Mardi 15 mars 2005 : 10h15-10h40

Imputation de distributions dans les données fiscales administratives.

Rong HUANG[1], *Dominique LADIRAY*[2]
[1] Statistique Canada
[2] INSEE, DESE

Statistique Canada s'est résolument engagé depuis plusieurs années dans l'utilisation de données fiscales administratives pour diminuer la taille des échantillons des enquêtes et alléger ainsi la charge de réponse des entreprises.

Les données fiscales, qui sont transmises régulièrement par l'Agence du Revenu du Canada, sont en principe exhaustives. Mais elles présentent les défauts habituels - non réponse partielle, valeurs atypiques etc. - et doivent être analysées et redressées avant d'être utilisables par les divisions clientes.

En particulier, les déclarations fiscales portent en général plus d'attention aux totaux qu'aux ventilations, souvent facultatives. Lorsque qu'un poste de la déclaration ne comporte que le total (appelé « générique ») de la recette ou de la dépense, une procédure d'imputation est alors mise en œuvre pour estimer la répartition en sous-postes (appelés « détails »).

La procédure actuelle est basée sur un découpage a priori de la population des entreprises répondantes en groupes définis par le code activité (code SCIAN) et par la taille de l'entreprise. La distribution marginale par détails des entreprises répondantes à l'intérieur d'une classe est alors utilisée pour les entreprises de la même classe qui ne fournissent pas les détails. Cette estimation par le ratio repose sur une hypothèse forte : les entreprises d'une classe donnée sont supposées avoir des comportements très voisins pour que la même répartition moyenne s'applique à toutes. Ce n'est malheureusement pas le cas et, de plus, l'utilisation d'une répartition moyenne peut aboutir à des imputations étranges : telle entreprise de restauration rapide s'est ainsi vu attribuer à tort des dépenses en boissons alcoolisées.

La communication proposée présente une méthodologie alternative laissant aux données le soin de définir les classes qui sont déterminées par une classification ascendante hiérarchique sur les valeurs des détails observées. Le problème est ensuite d'affecter une entreprise non répondante à l'une de ces classes homogènes par construction. Plusieurs procédures d'affectation, basées sur des variables explicatives disponibles dans la déclaration fiscale, sont comparées, sur données brutes ou discrétisées : analyses discriminantes paramétrique et non-paramétrique, modèles log-linéaires etc.

Les règles d'affectation ainsi obtenues sont évaluées et comparées par validation croisée et bootstrap. Enfin, l'efficacité globale des différentes procédures d'affectation associées elles-mêmes à des méthodes d'imputation différentes (ratio ou donneur) sont validées sur les données de l'année précédente.