Journées de méthodologie statistique de l'Insee 2022

TROIS PETITS PAPIERS EN QUÊTE D'AUTEUR

Luigi PIRANDELLO (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

luigi.pirandello@insee.fr

Mots-clés : Échantillonnage, calage, imputation, modèle linéaire, revenu

Domaine concerné : calage sur marges, imputation, modélisation stochastique

Résumé

Papier n°3 : Du revenu déclaré en tranches au vrai revenu fiscal, ou comment estimer un modèle économétrique sans disposer d'observations

Une variable essentielle explicative du comportement des ménages dans de multiples domaines est le revenu du ménage. La difficulté est de disposer d'une observation correcte de cette variable. Lorsque les enquêtes sont échantillonnées dans les bases fiscales, on peut disposer d'une donnée *vraie*, le revenu déclaré dans l'IRPP.

Mais des enquêtes, soit plus anciennes, échantillonnées dans les bases issues du recensement de la population (systèmes Octopusse, puis Nautile) ou sur des sujets spécifiques nécessitant des bases de sondage appropriées à la thématique ne permettent pas de disposer de cette information.

Celle-ci est alors collectée dans l'enquête elle-même ou dans le cadre commun qui l'enserre, tel que le Tronc Commun des Enquêtes Ménages (TCM). Récemment, l'auteur s'est penché sur les enquêtes sur le Suivi de la demande Touristique, actuellement sous-traitées à un prestataire privé, tant au niveau de l'échantillonnage que de celui de la collecte, et tirées dans des bases de sondages tout à fait imparfaites.

Dans ce cadre, la donnée *observée* est évidemment entachée de plusieurs erreurs ou imperfections.

- le revenu déclaré dans le TCM ou dans l'enquête présente en général des sous-estimations systématiques par rapport au revenu vrai (non observé).
- il pose des questions de concept, de champ, de définition, de périmètre (inclut-il ou non les revenus de redistribution, mobiliers, sans parler de la prise en compte ou non de la retenue à la source..), quel contour du ménage utilise-t-il ?
- Il est sujet à de la non-réponse.
- **le déclaratif est incomplet** ; les ménages ont la possibilité de répondre en clair ou, à défaut, en tranches ou seulement sur le second mode.

Ce papier se propose de fournir des solutions à deux questions :

- peut-on, au niveau individuel, à partir de données déclaratives observées sur le revenu, « reconstituer » un revenu vrai au sens du revenu fiscal (supposé *inobservable*) ? On montrera que, si l'on dispose seulement de la *distribution* des revenus vrais et de l'observation des revenus déclarés en tranches, il est possible, à partir d'un modèle linéaire, d'imputer un revenu vrai estimé à chaque unité statistique.

- si l'on dispose de sources fiscales constituant une donnée exogène exacte sur la distribution vraie des revenus, peut-on et comment utiliser ces informations pour les incorporer dans le processus de calage des enquêtes ? De surcroît, on s'intéresse en général non seulement au revenu total (ou revenu moyen) des ménages, mais aussi à la *distribution* des revenus et le calage doit intégrer ces deux dimensions, ce que montrera le papier.

On n'abordera dans ce papier ni les questions conceptuelles ou de champ (on supposera que données collectées et données vraies correspondent au même concept et au même champ), ni celles relatives au traitement de la non-réponse[1] et encore moins celles de l'analyse et la correction des points aberrants : on se placera ici en aval de ce processus d'imputation.

Ces petits papiers cherchent un (co-)auteur pour mettre en application les méthodes exposées sur des données réelles, tester leur pertinence et apporter tout complément utile….

Bibliographie

[1] Christine M., rapport du Groupe Marges, version révisée novembre 2013, note interne UMS, *unpublished paper*
[2] Vincent L., Faivre S., "Le projet Nautile (Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes)", Actes des 13èmes Journées de méthodologie statistique de l'Insee

[1] En général, un processus spécifique est tout d'abord opéré sur les données d'enquête (à l'aide de la méthode des résidus simulés, au moyen d'équations économétriques mettant en jeu diverses caractéristiques socio-démographiques des ménages, estimées sur les répondants en clair et appliquées aux autres ménages) : ce processus réalise une **imputation**, à partir des données déclarées par les répondants en clair, à la fois pour les répondants en tranches et pour les non-répondants. Dans le cas d'une réponse en tranches, on s'assure que la valeur simulée se situe bien dans la tranche déclarée.