Esane : À la recherche d'une cohérence maximale des données multi-sources sur les entreprises par le biais de micro et macro contrôles

Olivier HAAG[1]

En 2006, un grand projet de refonte de l'élaboration des statistiques structurelles (le programme Resane) a été mis en œuvre à l'Insee. Ces statistiques se fondent sur la prise en compte de plusieurs sources d'informations ;

- des sources administratives (déclarations fiscales et sociales)
- une source statistique : l'Enquête Sectorielle Annuelle (ESA) qui se décline par des questionnaires personnalisés selon les différents secteurs économiques.

Afin de conserver une certaine adaptabilité du système face à cette diversité de sources qui peuvent évoluer indépendamment de l'Insee, le système d'information est conçu de façon modulaire. Chaque source subit de façon indépendante, une chaine de contrôle. L'objectif de cette chaîne est d'optimiser le partage des tâches entre l'homme et la machine pour rechercher et corriger les erreurs qui se sont glissées dans l'information collectée. Ne sont confiés à l'action humaine que les dossiers pour lesquels l'automatisation est moins performante. L'automatisation trouve ses limites surtout en matière d'entreprises dont le poids économique relativement à l'ensemble de la population est important.

Ainsi cette chaîne de contrôle se déroule en deux étapes :

- la première a pour objectif de traiter la non-réponse partielle et de corriger des valeurs jugées aberrantes. Elle est entièrement automatique et met en œuvre des micro-contrôles de cohérence et de vraisemblance puis, le cas échéant, des redressements automatiques se basant sur la réponse de l'année précédente ou sur les réponses des autres entreprises appartenant à la même strate de contrôle (obtenue par croisement entre l'activité principale et la tranche de taille de l'entreprise).
- la seconde vise à identifier les unités qui doivent être contrôlées par un gestionnaire. Cette sélection s'opère à l'aide de macro-contrôles qui vont permettre d'identifier les unités les plus fortement contributrices aux agrégats diffusés.

Une fois les caractéristiques des différentes sources validées, il convient d'assurer une cohérence globale de ces caractéristiques multi-sources. C'est le but du processus de réconciliation des données individuelles. Il s'assure que les caractéristiques communes à deux sources ont in fine la même valeur.

Dans le cas où il existe une divergence initiale sur ces caractéristiques d' « accrochages », le processus affecte une valeur unique après un arbitrage automatique qui privilégie une source selon différentes règles de priorité. Cette réconciliation concerne le chiffre d'affaires et sa ventilation en type de ventes mais également l'emploi et les rémunérations.

Ce processus automatique s'accompagne d'un contrôle manuel des plus gros écarts initiaux. Ces derniers sont identifiés par le biais de macro contrôles. Une fois la valeur unique

[1] olivier.haag@insee.fr

Insee : Département« Répertoires, Infrastructure et statistiques structurelles »

déterminée, elle est réinjectée dans chaque source dans laquelle elle figure et d'éventuels redressements automatiques ont lieu afin de rétablir les cohérences intra-source.

Enfin, en statistique d'entreprises, les restructurations économiques (fusion, scission etc.) ont une très forte influence sur les résultats diffusés. Le processus de production ESANE essaye d'appréhender ce phénomène dès le contrôle des données en analysant les réponses des entreprises par le biais d'une nouvelle unité statistique spécifique : l'enveloppe de restructuration. Cette nouveauté permet d'améliorer la diffusion des résultats en évolution.