Mise à jour de répertoires d'entreprises

Une des tâches fondamentales que doit assurer un Institut de statistique est de gérer différentes bases de sondage, celles-ci étant ensuite utilisées lors de la mise en place des processus de production statistique. Et, problème bien connu des statisticiens, chaque base de sondage est potentiellement affectée par un certain nombre de défauts : problèmes de sous-couverture, problèmes de sur-couverture (avec des unités cessées mais dont on ne sait pas qu'elles le sont), problèmes de doublons, mauvaise qualité des informations disponibles qui sont candidates pour être des variables utilisées pour la stratification d'un plan de sondage.

Face à cette difficulté, le statisticien est rarement totalement démuni, car il dispose d'informations permettant de cibler, de manière plus ou moins précise, les « zones » qu'il serait bon de contrôler en priorité : par exemple, informations sur la fraîcheur des données contenues dans la base de sondage (à partir du moment où on a conservé trace de celle-ci, bien sûr), informations sur la plus ou mois grande fiabilité des caractéristiques contenues dans la base de sondage (par exemple suite à une enquête qui va permettre de mettre en regard ces informations avec l'information actualisée, obtenue via l'enquête).

Si l'on s'intéresse au domaine des statistique d'entreprises, la base de sondage est généralement constituée à partir d'un répertoire d'entreprises. Celui-ci contient des éléments d'identification (identifiant de l'entreprise, raison sociale, adresse), mais également des critères permettant de stratifier les plans de sondage : souvent, la stratification croise un critère qualitatif (classement de l'entreprise dans un secteur d'activités, en référence à la nomenclature d'activités, NAF au niveau français ou NACE au niveau européen) et un critère de taille établi à partir d'une variable quantitative (chiffre d'affaires, ou effectif salarié). C'est le cas à l'INSEE pour le répertoire SIRENE. Celui-ci est un répertoire inter-administratif : en particulier l'identifiant géré par l'INSEE est utilisé par l'ensemble des administrations françaises ayant des contacts avec les entreprises.

Ceci a pour conséquence d'éviter les problèmes de sous-couverture et de doublons. En revanche, les problèmes de sur-couverture et de qualité des caractéristiques contenues dans le répertoire sont présents. Un certain nombre d'opérations de nettoyage sont donc menées chaque année, en fonction d'objectifs spécifiés a priori (par exemple, vérification du bon classement dans la NAF pour les unités contenues dans deux postes de la nomenclature qui présentent une certaine proximité). Mais, à notre connaissance, il n'existe pas de démarche formalisée permettant de guider la mise en place de ces opérations qualité.

L'objectif du papier présenté est de proposer une formalisation tenant compte d'éléments quantifiés disponibles pour guider le travail des gestionnaires du répertoire. Plus précisément, l'idée est de s'appuyer sur des travaux anciens menés sur la probabilité pour une entreprise d'être active (document à lucarne E2008/06, Henri Mariotte), et sur des résultats d'enquêtes statistiques, ceci afin de définir des populations d'entreprises sur lesquelles cibler des opérations qualité, en fonction d'objectifs définis a priori (par exemple, mise en place d'une enquête destinée à estimer le chiffre d'affaires de tel ou tel secteur).