La recodification du recensement 1999 en Naf rev.2 Principes et élaboration d'un outil pour calculer la précision

Benoît BUISSON[1]

Le bureau du Comité d'Orientation Pour l'Action Régionale (COPAR) de l'INSEE a passé commande, en septembre 2009, au pôle ingénierie statistique entreprises (PISE) d'une étude pour codifier les codes d'activités principales des établissements (APET) du recensement 1999 en Naf rev.2 à un niveau agrégé. Cette codification était jugée nécessaire pour calculer des évolutions d'emploi entre les années postérieures à 2006 et 1999. Pour cela il était demandé au pôle ISE de se servir directement de la table de passage utilisée pour double coder le recensement 2006, sans retourner au niveau individuel des établissements.

Cet article se propose de détailler, dans un premier temps, la méthode utilisée pour assurer cette codification en nouvelle nomenclature. Une fois la table de passage complètement définie, le principe général a été d'affecter un code d'activité en Naf rev.2 selon le même principe que le traitement de la non-réponse partielle dans une enquête, à savoir par un mécanisme d'imputation aléatoire. Cette imputation aléatoire a été réalisée à différents niveaux de la nomenclature. L'analyse des taux d'imputation par zone géographique de référence a permis de choisir le niveau de nomenclature à privilégier pour cette recodification (niveau A_129). Il a également été nécessaire de réaliser des imputations déterministes sur certaines zones, avec des employeurs particulièrement importants dont nous avons déterminé individuellement l'activité.

Une fois l'imputation réalisée, il était important de définir des critères de fiabilité statistique de cette opération vis à vis des utilisateurs internes de l'INSEE (chargés d'études en région notamment). Des règles prudentielles ont été établies pour juger de la fiabilité de l'imputation, en prenant en compte, sur la zone étudiée, la taille de la population à estimer (nombre d'emplois par activité) ainsi que le taux d'imputation non presque sure. Ces règles ont été établies par le pôle ISE à partir de l'examen d'un certain nombre de cas choisis sur le territoire national, cas jugés représentatifs de l'éventail des situations qui pourraient être rencontrées. Afin d'aller plus loin pour juger de la fiabilité statistique des estimateurs, le pôle ISE a conçu et mis à disposition une macro SAS qui permet de calculer à la demande des indicateurs de précision à partir des techniques du bootstrap. L'idée générale de ce type d'outil est de réaliser l'imputation un grand nombre de fois (500) sur une zone géographique et à un niveau de nomenclature choisis par l'utilisateur. La macro SAS fournit en sortie des indicateurs de fiabilité et de dispersion : moyenne, médiane, valeurs extrêmes et coefficient de variation notamment. Des exemples d'application seront décrits pour montrer l'intérêt complémentaire de cet outil par rapport aux règles prudentielles. Il est à noter que cet outil permet uniquement de juger de la fiabilité statistique de la codification et non de son adéquation économique. Pour cela il est nécessaire de mener une expertise locale à partir de connaissances sur le tissu productif.

[1] benoit.buisson@insee.fr