Algorithme CURIOS et méthode de priorisation pour les enquêtes en face-à-face - Application à l'enquête Patrimoine 2014

Domaines : Échantillonnage, Méthodologie de collecte

Dans un contexte de dégradation des taux de collecte dans les enquêtes ménages, l'INSEE cherche à utiliser au mieux les ressources disponibles. Il s'agit, étant donnés les moyens alloués à une enquête, d'obtenir l'échantillon collecté contenant le plus d'information possible (et conduisant *ex post* à la variance la plus faible possible). Un point de départ est le concept de R-indicateur, développé dès 2009 par Schouten, qui permet de construire dans des enquêtes par téléphone des échantillons de répondants représentatifs de la population.

Les R-indicateurs ont été initialement développés pour permettre la priorisation d'efforts de relance dans des enquêtes dont la collecte s'effectue par téléphone. Contrairement à ce cadre, la collecte en face-à-face ne permet pas un réajustement réactif des efforts de collecte, principalement car les enquêteurs organisent sur plusieurs semaines la collecte des unités qui leurs sont affectées.
La solution choisie est de réaliser l'enquête **en deux vagues**. L'échantillon de vague 2 est tiré en prenant en compte le portrait de la collecte réalisée en vague 1. Il est tiré avec l'objectif d'équilibrer la collecte (et de minimiser la dispersion des poids) à la fin de la vague 2, en supposant que les conditions de collecte restent identiques entre les deux vagues.
Ce principe a été mis en place pour l'enquête Patrimoine 2014 en région Île-de-France.

D'un point de vue technique, il s'agit d'un problème d'exercice optimal à la date 0 d'une stratégie dont les résultats sont attendus à la date T, la quantité optimisée pouvant évoluer entre les dates (ultérieures) 0 et T. L'idée est d'utiliser les R-indicateurs de manière à tirer l'échantillon donnant la prévision de collecte optimale, c'est-à-dire équilibrée au sens des R-indicateurs, et avec la dispersion des poids corrigés de la non-réponse la plus faible possible, en simulant la collecte avec les nouvelles probabilités de tirage. On intègre à l'algorithme une phase de correction de la non-réponse par Groupes de Réponse Homogène de façon anticipée. La fonction définie par ces paramètres en fonction du vecteur de "sur-représentation" possède de bonne propriétés, et son optimum peut être déterminé par optimisation linéaire.
L'algorithme est baptisé CURIOS (Curios Uses Representativity Indicators to Optimize Samples).

Cette méthode permet de s'affranchir du risque de "trous de collecte" (modalité ou zone géographique pour laquelle le taux de collecte est tellement bas que les risques de biais ne sont plus négligeables) et enrichit le monitoring de collecte par R-indicateur d'un contrôle de la dispersion des poids corrigés de la non-réponse, ce qui permet de diminuer la variance *ex post*.