Journées de méthodologie statistique de l'Insee 2022

LE REDRESSEMENT D'UNE ENQUÊTE ENTREPRISES AUPRÈS D'UNE POPULATION ATYPIQUE : LE CAS DE L'ENQUÊTE R&D AUPRÈS DES ENTREPRISES

Thomas BALCONE (*), Charles DEULIN (*), Lisa KERBOUL (*)

(*) SIES, Département des études statistiques de la recherche

thomas.balcone@recherche.gouv.fr
charles.deulin@recherche.gouv.fr
lisa.kerboul@recherche.gouv.fr

Mots-clés : non-réponse, imputation, repondération, calage, groupe de réponse homogène (GRH)

Domaine concerné : Statistique d'entreprises ; Redressement, pondération et repondération, calage sur marges

Résumé

La loi de programmation de la recherche (LPR) présentée le 19/03/2020 a pour objectif de porter l'investissement dans la recherche et le développement expérimental (R&D) à 3% du PIB. Cet objectif s'inscrit dans la continuité de la stratégie Europe 2020 qui portait déjà cet objectif de 3 %. Or la R&D représente 2,2 % du PIB et 462 000 emplois en équivalent temps plein (ETP) en France en 2019. Les enjeux autour de la recherche n'ont pas échappé à l'Etat et celui-ci a décidé de soutenir l'innovation (par exemple via le crédit impôt recherche ou plus récemment avec la LPR). Pour que cette politique soit efficace, il faut s'appuyer sur des statistiques fiables concernant entre autres les moyens consacrés à la R&D par les entreprises.

C'est la sous-direction des systèmes d'information et des études statistiques (SIES) qui est chargée de produire ces statistiques. Ces statistiques sont réalisées à partir de l'enquête annuelle sur les moyens consacrés à la R&D dans les entreprises implantées en France. Afin de produire les statistiques les plus fiables possibles à partir des données collectées dans le cadre de cette enquête, de nombreux traitements post-collecte ont été mis en place. Parmi ces traitements figurent ceux permettant de corriger la non-réponse des entreprises. Il est légitime, dans un souci de potentiellement mieux faire, de se poser la question suivante :

Quelles améliorations apporter aux traitements post-collecte actuellement mis en œuvre ?

Dans cet article, nous allons tout d'abord faire un état des lieux des traitements actuels visant à corriger la non réponse. Pour réaliser cet état des lieux, il est nécessaire de prendre

du recul sur les données. Nous proposons donc dans un premier temps de s'intéresser aux entreprises étudiées. La population d'intérêt de cette enquête est constituée des sociétés ayant réalisé effectivement des travaux de R&D en interne au cours de l'année. Cependant, afin de minimiser le risque de défaut de couverture, la base de sondage est constituée des entreprises susceptibles de réaliser de la R&D en interne. Ce recul sur les données permet également de présenter le plan de sondage ainsi que l'échantillonnage.

Ceci fait, on peut décrire les traitements post-collecte actuels. Ils s'articulent autour de 5 grandes phases :

- une première phase durant laquelle les bases de production sont rassemblées en une seule base de travail,
- une phase de correction de la non-réponse totale. Elle s'effectue par repondération,
- une phase de changement de clef primaire,
- une phase de correction de la non-réponse partielle,
- une phase de mise en forme de la base pour la diffusion.

Plusieurs problèmes potentiels ressortent de cet état des lieux des traitements post-collecte actuels. Nous avons tenté de les résoudre en proposant nos propres traitements. Ainsi, les deux dernières parties de cet article décrivent les traitements que nous proposons. Ce sont :

- les traitements préalables à la correction de la non-réponse. Ils diffèrent des traitements actuels en présentant une autre approche de la construction de la base de travail,
- les traitements destinés à la correction de la non-réponse partielle. L'ensemble des imputations existantes a été repris. Nous avons adapté certains traitements et proposé de nouveaux types de correction comme l'imputation par la régression,
- les traitements destinés à corriger la non-réponse totale. Nous avons gardé la stratégie de repondération en l'affinant en mettant en place des groupes de réponse homogène (GRH),
- le traitement des valeurs influentes. Celles-ci étaient traitées arbitrairement. Nous avons utilisé la technique de winsorisation de type II avec un calcul des seuils à l'aide de la méthode présentée par Kokic et Bell,
- le calage sur marges. Ce traitement n'était pas présent dans les traitements post-collecteactuels.

Bibliographie

[1] Béatrice NEITER et Benoît BUISSON. Comment redresser une enquête thématique ? *Série des documents de travail de la Direction des Statistiques d'Entreprises*, pages 8–10, Janvier 2010.

[2] Thomas DEROYON. La correction de la non-réponse par repondération. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[3] Thomas DEROYON et Cyril FAVRE-MARTINOZ. La correction de la non-réponse par imputation. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[4] Olivier SAUTORY. Les enjeux méthodologiques liés À l'usage de bases de sondage imparfaites. *Journées de méthodologie statistique*, pages 4–5, mars 2015.

[5] Nathalie CARON et Pascale PIETRI-BESSY Philippe BRION. Redresser la non-réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter. *Journées de méthodologie statistique*, pages 3–7, mars 2005.

[6] Cyril FAVRE-MARTINOZ et Thomas DEROYON. Traitement des valeurs influentes dans les enquêtes. *Note méthodologique du département des méthodes statistiques de l'Insee*, pages 1–4, octobre 2017.

[7] Thomas DEROYON. Traitement des valeurs atypiques d'une enquête par winsorization – application aux enquêtes sectorielles annuelles. *Journées de méthodologie statistique*, pages 3–5, 2015.

[8] Arnaud FIZZALA. Adaptation of winsorization caused by weight share method. *Article ESANE*, octobre 2018.

[9] David HAZIZA et Jean-François Beaumont Cyril FAVRE-MARTINOZ. Une méthode de détermination du seuil pour la winsorisation avec application à l'estimation pour les domaines. *Techniques d'enquête Vol. 41, Statistique Canada, n° 12-001-X*, pages 59–79, juin 2015.

[10] Olivier Sautory. La macro calmar redressement d'un échantillon par calage sur marge. *Série des documents de travail de la direction des statistiques démographiques et sociales*, novembre 1993.