Le nécessaire échantillonnage des données fiscales : le cas de l'impôt sur le revenu

L. Aeberhardt, L. Guenchi, R. Lardellier (DGFiP, bureau GF-3C)

Le bureau GF-3C est depuis 2011 reconnu en tant que service statistique ministériel. Il produit des analyses sur l'ensemble du domaine de la fiscalité, aussi bien en fiscalité des entreprises (TVA, impôt sur les sociétés), fiscalité des particuliers (impôt sur le revenu, droits de mutations, ISF) qu'en fiscalité locale (taxes d'habitation et foncière). L'étude que nous proposons ici porte sur les outils et les données utiles à la production des travaux en matière d'impôt sur le revenu (IR). La section en charge de l'IR a notamment pour missions, d'une part, de participer aux travaux de prévisions de recettes fiscales d'IR (*i.e.* prévision d'un montant global) et, d'autre part, d'estimer le coût de dispositifs fiscaux spécifiques (évaluations *ex ante* ou *ex post*). Pour tous ces travaux, le premier enjeu consiste à concilier précision des résultats et respect de délais contraints.

Pour chaque foyer fiscal français soumis à l'IR, la base de données de la DGFiP recense l'ensemble des informations déclarées, variables de gestion et variables calculées. Elle est ainsi constituée de plus de 36 millions de foyers et d'environ 3 000 variables. Les données portant sur les revenus n sont disponibles par vagues (« émissions »), de juillet n+1 à février n+2. La richesse d'une telle base est évidente mais elle ne va pas sans difficultés dès qu'il s'agit de mettre en place des procédures complexes.

Ainsi, en matière de prévision des recettes d'IR ou d'évaluation de dispositifs fiscaux, le bureau GF-3C procède par microsimulation. Un programme retranscrit fidèlement la législation afin de simuler, pour chaque foyer, le calcul de l'impôt. La législation de l'IR étant complexe, ce programme implique de très longs temps de traitement (1h/100 000 foyers en législation 2013, 1h45/100 000 foyers en législation 2014). Dès lors, en vue d'effectuer ces travaux de microsimulation, il est impensable de travailler à partir de la base exhaustive.

Contrairement aux problématiques courantes en théorie des sondages, nous ne cherchons pas à estimer au mieux des grandeurs inconnues, mais à réduire le plus fidèlement possible notre population afin de recalculer l'IR le plus précisément possible et d'en déduire les coûts de dispositifs fiscaux. En plus de la nécessité de réduire les temps de traitements, nos objectifs sont à la fois le maintien d'une bonne précision globale, une reproduction fidèle de la distribution de certaines variables fondamentales dans le calcul de l'impôt, mais également un court délai pour la production de l'échantillon.

Cette problématique n'est pas nouvelle au sein du bureau GF-3C. En 1997, les premiers travaux d'échantillonnage ont été réalisés, adaptés aux contraintes informatiques de l'époque. Ils ont abouti à la mise en place de deux échantillons, un échantillon dit « léger » (50 000 foyers) et un second dit « lourd » (500 000 foyers). Ces deux échantillons reposaient sur une méthode identique, à savoir un sondage stratifié avec allocation de Bankier et tirage systématique avec tri préalable. Hormis quelques modifications mineures et l'abandon de l'échantillon léger, cette méthode est toujours utilisée pour produire les échantillons d'IR.

Aujourd'hui, l'intérêt d'une remise à plat de cette méthodologie est multiple. Il convient d'abord de l'adapter aux nouvelles caractéristiques de l'IR ainsi qu'aux utilisations devenues courantes de l'échantillon. Dans ce cadre, il est nécessaire de se poser les bonnes questions concernant les déterminants de l'IR dans un contexte législatif évoluant chaque année, en particulier sur la sélection des variables auxiliaires dont nous souhaitons reproduire la distribution. L'idée est surtout de pouvoir profiter des nouveaux outils disponibles pour la production d'échantillons, notamment les méthodes d'échantillonnage équilibré développées par l'Insee (Cube et FastCube). Par ailleurs, contrairement à ce qui était fait précédemment, une production automatisée de l'échantillon permettrait un gain de temps non négligeable lors de la réception des différentes données. Enfin, à long terme, une recherche sur la taille optimale de l'échantillon réduirait les temps de traitement lors des microsimulations.

Dans un premier temps, nous présenterons les bases de données et les différents travaux qu'il est possible de réaliser, soit à partir des données exhaustives, soit à partir d'un échantillon. Dans un deuxième temps, en partant de l'ancienne méthode d'échantillonnage nous montrerons les évolutions qui peuvent être apportées à la marge et nous questionnerons les critères fondamentaux à prendre en compte pour refondre la méthode. Enfin, nous testerons le passage à un tirage équilibré, en vue de créer une procédure d'échantillonnage valable pour d'autres données fiscales.