Journées de méthodologie statistique de l'Insee 2022

Estimation des montants manquants de versements de TVA : exploitation des données du contrôle fiscal

Cécile WELTER-MÉDÉE (*), Simon QUANTIN (**)

(*) Insee, Département des Études Économiques, Division Marchés et Entreprises, CREST

(**) Insee, Département des Études Économiques, Division Marchés et Entreprises

simon.quantin@insee.fr
cecile.welter-medee@insee.fr

Mots-clés. : *Machine Learning*, biais de sélection, écart de TVA, contrôles fiscaux.

Domaines. Théorie des sondages aval, *Machine learning* / Apprentissage statistique

Résumé

L'estimation du manque à gagner de l'administration fiscale dans son activité de recouvrement des impôts est un enjeu important, mais est difficile à effectuer. L'extrapolation à l'ensemble des entreprises redevables d'un impôt en s'appuyant sur des informations issues des contrôles fiscaux, nécessite de prendre en compte le processus qui a conduit à sélectionner des entreprises contrôlées. Or ce processus est particulièrement complexe et non formalisé : il est engendré par un programme de contrôle qui concerne l'ensemble des secteurs d'activité, qui est lui-même issu d'un travail d'expertise fouillé conduit par les contrôleurs fiscaux sur la base de nombreuses informations. En découle un biais de sélection certainement important, qui empêche toute extrapolation hâtive.

Ce travail s'appuie sur les données de gestion du contrôle fiscal transmises par la DGFiP pour estimer les montants manquants de TVA. Nous adoptons une méthodologie en deux étapes (appliquée également par Tagliaferri dans le même contexte d'estimation de l'écart de TVA [10]), appliquée sur une partition des entreprises redevables de la TVA selon l'administration en charge du contrôle afin de prendre en compte la différence structurante d'organisation des contrôles fiscaux.

La première étape, qui s'inspire des méthodes de redressement de la non réponse par repondération, consiste à attribuer à l'ensemble des entreprises faisant une déclaration de TVA l'année considérée, une probabilité de contrôle afin de repondérer leurs poids de sondage pour mener l'extrapolation. Les probabilités de contrôle sont inconnues (comme les probabilités de réponse à une enquête le sont) et sont estimées à partir des informations contenues dans les déclarations de TVA adressées par les entreprises à la DGFiP. Une telle démarche est fréquente dans la correction du biais induit par le phénomène de non-réponse aux enquêtes, par exemple en prenant en compte la probabilité de réponse pour redresser les probabilités d'inclusion (Sarndal 2003 [9],

Deroyon 2018[11]), mais aussi dans les méthodes économétriques d'évaluation comme dans le cas de l'ajustement par pondération de l'inverse du score de propension (Austin 2015[1], Imbens 2003[5]). La correction de la non-réponse par repondération dans les enquêtes s'appuie sur une repondération des probabilités d'inclusion initiales par les probabilités de réponse estimées. Par analogie, on considère ici que les entreprises qui font une déclaration de TVA sont un échantillon (en l'occurrence exhaustif) dont une partie seulement a été contrôlée. On repondère alors chaque entreprise qui réalise une déclaration de TVA par une estimation de sa probabilité d'être contrôlée. Comme il est d'usage en redressement de la non-réponse dans les enquêtes menées auprès des entreprises, des probabilités de contrôle sont estimées à l'aide d'algorithmes de machine learning, de la famille des arbres de classification (les algorithmes sont entraînés sur des échantillons rééquilibrés à l'aide de la méthode SMOTE [2]). Ce type de méthode permet d'intégrer des effets non linéaires des variables explicatives et permet par ailleurs de prendre en compte des caractéristiques structurantes des entreprises, permettant en plus de tenir compte, en plus de leur propension à être contrôlée, de leur hétérogénéité (en termes de taille notamment). Par suite, et comme préconisé par Haziza et Beaumont (Haziza et Beaumont 2007[3]) des « groupes de contrôle homogènes » sont constitués, comme le seraient des groupes de réponse homogène dans le cas du redressement de la non-réponse par repondération, à partir des quantiles de la distribution des probabilités prédites. Les groupes d'entreprises ainsi constitués sont donc considérés comme homogènes au sens de la probabilité d'être contrôlée. Chaque entreprise se voit enfin attribuer la probabilité empirique d'être contrôlée observée dans son groupe de contrôle homogène. Cette dernière étape permet de se départir de la forme fonctionnelle sous-jacente de l'algorithme retenu pour la prédiction des probabilités de contrôle et de limiter l'apparition de points influents.

Dans une seconde étape, les montants de manque à gagner sont extrapolés par un estimateur par le ratio qui est appliqué à des domaines d'entreprises considérées cette fois-ci comme homogènes au sens de la fraude. La partition de la population d'entreprises faisant une déclaration de TVA en domaines permet de tenir compte de l'hétérogénéité des comportements de fraude, au sens des redressements notifiés semblables de par leur motif et/ou le montant éludé. Ce découpage en domaines permet d'extrapoler directement les montants de fraude en leur sein, en s'appuyant sur une hypothèse de comportement de fraude homogène des entreprises qui les constituent. Les déterminants de la fraude ont été identifiés à l'aide de l'estimation d'une régression logistique avec pénalisation LASSO sur la notification d'un montant strictement positif par l'administration fiscale, parmi les seules entreprises contrôlées. Un estimateur classique de la théorie des sondages est enfin utilisé : l'estimateur par le ratio, qui s'appuie sur de l'information auxiliaire corrélée aux montants d'impôts éludés.

Dans l'ensemble, les différentes estimations obtenues à partir des exercices comptables de 2012 semblent assez peu dépendantes des variations méthodologiques adoptées et sont globalement comprises entre 20 et 26 milliards d'euros.

Références

[1] Austin, P., and Stuart, E. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine 34* (08 2015).

[2] Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16* (June 2002), 321–357.

[3] Haziza, D., and Beaumont, J.-F. On the Construction of Imputation Classes in Survey. *International Statistical Review 75*, 1 (2007), 25–43.

[4] Haziza, D., Chen, S., and Gao, Y. Targeting Key Survey Variables at the Unit Nonresponse Treatment Stage. *Journal of Survey Statistics and Methodology* (11 2020).

[5] Imbens, G., Hirano, K., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 4 (2003), 1161–1189.

[6] Louvot, C. L'évaluation de l'activité dissimulée des entreprises sur la base des contrôles fiscaux et son insertion dans les comptes nationaux. *Documents de travail, Insee*, G2011/09 (2011).

[7] Sautory, Olivier. Les méthodes de calage, Note méthodologique INSEE, 2018.

[8] Särndal, C.-E., and Lundstrom, S. *Estimation in surveys with nonresponse.* John Wiley & Sons, 2005.

[9] Särndal, C.-E., Swensson, B., and Wretman, J. *Model assisted survey sampling.* Springer Science and Business Media, 2003.

[10] Tagliaferri, G., Scacciatelli, D., and Alaimo Di Loro, P. VAT tax gap prediction : a 2-steps Gradient Boosting approach, 12 2019.

[11] Thomas Deroyon. La correction de la non-réponse par repondération, Note méthodologique INSEE, 2018.