COMPOSITION , FACTORISATION ET CONDITIONS D'OPTIMALITE (FAIBLE, FORTE) DANS LA METHODE DE PARTAGE DES POIDS ; APPLICATION A L'ENQUETE SUR LE TOURISME EN BRETAGNE.

Jean-Claude DEVILLE(*), Pierre LAVALLEE (**) et Myriam MAUMY(***)
(*) CREST/ENSAI
(**)Statistique Canada
(***)Université de Strasbourg

1-Introduction rapide

La méthode de partage des poids (ou échantillonnage indirect) peut s'appliquer quand deux populations *A* et *B* sont liées par un graphe dont les flèches relient des éléments de *A* à des éléments de *B*. On possède, pour *A*, une 'bonne' base de sondage permettant de faire un 'bon' sondage probabiliste conduisant à un système de poids w_i . En revanche, on ne dispose pas de base pour *B*, et, seule, les flèches du graphe nous livrent un moyen d'attraper des unités *k* de *B*. Malheureusement, les probabilités d'inclusion sont inconnues, voir incalculables. La figure 1 donne un exemple simpliste de cette situation. Si les flèches du graphe sont chargées par des nombres positifs θ_{ik} , tout système de poids sans biais w_i sur A se transforme en un système de poids w_k sur *B*. Matriciellement, si on représente les poids par des vecteurs-lignes w^A et w^B et que la matrice *T* est celle des θ_{ik} , on pose $w^B=w^A T$. Il est bien connu [3,6], que les poids sur B sont sans biais si et seulement si $\mathbf{1}^A T=\mathbf{1}_B$, où $\mathbf{1}^{A \text{ resp } B}$ est le vecteur ligne de 1 construit sur *A resp B.* Autrement dit la somme en *i* des θ_{ik} à *k* fixé est égale à 1 pour tout *k* .

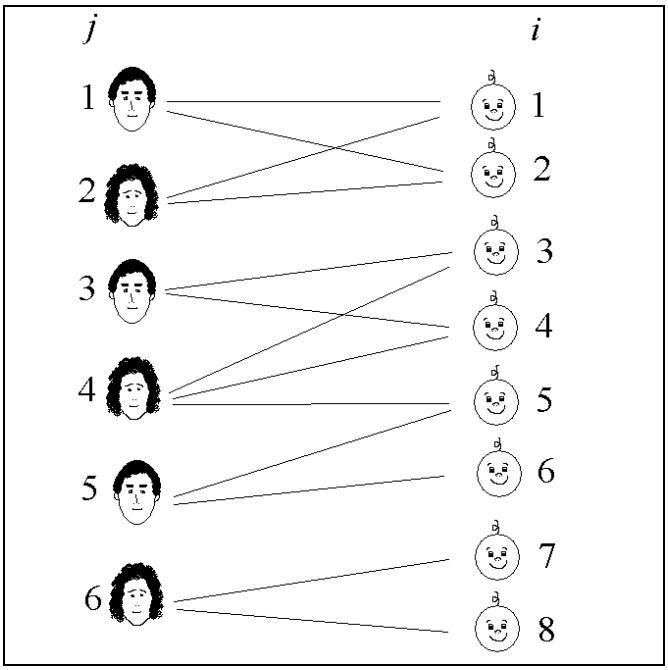
Si on connaît la matrice D^A des variances/covariances sur *A* (c'est à dire $Var(w^A)$) alors la variance du système de poids sur *B* vaut $D^B=T'D^AT$.

Certains graphes ont des structures particulières correspondant à des situations bien connues. Pour notre propos, nous aurons besoin des graphes du type 'Grappes en un pour tous' (dans la terminologie de [2] traduite librement) où *B* est partagée en grappes. Tous les éléments de *B* sont pointés par une seule flèche, provenant du même élément de *A* pour une grappe donnée. C'est le cas, classique, du sondage 'en grappes' pour lequel la seule charge sans biais du graphe de liens est donnée par $\theta_{ik}=1$. Le cas typique de ce genre de sondage est celui où, à partir d'une base de logements, on s'intéresse aux individus qui l'occupent à titre de résidence principale.

Nous aurons aussi besoin des graphes du type 'Grappes en tous pour un' où *A* est partitionnée en grappes, chaque élément de *A* pointant sur un unique élément de *B* , le même pour toutes les unités d'une même grappe.

Si, de plus, *B* est reliée à une troisième population *C* par un second système de flèches, la méthode peut être itérée : on définit des charges σ_{km} et la matrice *S* correspondante ayant les mêmes propriétés (en particulier de donner naissance à des poids sans biais, soit $\mathbf{1}^C=\mathbf{1}^B S$). L'estimateur définit par les poids $w^C=w^BS=w^ATS$ est sans biais et sa variance vaut $D^C=S'T'D^ATS$. .

Figure 1: Population *A* de parents et population *B* d'enfants et liens entre elles.



2-Factorisation et optimalité

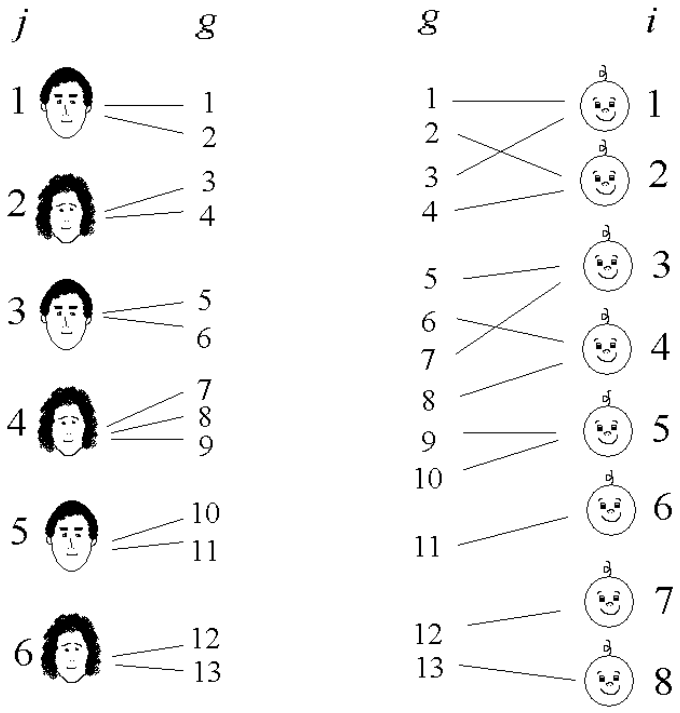
Inversement, tout graphe de liens peut être factorisé de façon canonique. Il suffit de s'intéresser à la population *G* des flèches du graphe de liens entre *A* et *B*. Clairement, chaque unité *i* de *A* est liée de façon évidente à une grappe d'unités de *G* (les flèches d'origine *i*) en un graphe du type 'Grappes en un pour tous'. De même, les unités de *G* sont liées aux unités *k* de *B* en un graphe du type 'Grappes en tous pour un' (flèches aynt pour extrémité l'unité *k*). La composition des deux graphes redonne bien la graphe initial liant *A* à *B* . Pour l'exemple de la figure 1, on obtient la factorisation de la figure 2.

Bien qu'assez artificielle, cette factorisation permet de simplifier grandement la question qui nous occupe dans cet exposé : comment choisir au mieux les charges θ_{ik} du graphe de liens entre les deux populations ?

Un élément de réponse est que nous tenons à conserver la caractère sans biais de l'estimation, ce qui nous donne donc un système de contraintes sur les θ_{ik} : $\mathbf{1}^A T = \mathbf{1}^B$.

La factorisation nous montre qu'on peut se contenter de considérer les graphes du type 'Grappes en tous pour un' , puisque le 'facteur' du type 'Grappes en un pour tous' ne nous laisse aucune liberté de choix.

Figure 2: Résultat de la factorisation pour les populations parents-enfants.



3-Optimalité faible

Avant de chercher des conditions générales d'optimalité, nous allons nous intéresser à un problème plus simple que nous appellerons l'optimalité faible. L'idée est assez simple, voire simpliste. Comme les poids sont sans biais, nous savons que l'espérance de w_k , si k est dans l'échantillon, est égale à l'inverse de sa probabilité d'inclusion qui est, malheureusement, inconnue : $E(w_k \,|\, k \in s_B) = 1/\pi_k$.

Le poids va donc dépendre de l'échantillon s_B particulier qui contient k et nous allons chercher à minimiser ses variations, au sens de la variance, bien sûr, car on ne sait faire que ça (et pas toujours très bien !). Utilisant la factorisation, on doit donc minimiser $\theta_k^{'} D_{kk}^G \theta_k$ sous la contrainte $\mathbf{1_k}\theta_k$ =1. Dans cette expression, θ_k est le vecteur des θ_{gk} cherché, D_{kk}^G la sous-matrice de variance de G correspondant aux lignes et colonnes qui ont des indices g pointant sur k, et $\mathbf{1_k}$ le vecteur ligne de 1 ayant la dimension du nombre de flèches pointant sur k. Sans restreindre la généralité, D_{kk}^G sera inversible, et le problème de minimisation se résout facilement par $\theta_k = \lambda_{kk} D_{kk}^{G\,-1} 1_k$. Ici λ_{kk} est un multiplicateur de Lagrange dont la valeur est :

$$(1_k^{'} D_{kk}^{G\,-1} 1_k)^{-1} = \theta_k^{'} D_{kk}^G \theta_k = Var(w_k) .$$

Ce résultat est toujours relativement facile à appliquer . Nous renvoyons le lecteur à [2] pour le calcul dans des cas particuliers. On verra aussi au point 5 comment on peut en obtenir une approximation immédiate, simple et naturelle, valide dans un large éventail de cas.

4-Optimalité forte

On aimerait bien, cependant, trouver des critères d'optimalité concernant une variable d'intérêt, a priori quelconque, *y*, ce que nous appellerons l'optimalité forte. Le problème est facile à poser, toujours dans le cas d'un graphe de type 'Tous pour un'. Appelons D^G_{kl} la matrice extraite de D^G dont les lignes correspondent aux flèches qui pointent vers *k*, les colonnes à celles qui pointent vers *l*. On cherche à minimiser :

$$\sum_k \sum_l y_k y_l \theta_k^{'} D^G_{kl} \theta_l$$

sous les contraintes $\mathbf{1}_k \theta_k = 1$ pour tout *k*.

On peut poursuivre la recherche d'optimisation avec la méthode de Lagrange, ce qui conduit aux équations :

$$y_k \sum_l D^G_{kl} \theta_l y_l = \lambda^y_k \mathbf{1}_k' \qquad (*).$$

Malheureusement, la solution de ce système dépend, en général, explicitement de la variable *y*, et sa résolution est donc impossible en pratique. On peut cependant chercher à quelles conditions l'optimisation peut conduire à un résultat (les vecteurs θ_k) indépendant de *y*, et donc calculable à partir uniquement des D^G_{kl} .

Pour cela, on peut observer que l'optimalité faible peut se concevoir comme relative à une variable très particulière qui vaudrait 1 pour une certaine valeur de *k*, et 0 sinon. La valeur de θ_k, pour l'optimalité forte, est donc nécessairement celle de l'optimalité faible.

Compliquant légèrement les choses, intéressons nous à une variable valant 1 pour les valeurs *k* et *l*, et 0 sinon. En utilisant toujours la technique de Lagrange, on obtient les conditions nécessaires suivantes :

$$D^G_{kk} \theta_k + D^G_{kl} \theta_l = \lambda^{kl}_k \mathbf{1}_k$$
$$D^G_{ll} \theta_l + D^G_{lk} \theta_k = \lambda^{kl}_l \mathbf{1}_l$$

Comme nous avons l'optimalité faible, on a $D^G_{kk} \theta_k = \lambda^{kk}_k 1_k$, et compte tenu du fait que les deux égalités précédentes sont équivalentes (avec en particulier égalité des multiplicateurs de Lagrange), ceci nous conduit aux conditions :

$$D^G_{kl} \theta_l = \phi^{kl} \mathbf{1}_k \quad \text{(valides aussi si } k=l\text{)} \qquad (**)$$

Soit $\mathbf{\Phi}$ la matrice d'éléments $\Phi_{ii'}$. On peut maintenant facilement montrer que la variance optimale est donnée par $\mathbf{Y}'\mathbf{\Phi}\mathbf{Y}$, où $\mathbf{Y}$ est le vecteur des y_k .

De plus, un raisonnement assez simple (refaire les calculs 'à l'envers') montre que les conditions (**) sont suffisantes pour avoir l'optimalité forte.

Bien qu'apparemment assez particulières, ces conditions se rencontrent assez souvent dans la nature, mais elle dépendent beaucoup des liaisons entre les flèches du graphe de correspondance et le plan de sondage sur *A*. On renvoie le lecteur à [2] pour une étude approfondie. Pour ce qui nous concerne ici, nous allons nous contenter d'un exemple concret.

De façon opérationnelle, nous avons donc établi des conditions d'optimalité faible assez faciles à mettre en œuvre, et qui donnent les valeurs nécessaires des charges. Si, éventuellement, on est, en plus , dans les conditions d'optimalité forte, ces valeurs restent les mêmes. On se contentera donc de cela, avec donc un fort bonus dans le cas où, par chance, on se trouverait dans le cas fort sans avoir les moyens de le vérifier !

5- Application à l'enquête sur le tourisme en Bretagne

Pour attraper des touristes en Bretagne (ou ailleurs si on en a envie, voir l'exposé de Myriam Maumy, JMS 2005), on se sert de plusieurs sondages relatifs à leurs activités, qui, après études, tests et hésitations, se ramènent au schéma suivant :

- Un sondage sur les visites de certains sites (Carnc,Fréhel,le péage de La Gravelle !).On réalise un sondage à deux degrés par site : premier degré le jour d'enquête, deuxième degré : le ménage touristique.
- Un sondage à trois degrés sur les achats en boulangerie (premier degré : la boulangerie, puis le jour, puis le ménage).

Bien que ce ne soit pas nécessaire, on va supposer que chacun de nos touristes k est attrapé par un seul de ces dispositifs d, ce qui lui confère un 'poids de base' w_d . Ici, encore pour simplifier, nous allons supposer que ces poids sont ceux d'Horvitz-Thompson, soit l'inverse de la probabilité de sélection dans le dispositif d.

Nous savons que ce touriste a effectué n_b achats en boulangerie, ainsi que la liste des sites visités (une seule fois aussi par hypothèse, mais ce n'est pas indispensable). Soit L_k la liste des occurrences où notre ménage touristique a pu être attrapé, n_k son cardinal. Nous devons construire la matrice $n_k \, x \, n_k$ des variances et covariances entre ces occurrences. C'est assez facile car :

- les trois dispositifs sont indépendants, donc les éléments de matrices hors diagonale correspondants sont nuls. Sur la diagonale, nous avons une quantité du type $\frac{\pi_i - \pi_i^2}{\pi_i^2}$ connue par le plan dans le dispositif correspondant. Pour les $n_b \, x \, n_b$ entrées venant des boulangeries, la partie diagonale s'obtient de la même façon. Comme n_b est très petit devant le nombre total d'achats de la 'base de sondage', les éléments hors diagonale sont négligeables devant les éléments diagonaux. Notre matrice est donc très voisine d'une matrice diagonale d'éléments $\frac{\pi_i - \pi_i^2}{\pi_i^2} \cong \frac{1}{\pi_i}$. On obtient donc, approximativement, θ_{ik} proportionnel à π_i et donc égal à $\theta_{ik} = \frac{\pi_i}{\sum_{l \in L_k} \pi_l}$. Le poids du ménage touristique sera donc égal à $w_d \;\; \theta_{dk} = \frac{1}{\sum_{l \in L_k} \pi_l}$. On trouve un résultat somme toute assez naturel. En particulier, si un touriste ne déclare qu'une seule occurrence ayant permis de l'attraper, son poids est celui d'Horvitz-Thompson .

- Si notre touriste a été attrapé q_k fois dans le dispositif, avec les mêmes approximations, il recevra le poids $\frac{q_k}{\sum_{l \in L_k} \pi_l}$. On voit, alors, qu'il n'est même pas nécessaire de vérifier ce doublon : si on crée q_k enregistrements dans le fichier de dépouillement, chacun muni du poids $\frac{1}{\sum_{l \in L_k} \pi_l}$, les résultats de toute analyse statistique seront les mêmes, y compris les estimations de variance !

.

Pour donner une image à titre d'exemple, supposons qu'on ait $n_b=4$ achats en boulangerie, avec des poids, pour simplifier, de sondage aléatoire simple N/n , et deux sites visités avec des poids de sondages aléatoires simples N_{1ou2}/n_{1ou2} . la matrice extraite aura la forme suivante :

$$D_{kk}=$$

$$\begin{pmatrix}
\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & 0 & 0 \\
-\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & 0 & 0 \\
-\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & \frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & 0 & 0 \\
-\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & -\frac{1}{N-1}\frac{N}{n}(1-\frac{n}{N}) & \frac{N}{n}(1-\frac{n}{N}) & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{N_1}{n_1}(1-\frac{n_1}{N_1}) & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{N_2}{n_2}(1-\frac{n_2}{N_2})
\end{pmatrix}$$

soit

$$D_{kk}\cong\begin{pmatrix}
N/n & -1/n & -1/n & -1/n & 0 & 0 \\
-1/n & N/n & -1/n & -1/n & 0 & 0 \\
-1/n & -1/n & N/n & -1/n & 0 & 0 \\
-1/n & -1/n & -1/n & N/n & 0 & 0 \\
0 & 0 & 0 & 0 & N_1/n_1 & 0 \\
0 & 0 & 0 & 0 & 0 & N_2/n_2
\end{pmatrix}$$

On doit résoudre $D_{kk}\theta_k=\mathbf{1_k}$ à un facteur près pour que la somme des éléments du vecteur θ_k fasse 1, ce qui nous donne approximativement $(n/N)/S$ pour les 4 premières coordonnées, puis $(n_1/N_1)/S$ et $(n_2/N_2)/S$ pour les deux dernières, avec $S=4*n/N+n_1/N_1+n_2/N_2$. Où que notre client ait été attrapé, il recevra donc le poids $1/S$.

Bien évidemment, ces approximations sont valides dans un champ bien plus général que l'enquête auprès des touristes en Bretagne, à savoir les cas où le nombre de flèches pointant sur un individu de B est négligeable devant l'inverse des taux de sondage.

Références :

[1] Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Editors), John Wiley and Sons, New York, pp. 135-159.

[2] Deville, J.C.,et Lavallée, P. (2005). Indirect Sampling: The Foundations of the Generalized Weight Share Method, *Manuscrit soumis à une éventuelle et aléatoire publication*

[3] Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes, n°84-85-86, Actes des Journées de methodologie statistique des 17-18 mars 1998,* pp 63-82.

[4] Deville, J.C.,et Maumy, M., (2005) Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme, *Manuscrit soumis à une éventuelle et aléatoire publication*

[5] Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, Vol. 21, No. 1, pp. 25-32.

[6] Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.