Journées de méthodologie statistique de l'Insee 2022

Forêts aléatoires : d'une approche par modélisation assistée au traitement de la nonréponse

Mehdi Dagdoug (*), Camelia Goga (*) et David Haziza (**)

(*) Université de Bourgogne Franche-Comté,
Laboratoire de Mathématiques de Besançon, Besançon, FRANCE

(**) University of Ottawa, Department of Mathematics and Statistics,
Ottawa, CANADA

Mots-clés. Echantillonnage, forêts aléatoires, estimation par modélisation assistée, données manquantes, imputation.

Domaines. Théorie des sondage aval, contrôle et redressement des données.

1 Résumé

De nos jours, les enquêtes par sondage font face à l'émergence de jeux de données complexes et de très grandes tailles. Lorsque les mesures de la variable d'intérêt sont connues pour tout les éléments sélectionnés dans l'échantillon, il est pratique courante de recourir à la modélisation pour construire de meilleurs estimateurs (approche par modélisation assistée, voir Särndal et al. (1992)). Ces estimateurs sont particulièrement efficaces lorsque le modèle utilisé est capable de produire des prédictions proches des valeurs réelles, non observées.

Dans un cadre en grande dimension, il est fréquent que les modèles paramétriques deviennent particulièrement instables et inefficaces. Bien que les estimateurs assistées par le modèle restent approximativement sans biais dans ces cas là, ils peuvent tout de même souffrir d'une variance élevée. Ces dernières décennies, de nombreux algorithmes provenant du domaines de l'apprentissage statistique ont été proposés pour répondre à des problématiques de prédictions, notamment lorsque le nombre de covariables est important. L'algorithme de forêts aléatoires (Breiman, 2001) en est un exemple particulièrement prometteur. Dans ce travail, nous proposons une nouvelle classe d'estimateurs assistés par un modèle et basé sur des forêts aléatoires pour l'estimation du total d'une variable d'intérêt. Sous certaines conditions, l'estimateur proposé est asymptotiquement sans biais et convergent pour l'estimation d'un total. Un estimateur convergent de la variance est suggéré et la distribution asymptotique de l'estimateur est obtenue également permettant ainsi la construction des intervalles de confiance asymptotiques. Pour plus de détails, voir Dagdoug et al. (2021).

Lorsque certains éléments sélectionnés dans l'échantillon refusent de répondre, les statisticiens ont souvent recours à l'imputation : un procédé au cours duquel les valeurs manquantes sont remplacées par des valeurs prédites. Pour effectuer ces prédictions, un modèle d'imputation est généralement postulé. Ici encore, lorsque le nombre de covariable est important, les méthodes

paramétriques peuvent ne plus convenir. L'estimateur imputé qui en résulte peut alors souffrir d'un biais de nonréponse relativement important et d'une variance accrue. Nous proposons, ici aussi, d'utiliser les forêts aléatoires pour obtenir les valeurs imputés. Les propriétés en échantillon finie de l'estimateur imputé par forêt aléatoires sont étudiées. Sous quelques conditions, la convergence en moyenne quadratique de l'estimateur est obtenue.

Après une introduction aux forêts aléatoires dans un contexte d'échantillonage, nous ferons l'hypothèse des données complètes et présenterons les propriétés principales de l'estimateur par modélisation assistée. Nous considérerons ensuite un cadre de travail plus réaliste dans lequel certaines des mesures de la variable d'intérêts seront supposées manquantes. Les propriétés de l'estimateur imputé seront étudiées dans ce contexte. Nous présenterons les résultats de différentes études par simulations qui semblent illustrer le bon comportement des estimateurs proposés.

Références

Breiman, L. (2001). Random forests. *Machine learning*, 45 :5–32.

Dagdoug, M., Goga, C., and Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *To appear in Journal of the American Statistical Association*, pages 1–50.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.