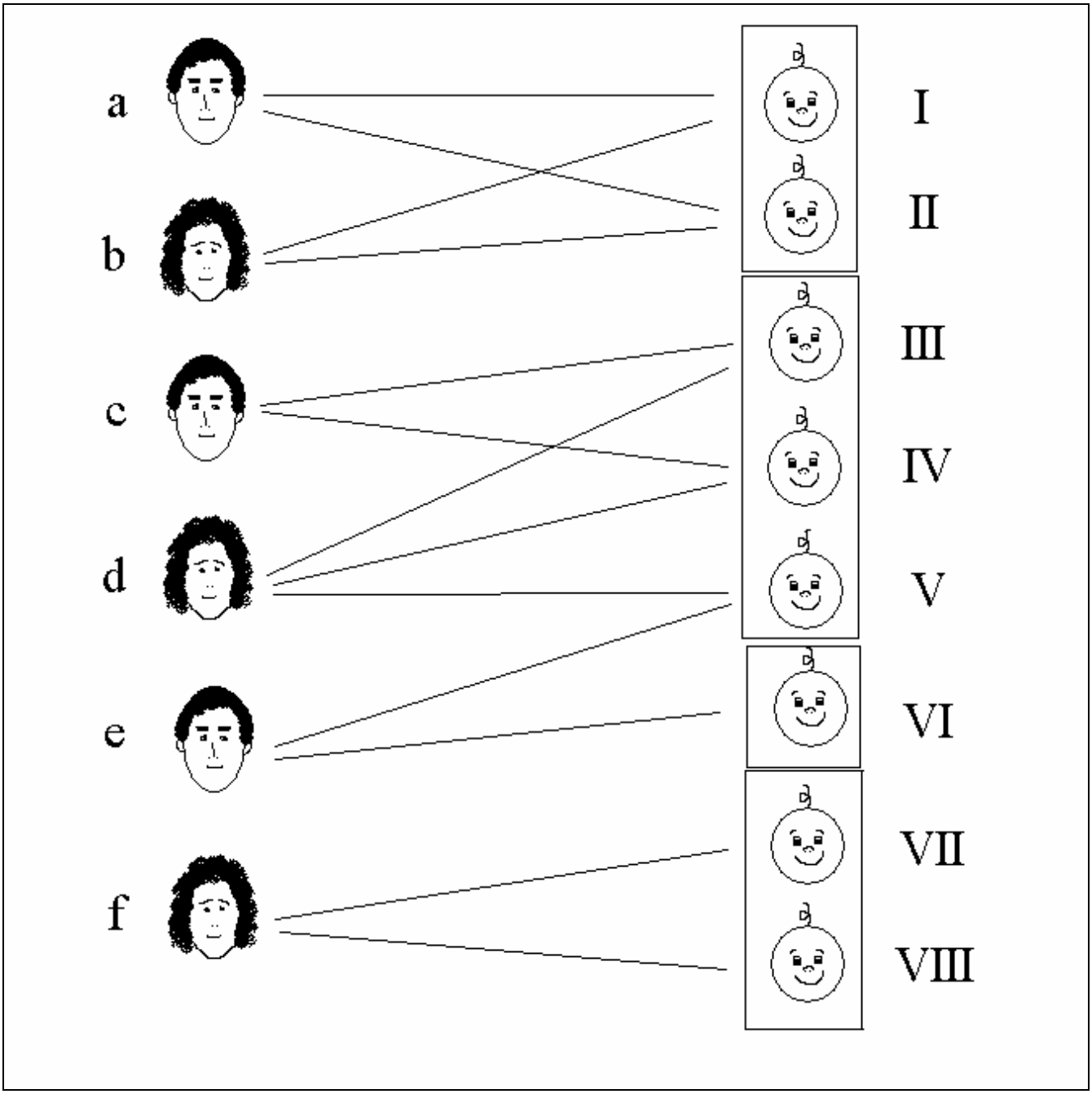
Pour une enquête donnée,
on dispose d'une liste (base de sondage) contenant les unités de collecte désirées.

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

Problème étudié ici:

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 désirée.


a
b
c
d
e
f
I
II
III
IV
V
VI
VII
VIII

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**

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

Difficile si les liens entre U^A et U^B ne sont pas bijectifs (un pour un).

Difficulté :

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

Solution :

MÉTHODE GÉNÉRALISÉE DU PARTAGE DES POIDS

Obtention d'un poids d'estimation pour chaque unité enquêtée de la population cible U^B (Lavallée, 1995, Lavallée, 2002).

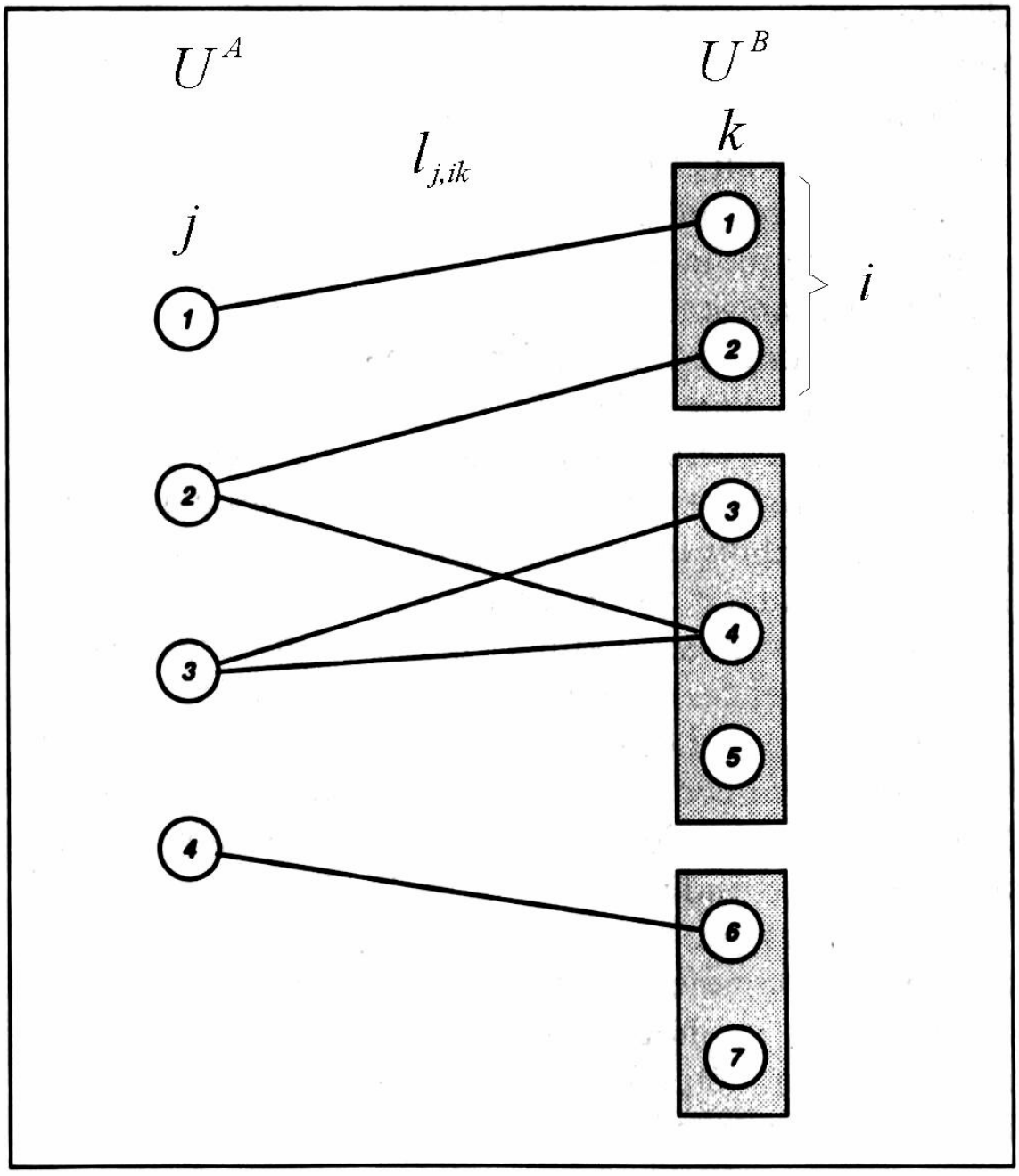
2. MGPP
2.1 Description

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

$l_{j,ik} = 1$ si lien entre l'unité $j \in U^A$ et l'unité $ik \in U^B$, 0 sinon.


U^A
U^B
$l_{j,ik}$
k
j
i
1
2
3
4
5
6
7

Processus d'enquête:

Pour chaque unité *j* de s^A,
on identifie les unités *ik* de U^B qui ont un lien, c.-à-d. $l_{j,ik} = 1$.

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

On enquête auprès de toutes les unités des grappes identifiées.

On mesure notamment le nombre de liens L_i^B qui pointent sur les unités de U^A de chaque grappe *i* identifiée, c.-à-d.

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} .$$

Considérer toutes les unités d'une grappe :

- Réduction des coûts de collecte
- Production d'estimations sur les grappes

On cherche à estimer le total Y^B :

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

Solution classique (Horvitz-Thompson, 1952):

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

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

Demande de connaître π_{ik}^B.

MGPP:

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

Étape 1 : Poids initial w'_{ik} :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$$

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

Étape 2 : Nombre total de liens L_i^B :

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik}$$

Étape 3 : Poids final w_i :

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{L_i^B}$$

Étape 4 : $w_{ik} = w_i$

2.2 Estimation

Estimateur de Y^B :

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

Autres formes de $\hat{Y}^B$:

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

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

$\hat{Y}^B$:Estimateur Horvitz-Thompson avec variable dérivée Z_j.

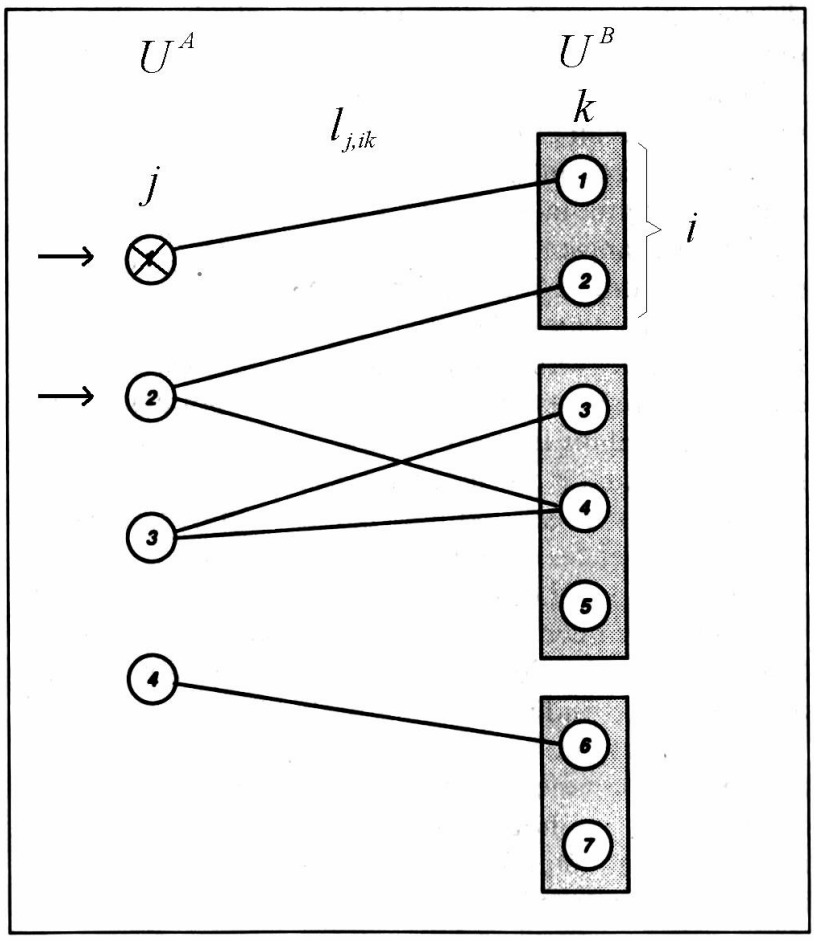
$\hat{Y}^B$ est sans biais pour Y^B.

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

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

3. NON-RÉPONSE TOTALE
3.1 Non-réponse au sein de s^A

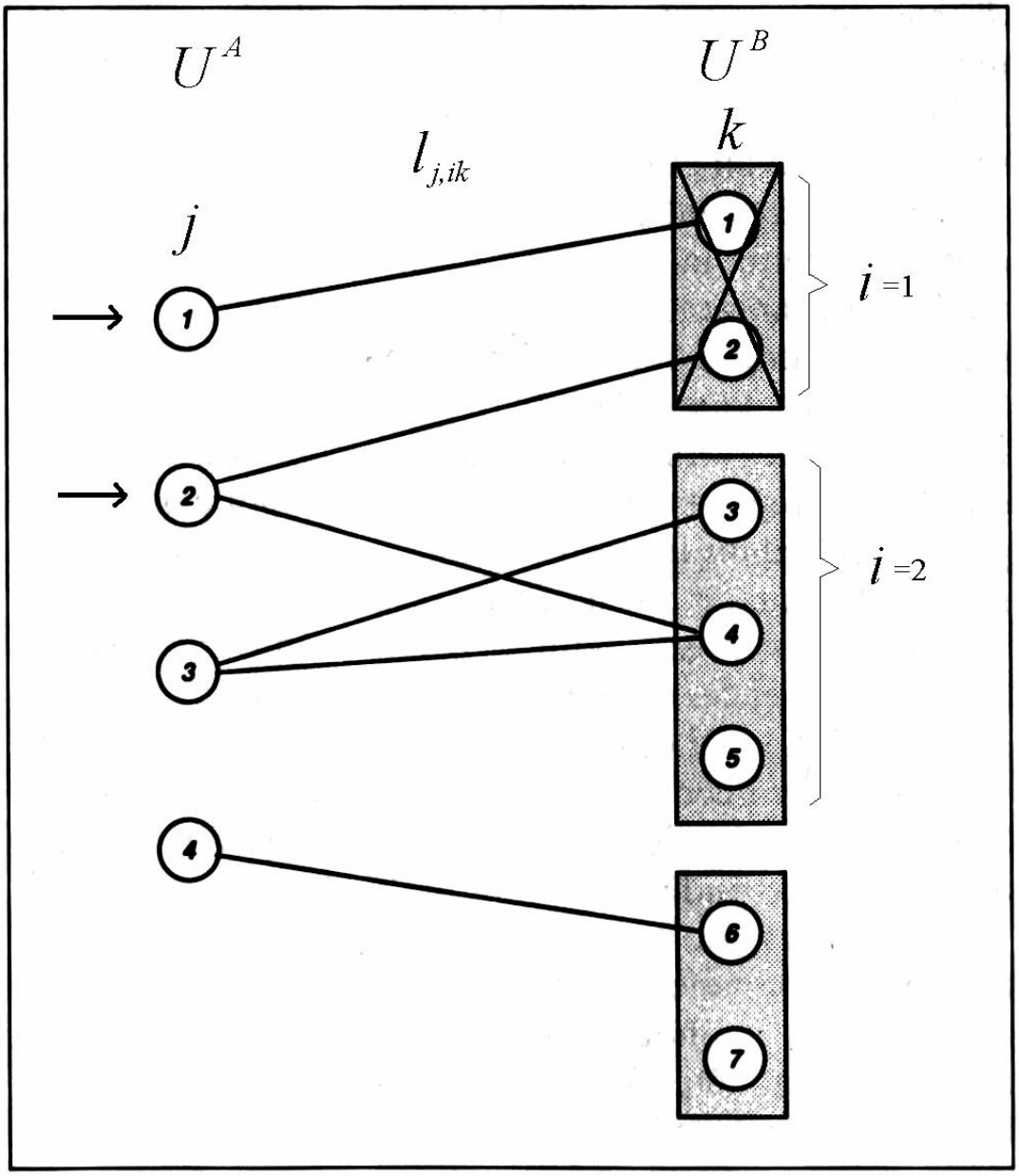
Cas classique de non-réponse.



Se traite comme le cas où on a tiré l'échantillon s^A dans le but de produire une estimation pour U^A.

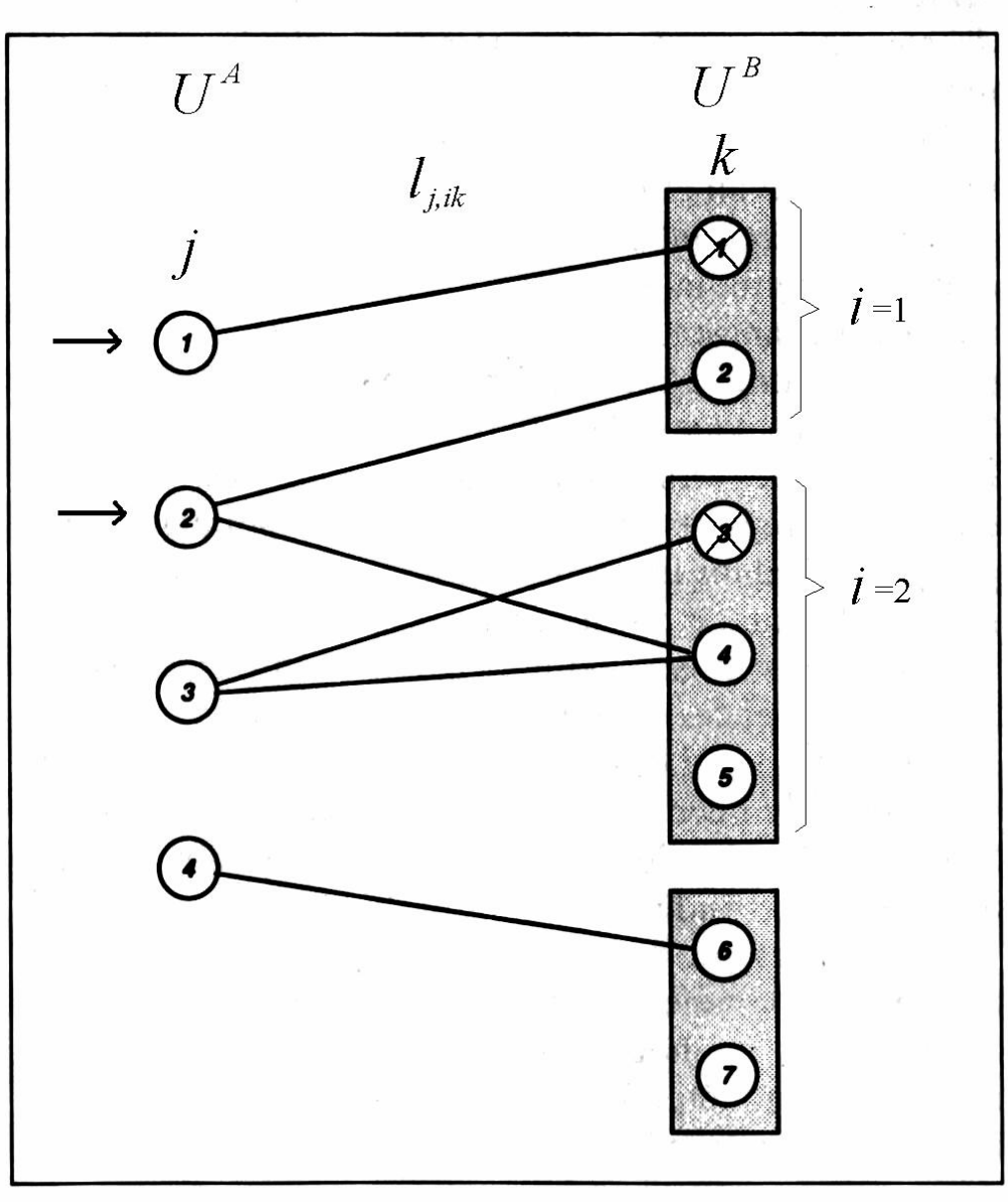
3.2 Non-réponse de grappes

Pour certaines grappes entières, on ne peut obtenir de données.



3.3 Non-réponse d'unités

Non-réponse totale où une ou plusieurs unités de la grappe, mais pas toutes, n'ont pas répondu.

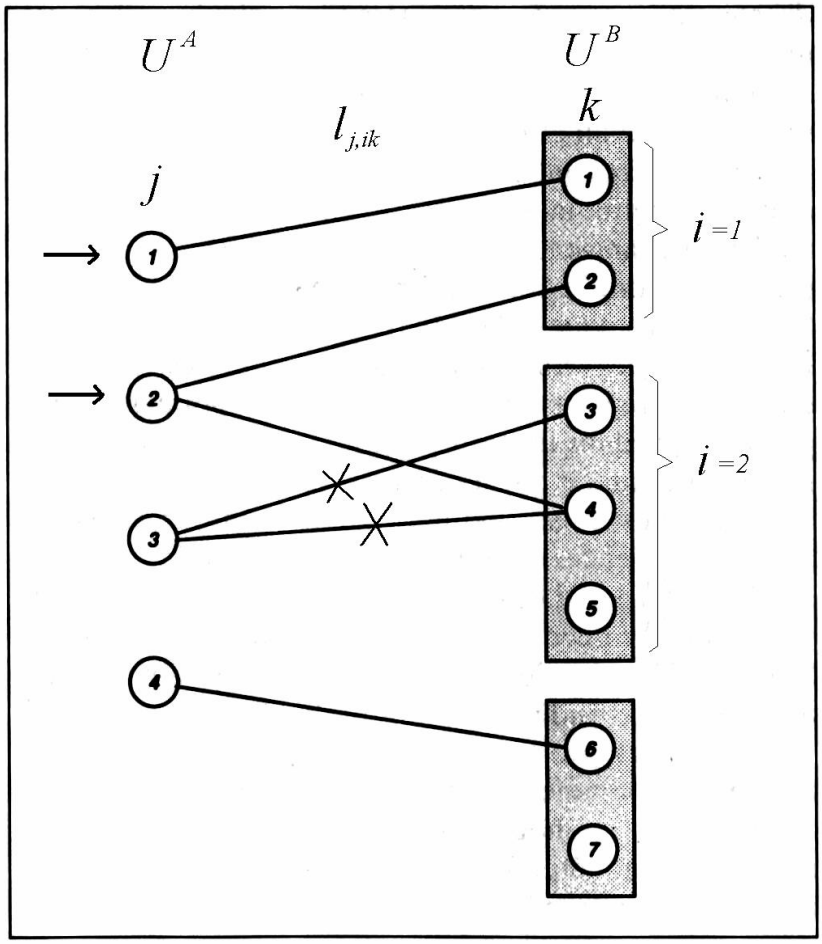


4. ERREURS D'OBSERVATION DES LIENS

Impossibilité d'établir si une unité j de U^A est liée à une unité ik de U^B.

Particulièrement problématique pour la MGPP.

Peut créer de sérieux problèmes de biais dans les estimations.



Problèmes types :

1. Impossibilité d'établir en général les liens $l_{j,ik}$ entre les unités i de U^A et les unités ik de U^B.

2. Impossibilité d'établir le nombre total de liens L_i^B de la grappe i de U^B.

3. Impossibilité d'établir le nombre total de liens L_i^B de la grappe i de U^B, mais contraintes connues sur L_i^B.
   (ex : $L_i^B \leq M_i^B$)

4.1 Obtention des liens entre U^A et U^B

Supposons qu'on ne connaît pas les liens individuels entre les unités des deux populations U^A et U^B, c.-à-d. $l_{j,ik}$=?.

Pour la MGPP, seuls les liens des unités j de s^A et les unités des grappes identifiées de U^B sont nécessaires.

En pratique, on peut souvent obtenir ces liens lors des entrevues de unités sélectionnées dans s^A, ainsi que des unités k des grappes i identifiées dans U^B.

Si on dispose cependant de deux fichiers A et B contenant respectivement les populations U^A et U^B, on peut chercher à obtenir tous les liens entre les deux populations.

Une façon d'obtenir les valeurs de $l_{j,ik}$ est d'effectuer un *couplage d'enregistrements*.

Couplage d'enregistrements :

Deux fichiers A et B.

j : un enregistrement (ou unité) du fichier A (ou population U^A).

k : un enregistrement (ou unité) du fichier B (ou population U^B).

Pour chaque paire *(j,k)* de l'espace ***AXB,*** on calcule un poids de couplage θ_{jk}.

θ_{jk} reflète la propension à considérer la paire *(j,k)* comme un véritable lien.

Important : θ_{jk} est une fonction de différentes probabilités. On peut donc chercher à estimer θ_{jk} avec un modèle de type logistique en utilisant des vecteurs de variables auxiliaires $\mathbf{x}_j^A$ et $\mathbf{x}_k^B$.

Règle de décision de Fellegi et Sunter (1969):

$$D(j,k) = \begin{cases} \text{lien} & \text{si } \theta_{jk} \geq \theta_{High} \\ \text{lien possible} & \text{si } \theta_{Low} < \theta_{jk} < \theta_{High} \\ \text{lien nul} & \text{si } \theta_{jk} \leq \theta_{Low} \end{cases}$$

Application de la règle de décision:

Besoin d'intervention manuelle lorsque $\theta_{Low} < \theta < \theta_{High}$.

Erreurs possibles.

Variable indicatrice:

$l_{jk} = 1$ si la paire *(j,k)* est considérée comme un lien,
0 sinon.

N.B.: La règle de décision n'empêche pas d'avoir des liens surjectifs ou injectifs.

4.2. Estimation de L_i^B

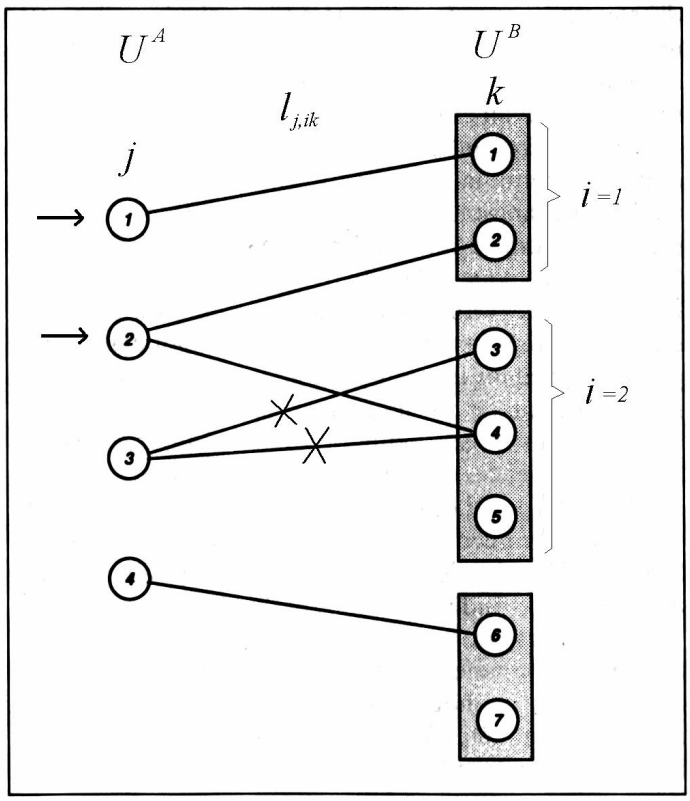
On suppose qu'on connaît les liens $l_{j,ik}$ pour toutes les unités j de s^A.

Cependant, pour certaines unités des grappes i de U^B identifiées par s^A, on ne connaît pas tous les liens $l_{j,ik}$ nous ramenant à U^A.

Autrement dit :

On connaît $l_{j,ik}$ pour $j \in s^A$.

On ne connaît pas tous les $l_{j,ik}$ pour $j \notin s^A$.



Conséquence : On ne peut établir la valeur de L_i^B essentielle à la MGPP.

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

où

$$L_i^B = \sum_{j=1}^{M^A} L_{j,i} = \sum_{j \in s^A} L_{j,i} + \sum_{j \notin s^A} L_{j,i}$$
$$= L_{j,i}^{(s)} + L_{j,i}^{(\bar{s})}$$

L_i^{B*} : Nombre total de liens établis entre la grappe i et la population U^A.

- $L_i^{B*} \leq L_i^B$
- $L_i^{B*} \geq L_i^{B(s)} > 0$ où $L_i^{B(s)} = \sum_{j \in s^A} L_{j,i}^{(s)}$

  Il y a au moins un lien connu d'une unité j de s^A vers la grappe i identifiée.

Produit une surestimation de Y^B.

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

Approche 1 : Modéliser $l_{j,ik}$ à partir de variables auxiliaires

Peut s'effectuer par couplage d'enregistrements entre les unités *k* des grappes *i* identifiées et la population U^A.

On peut aussi chercher à estimer la probabilité $\phi_{j,ik}$ d'un lien les unités *j* et *ik* avec un modèle de type logistique en utilisant des vecteurs de variables auxiliaires $\mathbf{x}_j^A$ et $\mathbf{x}_{ik}^B$.

$$\hat{l}_{j,ik} = \hat{\phi}_{j,ik}$$

Peut se voir comme l'imputation des liens $l_{j,ik}$ pour $j \notin s^A$ (Ardilly et le Blanc, 2000).

Avec $\hat{l}_{j,ik}$, on obtient

$$\hat{L}_{ik}^{B} = \sum_{j \in s^{A}} l_{j,ik} + \sum_{j \notin s^{A}} \hat{l}_{j,ik}$$

$\hat{L}_{ik}^{B}$: Nombre estimé de liens entre l'unité *ik* de U^{B} et la population U^{A}.

Puis,

$$\hat{L}_{i}^{B} = \sum_{k=1}^{M_{i}^{B}} \hat{L}_{ik}^{B}$$

Approche 2 : Estimer $L_{j,i}$ pour $j \notin s^A$

Peut s'effectuer par couplage d'enregistrements entre les grappes *i* identifiées et la population U^A.

On peut aussi utiliser la proportion de liens $\overline{L}_{\bullet,i}$ entre les unités de U^A et la grappe *i* de U^B.

$$\overline{L}_{\bullet,i} = \frac{\sum_{j=1}^{M^A} L_{j,i}}{M^A}$$

Ex : Proportion de la population entrant dans un ménage donné *i*.

Estimation de $\overline{L}_{\bullet,i}$ à partir de s^A :

$$\hat{\overline{L}}_{\bullet,i} = \frac{\sum_{j=1}^{m^A} L_{j,i}}{m^A}$$

Estimateur par prédiction de L_i^B en utilisant $\hat{\bar{L}}_{\bullet,i}$:

$$\hat{L}_i^{B,\text{Préd}} = \sum_{j \in s^A} L_{j,i} + \sum_{j \notin s^A} \hat{\bar{L}}_{\bullet,i}$$

$$= \sum_{j=1}^{m^A} L_{j,i} + (M^A - m^A) \frac{\sum_{j=1}^{m^A} L_{j,i}}{m^A}$$

$$= \frac{M^A}{m^A} \sum_{j=1}^{m^A} L_{j,i}$$

Approche 3 : Estimer directement L_i^B

On cherche à estimer L_i^B comme un tout, sans référence à la population U^A.

Modèle log-linéaire :

$$\log(L_i^B) = \boldsymbol{\beta}' \mathbf{x}_i^B$$

Ex : Nombre typique de personnes dans un ménage avec certaines caractéristiques.

On peut aussi utiliser la proportion estimée de liens $\hat{\bar{L}}_{\bullet,i}$ entre les unités de U^A et la grappe i de U^B :

$$\hat{\bar{L}}_{\bullet,i} = \frac{\sum_{j=1}^{m^A} L_{j,i}}{m^A}$$

Estimateur direct de L_i^B:

$$\hat{L}_i^{B,\text{Dir}} = M^A \times \hat{\bar{L}}_{\bullet,i}$$

$$= \frac{M^A}{m^A} \sum_{j=1}^{m^A} L_{j,i}$$

En généralisant,

$$\hat{L}_i^{B,\text{Gén}} = \sum_{j \in s^A} \frac{L_{j,i}}{\pi_j^A}$$

Approche 4 : Corriger la surestimation de $\hat{Y}^{B*}$ par calage

Dépend de la disponibilité de variables auxiliaires $\mathbf{x}_{ik}^{B}$ corrélées avec la variable d'intérêt y_{ik}, ainsi que de totaux de contrôle $\mathbf{X}^{B}$.

Le calage sur marges peut s'effectuer avant l'application de la MGPP, ou sinon après (Lavallée, 2002).

4.3. Estimation de L_i^B sous contraintes

Rappel :

- On connaît $l_{j,ik}$ pour $j \in s^A$.
- On ne connaît pas tous les $l_{j,ik}$ pour $j \notin s^A$.

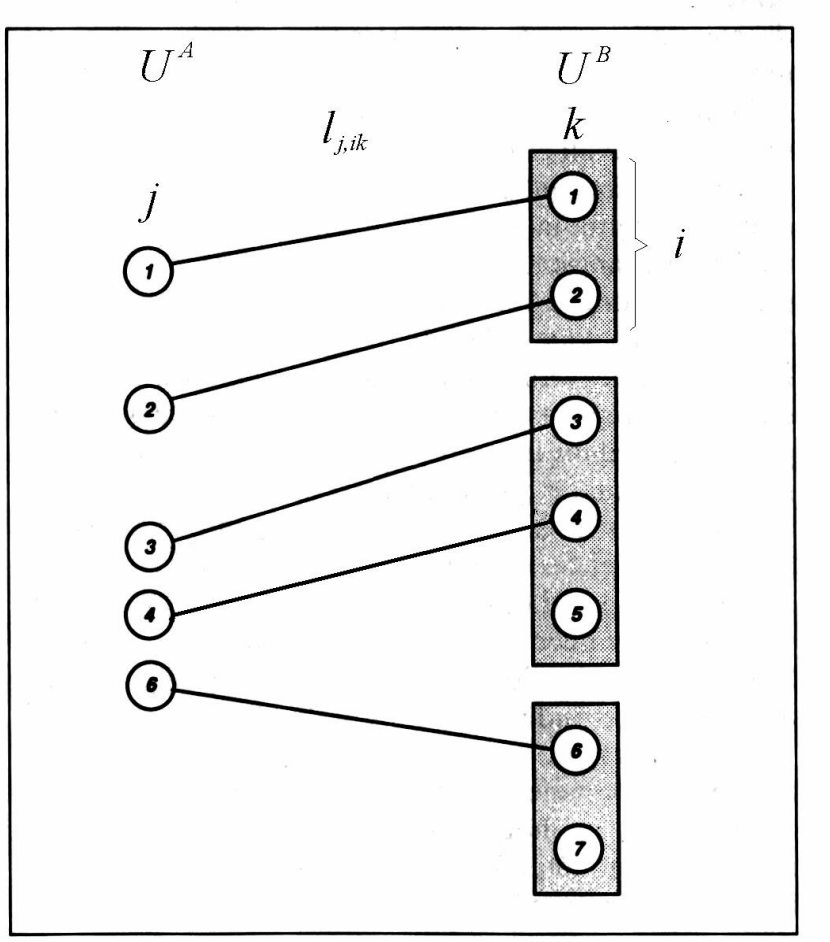
En général, on connaît la taille M_i^B des grappes i de U^B identifiées par s^A.

On peut souvent établir une relation entre M_i^B et L_i^B.

Exemple : $L_i^B = \gamma \times M_i^B$

Cas particulier: Enquêtes longitudinales

On note que $L_i^B \leq M_i^B$.



Estimation de L_i^B par $\hat{L}_i^{B,\text{Mén}} = M_i^B$:

Corresponds à ne supposer aucune naissance dans la population.

Produit une sous-estimation de Y^B.

$$\hat{Y}^{B,\text{Mén}} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{M_i^B} \quad \leq \quad \hat{Y}^B$$

Biais (Ardilly, 2004) :

Vu que

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

on a

$$E(\hat{Y}^{B,\text{Mén}}) - Y = \sum_{j=1}^{M^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{M_i^B} - \sum_{j=1}^{M^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{L_i^B}$$

$$= \sum_{j=1}^{M^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{L_i^B} \left[\frac{L_i^B - M_i^B}{M_i^B} \right]$$

$$= \sum_{i=1}^{N} \alpha_i^B Y_i$$

où $\alpha_i^B = \dfrac{L_i^B - M_i^B}{M_i^B}$

α_i^B représente l'ampleur du biais causé par la surestimation de L_i^B.

Estimation de L_i^B par $\hat{L}_i^{B,\text{Long}} = m_i^A$:

m_i^A : Nombre d'individus de s^A contribuant à la grappe i.

Correspond à supposer que les seuls individus de la grappe i sont les m_i^A unités de s^A qui ont mené à l'identification de cette grappe.

Produit une surestimation de Y^B si plus de m_i^A unités de U^A sont liées à la grappe i.

$$\hat{Y}^{B,\text{Long}} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{m_i^A} \quad \geq \quad \hat{Y}^B$$

Si on sélectionne des logements de U^A (au lieu de personnes) et on définit l'unité d'enquête de U^B comme le ménage au lieu de l'individu ($M_i^B = 1$ et $i=k$), la surestimation peut devenir négligeable (Ardilly, 2003).

5. CONCLUSION

MGPP :

Solution viable pour des problèmes de sondage indirect.

Non-réponse :

Phénomène inévitable.

Cas important à traiter :

Erreur d'observation des liens

Possible de combiner les différents traitements.