Journées de méthodologie statistique de l'Insee 2022

Imputation de valeurs manquantes avec des réseaux de neurones : prédiction des salaires dans l'enquête Emploi

Damien BABET (*)

(*) Insee, Direction des études et synthèses économiques

damien.babet@insee.fr

Mots-clés. : enquête Emploi, salaire, non-réponse, imputation, réseau de neurone.

Domaines. Non-réponse, Machine learning.

Résumé

Cet article reprend et développe un travail déjà présenté à l'Unece (Babet 2020) et dans un document de travail en cours d'achèvement (Babet, Deltour, Faria, Himpens).

La variable de salaire de l'enquête Emploi souffre d'une non-réponse importante. Environ 5% des salariés refusent de répondre, et 15% n'acceptent de répondre que par tranche. Ces observations font aujourd'hui l'objet d'une imputation par tirage aléatoire dans une distribution paramétrée par une équation de salaire de type "Mincer" (Mincer, 1974), et bornée par les limites de la tranche lorsqu'elles sont connues. Cela permet de livrer aux utilisateurs de l'enquête un fichier complet et plus facile d'usage, mais aussi de réduire ou corriger le biais de non-réponse, dans la mesure où cette non-réponse est ignorable (Rubin, 1976) et où le modèle d'imputation est bien spécifié (Chen et Haziza, 2017, 2019).

Nous expérimentons une autre méthode d'imputation : la prédiction par réseau de neurone (RN). Les méthodes d'apprentissage statistique telles que les réseaux de neurones sont particulièrement adaptées pour des usages purement prédictifs. Les problèmes d'imputation représentent donc un bon cas d'usage pour ces méthodes. Leur intérêt est de faciliter l'inclusion de plus de variables et l'estimation de modèles plus souples, en particulier non linéaires et comprenant des interactions, ce qui permet d'en espérer plusieurs avantages. D'une part, si le modèle classique est mal spécifié, l'imputation par RN est susceptible de réduire davantage le biais de non-réponse. D'autre part, on peut espérer un gain de précision de l'imputation. Enfin, lorsque les données, en partie imputées sont utilisées dans un second temps par des usagers du fichier statistique, un modèle d'imputation plus souple semble moins susceptible de générer des corrélations spécieuses.

Nous utilisons un corpus de 1,2 millions d'observations entre 1993 (date de la première question sur le salaire en clair) et 2018, divisé en un échantillon d'entraînement d'environ 1 million d'observation, un échantillon de validation et un échantillon de test d'environ 100000 observations chacun. La préparation des données est importante. La reconstitution de séries longues pour les variables explicatives n'a pas besoin d'être complète : les RN tolèrent bien en entrée des valeurs manquantes grossièrement imputées à la moyenne et signalée par une indicatrice. En revanche, les nomenclatures détaillées (profession, secteur, diplôme, etc.) sont difficiles à inclure

en raison de leur grand nombre de modalités. Nous adaptons des algorithmes de vectorisation de vocabulaire, issus de l'analyse du langage naturel (Mikolov et al. 2013), pour remplacer le vecteur des indicatrices de modalités par un encodage de plus faible dimension. Cette méthode, qui peut s'appliquer à d'autres types de variables catégorielles, présente un intérêt en elle-même (Doutreligne, Leduc, Nguyen, Vuagnat, 2020).

La qualité de l'ajustement est mesurée principalement par le R^2 sur l'échantillon test. Il est d'environ 0,6 pour une équation de salaire classique ou enrichie d'un grand nombre de variables, et de 0,7 pour la prédiction par RN. Outre ce surcroît de précision, d'autres enjeux du choix de la procédure d'imputation sont discutés, en particulier la stabilité des résultats, l'impact de l'imputation sur l'estimation de diverses quantités d'intérêt, et les éventuelles questions éthiques soulevées par le choix des variables explicatives.

Bibliographie

[1] Babet D. : « Wage Imputation with Deep Learning in the French Labor ForceSurvey », in UNECE workshop on statistical data editing, conference of european statisticians, 2020.

[2] Babet D., Deltour Q., Faria T., Himpens S., « Les réseaux de neurones, méthodes et cas d'usages pour la statistique publique », document de travail Insee, à paraître.

[3] Chen S., Haziza D., « Multiply robust imputation procedures for the treatment of item non-response in surveys », Biometrika, 104(2), 439–453, 2017.

[4] Chen S., Haziza D., « Recent developments in dealing with item non-response in surveys : a critical review », International Statistical Review, 87, S192–S218, 2019

[5] Doutreligne M., Leduc A., Nguyen D.-P., Vuagnat A., « Snds2vec, représentations continues pour les concepts médicaux du Système national des données de santé », Revue d'Épidémiologie et de Santé Publique, 68, S35, 2020.

[6] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J., « "Efficient estimation of word representations in vector space », arXiv preprint :1301.3781, 2013.

[7] Mincer J., « Schooling, Experience, and Earnings »Human Behavior & Social Institutions No. 2., 1974.

[8] Rubin D. B., « Inference and missing data », Biometrika, 63(3), 581–592, 1976.