Le kit de diffusion ESANE ou comment diffuser des agrégats à des niveaux fins tout en alertant l'utilisateur sur la pertinence des chiffres publiés

Julien SENG[1]

En statistique d'entreprises, la majorité des résultats est diffusée par secteur d'entreprises. Ainsi, on agrège les caractéristiques des entreprises selon leur secteur d'activité. Ce dernier est calculé à partir de l'activité principale exercée (APE) par l'entreprise. L'engagement pris par le programme REfonte des Statistiques ANnuelles d'Entreprises (RESANE) en termes de diffusion des résultats est une publication des agrégats sectoriels au niveau sous-classe de la nomenclature d'activité (5 positions) ou au niveau groupe de la nomenclature (3 positions) croisé avec des tranches d'effectifs (4 tranches).

La méthodologie utilisée pour le calcul des agrégats est complexe. Il s'agit de la mise en œuvre d'un estimateur dit « composite » qui combine les données de l'échantillon de l'enquête avec des données exhaustives provenant des liasses fiscales. Il donne des résultats différents de ceux obtenus par simple sommation des valeurs d'un fichier de données individuelles.
A partir de ce constat, et sachant que bon nombre d'utilisateurs avait l'intention de diffuser des résultats à des niveau plus fins que ceux proposés en standard, il a été décidé de leur mettre à disposition un kit de diffusion. Ce dernier, développé en SAS, met en œuvre la méthodologie de calcul ESANE (Elaboration des Statistiques Annuelles d'Entreprises) et permet donc à l'utilisateur de produire des estimations plus fines que les résultats officiels diffusés, et bien sûr « compatibles » avec celles-ci, au sens où elles permettent de retrouver les marges publiées.

Pour mettre en œuvre ce kit, l'utilisateur n'a qu'à :

- définir la population sur laquelle il souhaite travailler
- définir le domaine de diffusion qu'il souhaite. Il pourra le constituer en croisant des variables catégorielles présentes dans le fichier de diffusion en standard (tranche de taille par exemple) ou créer ses propres variables de stratification
-
- choisir le type de statistique souhaité (agrégat, moyenne, distribution etc.)
- choisir la ou les variables d'intérêt
- choisir la ou les sorties (SAS, Excel ou HTML)

Toutefois, comme ces statistiques sont calculées à partir d'un échantillon, il est nécessaire que l'utilisateur puisse juger de la fiabilité du résultat fourni par le kit au niveau du domaine de diffusion défini. Vu la taille de l'échantillon il est en effet possible, si l'utilisateur choisit un domaine de diffusion trop fin, que l'agrégat obtenu n'ait aucun sens statistique car calculé à partir de la réponse d'une seule entreprise par exemple.
Ainsi, le kit fournira, en plus de la statistique calculée, des indicateurs de qualité qui permettront de juger de sa pertinence.
Ces indicateurs peuvent être classés en trois groupes :

---

[1] Insee : Département « Répertoires, Infrastructure et statistiques structurelles »

Email : julien.seng@insee.fr

- des indicateurs statistiques traditionnels (coefficient de variation, intervalle de confiance etc.)
- des indicateurs plus basiques (nombre d'unités répondantes ou extrapolées prises en comptes, etc.)
- un indicateur synthétique composé de trois modalités qui résume les informations précédentes :

    - statistique diffusable
    - statistique fragile
    - statistique non diffusable

Le pari du kit a donc été de donner la valeur de l'agrégat à partir du moment où ce dernier est calculable en faisant confiance à l'utilisateur quant à sa diffusion.

Enfin le kit permettra également de gérer le secret primaire. Là encore, confiance est faite à l'utilisateur pour qu'il ne diffuse que ce qui peut légalement l'être