

## Construction d'un échantillon dispersé géographiquement

Marc CHRISTINE

INSEE, Unité Méthodes Statistiques (UMS)

Dans la pratique des tirages d'échantillons d'enquêtes ménages tels que les effectués par l'Insee, il est d'usage de chercher à concentrer l'échantillon sur certaines zones géographiques de manière à minimiser le déplacement des enquêteurs lors des enquêtes réalisées en face à face. En même temps, on a souvent le souhait d'assurer une certaine dispersion des points d'enquête, partant du pré-supposé intuitif qu'une bonne dispersion géographique de l'échantillon permettra d'assurer une couverture adéquate de la diversité du territoire sur lequel on échantillonne.

La réconciliation de ces deux souhaits conduit à la mise en place du système d'unités primaires qui sert de substrat aux échantillons d'enquêtes ménages réalisés par l'Insee.

Dans ce papier, on s'intéresse au problème de la construction et de la sélection d'un échantillon d'unités en général localisées (communes, adresses, logements...), assurant une dispersion géographique adéquate des unités tirées.

La dispersion géographique d'un échantillon sera évidemment mesurée par la quantité :

$$\frac{\sum_{i \in s} d_i g G_i^2}{\sum_{i \in s} d_i},$$
 où les  $G_i$  sont les points géométriques correspondant aux unités échantillonnées,

repérées dans un système de coordonnées orthométriques ad hoc et  $g$  leur centre de gravité vis-à-vis d'un système de poids  $d_i$ , qui seront en général les poids de sondage. Il s'agit donc tout simplement de la moyenne pondérée des carrés des distances des points échantillonnés par rapport au centre de gravité métrique de l'échantillon.

La dispersion géographique de l'échantillon sera considérée comme adéquate si et seulement si elle est égale à la vraie dispersion des unités dans la population de référence  $U$  (de taille  $N$ ), soit :

$$\frac{1}{N} \sum_{i \in U} G_i^2,$$
 où  $G$  est le centre de gravité de l'ensemble des points  $G_i$  lorsque  $i$  décrit  $U$ .

Le papier montre alors que la technique de sélection d'échantillons équilibrés au moyen de la méthode du CUBE permet d'atteindre cet objectif : le choix de variables d'équilibrage ad hoc permet, dans un premier temps, de tirer des échantillons dont la moyenne et la variance pondérées, pour des variables d'intérêt données, sont identiques aux vraies moyennes et variances des variables considérées dans la population. L'application au problème géométrique exposé ci-dessus s'en déduit alors naturellement.

Cette méthode a été employée lors du tirage des enquêtes devant être réalisées dans les départements d'Outre-mer en 2005 (Budget de famille et Emploi).