Qualité du codage de l'activité et de la profession dans le recensement de la population

Yves JACQUIN[1], Elodie MARTAL[2]

Connaître la qualité du codage de l'activité et de la profession est important pour les études et la diffusion des données du recensement de la population (RP). Pour mieux maitriser la qualité de son processus, une enquête de mesure de la qualité de la codification a été mise en place en 2006.

Le procédé de codage des données diffusées est le suivant : les variables sont d'abord codées de manière automatique (ce traitement automatique concernant un peu moins de 50% des bulletins pour l'activité, et 75% pour la profession actuelle), puis les rejets du codage automatique sont repris manuellement par des gestionnaires en directions régionales (via une application dénommée RECAP = REcherche et Codage de l'Activité et de la Profession).

Pour mesurer la qualité de ce codage, des opérations baptisées RECAP-QUALITE ont été régulièrement organisées sur un échantillon d'environ 100 000 individus. Elles sont articulées en 2 phases : une phase de recodage de l'activité et de la profession pour l'échantillon en question, suivie d'une phase d'arbitrage par des experts entre ce 2ème code obtenu et le code obtenu par le premier codage qui avait concerné tous les bulletins (et qui est le code conservé *in fine* dans le RP). Dans cette dernière phase, il est également possible d'arbitrer en faveur d'un codage différent des deux précédents. Le codage issu de la phase d'arbitrage est considéré comme le « bon codage » et sert de référence pour apprécier la qualité du premier codage et mesurer un taux d' « erreur de codage ».

Grâce au dispositif décrit ci dessus, la qualité du codage peut être mesurée de façon assez détaillée : par processus (traitement automatique ou manuel), par niveau ou domaine de nomenclature, etc. L'objet de la communication est de présenter les principaux résultats obtenus sur la qualité du codage de la profession et de l'activité. Les opérations RECAP-QUALITE ayant été réalisées pour chaque enquête annuelle de recensement depuis 2006, à l'exception de 2010, l'évolution de la qualité de la codification sera également abordée lors de la présentation.

Dans un premier temps, nous avons constaté que la distribution de l'échantillon des 100 000 individus selon la codification initiale présente dans les fichiers du RP pour la profession (et pour l'activité) est très proche de celle obtenue selon la codification définie lors de la phase d'arbitrage. Cette comparaison nous a également permis d'isoler les postes de nomenclature de la profession et de l'activité qui sont les plus fragiles. Pour chaque opération RECAP-QUALITE, des taux d'« erreurs de codage » ont été calculés pour la profession et l'activité : en 2009, celui de la profession actuelle est de 6,2% au niveau le plus agrégé de la nomenclature (niveau groupe) et de 14,7% au niveau le plus fin (niveau PCS).

[1] yves.jacquin@insee.fr, Division des méthodes et traitements des recensements, Direction des statistiques démographiques et sociales, Insee

[2] elodie.martal@insee.fr. Au moment de la réalisation de l'étude, E. Martal était stagiaire au sein de la division des méthodes et traitements des recensements

Dans un second temps, nous avons regardé l'évolution de ces taux d'« erreur de codage » depuis 2006 : ils diminuent pour les variables étudiées. Ainsi, entre 2006 et 2009, le taux d'erreur du codage de l'activité a diminué de 6,4 points au niveau classe, passant de 18,2% à 11,8%, et de 8,7 points au niveau division, passant de 17,4% à 8,7%. Concernant la profession actuelle au niveau groupe, le taux d'erreur de codage passe de 8,9% en 2006 à 6,2% en 2009, soit une baisse de pratiquement 3 points. Plus le niveau de la nomenclature est détaillé, plus la baisse est forte entre 2006 et 2009 : moins 7,6 points pour le niveau PCS.