Équilibrage et calage *a priori*

V. Loonis[1], X. Mary[2]

Hájek (1981) définit une stratégie d'échantillonnage et d'estimation d'un total comme un couple (P,w) où P est un plan de sondage et *w* un système de pondération. Il propose alors de choisir un tel couple de manière à ce que l'estimateur linéaire homogène pondéré par *w* conduise à des estimations parfaites, sous P, des totaux d'un ensemble de variables auxiliaires. La stratégie est alors qualifiée de représentative pour les variables considérées. Dans la pratique, les composantes d'une stratégie sont fixées séparément.

Le plan de sondage P est généralement choisi, sous contraintes organisationnelles ou statistiques, de manière à être représentatif d'un ensemble de variables pour un système de pondération w, non aléatoire, correspondant à celui de l'estimateur d'Horvitz-Thompson. Ce contexte est celui, notamment, du tirage équilibré décrit par Deville et Tillé (2004). Respectivement, à P fixé, un système de pondération aléatoire : *w(S)*, dépendant de l'échantillon aléatoire *S*, peut être défini, a posteriori, de manière à ce que l'estimation linéaire mobilisant w d'un ensemble de variables soit représentative. Ce contexte est celui du calage décrit par Deville et Särndal(1992).

On ne trouve pas dans la littérature, à notre connaissance, de méthodes permettant de fixer simultanément les deux composantes d'une stratégie. Cette situation provient d'une part de la difficulté qu'il y a à exprimer de manière analytique des quantités telles que les probabilités d'inclusion double, pour des probabilités d'inclusion simple quelconques (Chen et al. 1994), alors que ces quantités peuvent intervenir dans la recherche de stratégie représentative. Par ailleurs, quand le sytème de pondérations est également aléatoire, c'est la loi, encore plus inaccessible, du couple (P,w) qui doit être connue

Dans cette présentation on propose une méthode permettant de fixer simultanément P et w. On introduit pour cela une famille paramétrée P_θ : de plans de sondages dont les probabilités d'inclusion simple, double, triple…, sont des fonctions explicites du paramètre. On présente dans un premier temps cette famille ainsi qu'un algorithme de sélection respectant strictement P_θ . Dans un deuxième temps, on recherche une stratégie représentative pour un ensemble de variables en minimisant, en θ et w, un critère d'erreur quadratique moyenne, en contraignant θ à appartenir à un espace Θ, éventuellement de grande dimension, et w à ne pas être aléatoire.

Avec cette méthode, en imposant l'utilisation de l'estimateur d'Horvitz-Thompson, on trouve, en minimisant en θ, un problème d'équilibrage dans notre famille de plans. À θ fixé, la minimisation en w s'apparente à la recherche d'un estimateur calé, la grande différence avec la méthode de Deville et Särndal étant que les poids ne sont pas aléatoires et peuvent être définis a priori. Ces deux spécificités justifient, à nos yeux, le qualificatif d'équilibrage et calage a priori.

[1] Insee : Division des Méthodes et Référentiels Géographiques
[2] Université de Paris Ouest Nanterre La Défense