Imputation par forêts aléatoires en théorie des sondages

Mehdi Dagdoug [1] & Camelia Goga [1] & David Haziza [2]

[1] *Université de Bourgogne Franche-Comté*
Laboratoire de Mathématiques de Besançon, Besançon, FRANCE
mohamed_mehdi.dagdoug@univ-fcomte.fr
camelia.goga@univ-fcomte.fr

[2] *University of Ottawa, Department of Mathematics and Statistics*
Ottawa, CANADA
dhaziza@uottawa.ca

Résumé. La nonréponse partielle est généralement traitée à l'aide de méthodes d'imputations. Celles-ci ont pour objectif de prédire les valeurs non observées de la variable d'intérêt à partir d'un ensemble de variables auxiliaires. Dans de nombreux cas, les forêts aléatoires ont été reconnues comme particulièrement efficaces en matière de prédictions. Nous nous intéressons ici aux propriétés des estimateurs imputés lorsque ceux-ci sont basés sur des forêts aléatoires. Nous considérons un algorithme proche des forêts aléatoires de Breiman (2001). Nous construisons un estimateur imputé par forêts aléatoires à partir de cet algorithme. Ses propriétés en échantillon fini sont étudiées et nous montrons que cet estimateur est L^2-consistant pour le total de la variable d'intérêt. Les simulations effectuées suggèrent que l'estimateur imputé peut, dans certains cas (grande dimension, relations non-linéaires, ...) devenir plus performant que les méthodes d'imputations habituellement utilisées.
Mots-clés. Théorie des sondages, données manquantes, imputation, forêts aléatoires.

Abstract. Item nonresponse in surveys is usually handled through some form of imputation. Random forests provide flexible tools for obtaining a set of imputed values. We consider a random forest algorithm close to that of Breiman (2001) and establish its finite sample properties as well as its L^2-consistency. We present the results from a simulation study that investigates the performance of point and variance estimators based on random forest imputation in terms of bias, efficiency and coverage rate.
Keywords. Survey sampling, missing data, nonresponse, imputation, random forests.

1 Introduction

Dans la grande majorité des enquêtes, le statisticien doit faire face au problème de la non-réponse. Dans cette présentation, nous nous focalisons sur la non-réponse partielle qui est

caractérisée par l'absence d'une partie des variables collectées par l'enquête. Ce type de non-réponse est, la plupart du temps, traitée par imputation simple. L'imputation simple consiste à remplacer une valeur manquante par une valeur de remplacement construite au moyen d'une information auxiliaire disponible pour les répondants et les non-répondants. Il existe un éventail de méthodes d'imputation, allant de méthodes relativement simples (par exemple, l'imputation par la régression linéaire) à des méthodes plus complexes (méthodes basées sur des algorithmes d'apprentissage machine). Dans ces travaux, nous nous intéressons aux propriétés des estimateurs imputés lorsque les valeurs imputées ont été construites au moyen de forêts aléatoires. En section 2, nous définissons le cadre de travail ainsi que les notations utilisées dans cet article. En section 3, nous décrivons un algorithme proche des forêts aléatoires initialement proposées par (Breiman, 2001). La section 4 donne les propriétés de l'estimateur imputé construit sur l'algorithme décrit en section 3.

2 Cadre de travail et notations

Considérons une population $U = \{1, 2, ..., N\}$ de taille N. Le but est d'estimer le total sur la population $t_{\mathrm{y}} = \sum_{k \in U} y_k$, d'une variable d'intérêt Y. Pour ce faire, nous sélectionnons un échantillon S de taille n selon un plan de sondage $\mathcal{P}(S)$. Nous désignons les probabilités d'inclusions du premier et second ordre par $\{\pi_k\}_{k \in U}$ et $\{\pi_{k\ell}\}_{k \neq \ell \in U}$, respectivement. L'échantillon est caractérisé par un vecteur d'indicatrices $\boldsymbol{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top$, où $I_k = 1$ si $k \in S$ et $I_k = 0$, sinon. En l'absence de valeurs manquantes, l'estimateur d'Horvitz-Thompson

$$\widehat{t}_{\mathrm{ht}} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{1}$$

permet d'estimer t_{y} sans biais pourvu que $\pi_k > 0$, pour tout $k \in U$.

En pratique, la variable d'intérêt Y est sujette à une nonréponse. Soit $\boldsymbol{r} = (r_1, \ldots, r_N)^\top$ le vecteur des indicatrices de réponse, avec $r_k = 1$ si y_k est observée, et $r_k = 0$, sinon. Soit $S_r = \{k \in S;\ r_k = 1\}$, l'ensemble des répondants, de taille n_r, et $S_{nr} = \{k \in S;\ r_k = 0\}$ l'ensemble des nonrépondants, de taille n_{nr},. On a $S_r \cup S_{nr} = S$ et $n_r + n_{nr} = n$. Soit $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})^\top$, le vecteur des variables auxiliaires pour l'individu k. Les données observées sont donc $D = \{(y_k, \mathbf{x}_k), k \in S_r\}$ ainsi que $\mathbf{x}_k$ pour $k \in S_{nr}$.

Nous supposons que la relation qui lie la variable d'intérêt aux variables auxiliaires peut être décrite au moyen du modèle suivant:

$$\begin{aligned} \mathbb{E}(y_k \mid \mathbf{x}_k) &= m(\mathbf{x}_k), \\ \mathbb{V}(y_k \mid \mathbf{x}_k) &= \sigma^2, \\ Cov(y_k, y_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell) &= 0, \quad k \neq \ell, \end{aligned}$$

où $m(\cdot)$ est une fonction supposée inconnue et σ^2 un paramètre inconnu.

Une fois l'étape d'imputation réalisée, l'estimateur imputé de t_{y} est donné par

$$\widehat{t}_{\widehat{\mathrm{m}}} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_{nr}} \frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}, \tag{2}$$

où $\widehat{m}(\mathbf{x}_k)$ désigne la valeur prédite pour le nonrépondant $k \in S_{nr}$.

3 Forêts aléatoires

D'une manière générale, une forêt aléatoire est une méthode d'ensemble basée sur une collection d'arbres de régression. Un arbre de régression est un algorithme qui, dans un premier temps, construit une partition de $\mathbb{R}^p$, puis, utilise cette partition pour faire un ensemble de valeurs prédites. Plus précisément, la prédiction d'un arbre au point $\mathbf{x}$ est donnée par la moyenne des observations y_k, pour les éléments k tels que $\mathbf{x}_k$ soit dans le même élément de la partition que $\mathbf{x}$. En considérant une partition $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_T\}$ de $\mathbb{R}^p$ construite à l'aide des données disponibles D avec un mécanisme aléatoire modélisé par une variable aléatoire Θ, la prédiction d'un arbre au point $\mathbf{x}$ est donnée par

$$\widehat{m}_{\mathrm{tree}}(\mathbf{x}, \Theta) = \sum_{k \in S_r(\Theta)} \frac{\mathbb{1}_{A(\mathbf{x},\Theta)}(\mathbf{x}_k)}{\sum_{\ell \in S_r(\Theta)} \mathbb{1}_{A(\mathbf{x},\Theta)}(\mathbf{x}_\ell)} y_k = \sum_{k \in S_r} W_k^{(tree)}(\mathbf{x}) y_k,$$

où

$$W_k^{(tree)}(\mathbf{x}) = \frac{\psi_k \mathbb{1}_{A(\mathbf{x},\Theta)}(\mathbf{x}_k)}{\sum_{\ell \in S_r} \psi_\ell \mathbb{1}_{A(\mathbf{x},\Theta)}(\mathbf{x}_\ell)}$$

et $\psi_k = 1$ si l'élément k fait partie de l'échantillon considéré par l'arbre en question. Dans le cas d'un arbre de régression non randomisé (donc sans méchanisme de rééchantillonage), $\psi_k = 1$ pour tout les éléments de l'échantillon observé; en revanche, dans d'autre cas, voir notamment l'exemple des forêts aléatoires décrit ci-dessous, certains des éléments peuvent ne pas être considérés pour la construction de l'arbre.

Une prédiction obtenue au moyen d'une forêt aléatoire de taille B est définie comme la moyenne des prédictions issues de chacun des B arbres:

$$\widehat{m}_{rf}(\mathbf{x}, D, \Theta_1, ..., \Theta_B) = \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}, D, \Theta_b) = \sum_{k \in S_r} W_k^{(rf)}(\mathbf{x}) y_k, \tag{3}$$

où $W_k^{(rf)}(\mathbf{x}) = B^{-1} \sum_{b=1}^{B} W_k^{(tree,b)}$ et $\Theta_1, \Theta_2, \ldots, \Theta_B$ sont des variables aléatoires indépendantes et identiquement distribuées, indépendantes de l'ensemble des quantités aléatoires précédemment

mentionnées. Plus précisément, dans le cas des forêts aléatoires proposées par Breiman, la première étape de l'algorithme consiste en la sélection de B échantillons bootstrap, dénotés $S_r(\Theta_1), S_r(\Theta_2), \ldots, S_r(\Theta_B)$. Chaque arbre est construit à partir des données fournies par son échantillon bootstrap, de telle manière que la partition des arbres de régression est créé par différents splits successifs, où, chaque split conduit à la création de régions les plus homogènes possible. Désignons par A la node de cardinalité $\#(A)$ considérée pour le prochain split et $\mathcal{C}_A$ l'ensemble des splits possibles dans la node A. La procédure de split est exécutée en cherchant le split optimal (j^*, z^*) pour lequel le critère suivant est maximisé:

$$L(j,z) = \frac{1}{\#(A)} \sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A} \left\{ (y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L} \mathbb{1}_{x_{kj} < z} - \bar{y}_{A_R} \mathbb{1}_{x_{kj} \geq z} \right)^2 \right\},$$

où $A_L = \{k \in A; x_{jk} < z\}$, $A_R = \{k \in A; x_{jk} \geqslant z\}$ et $\bar{y}_A$ est la moyenne des observations y_k appartenant à la région A. Cette procédure continue jusqu'à ce qu'un split supplémentaire conduise à une feuille avec moins d'éléments qu'un nombre prédéterminé n_0. Ce critère est optimisé, non pas sur l'ensemble des p variables auxiliaires, mais sur un sous-ensemble de cardinalité p_0 de ces p variables auxiliaires, choisies aléatoirement selon Θ, avec p_0 choisi par l'utilisateur.

4 Propriétés de l'estimateur imputé par forêt aléatoire

L'estimateur imputé par forêt aléatoire est défini par

$$\widehat{t}_{\mathrm{rf}} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_{nr}} \frac{\widehat{m}_{\mathrm{rf}}(\mathbf{x}_k)}{\pi_k}, \tag{4}$$

où $\widehat{m}_{\mathrm{rf}} := \widehat{m}_{rf}(\mathbf{x}, D, \Theta_1, \ldots, \Theta_B)$ tel que défini en (3). Il est possible d'écrire $\widehat{t}_{\mathrm{rf}}$ comme une somme pondérée des observations y_k pour $k \in S_r$ comme nous l'établirons dans la proposition suivante.

Proposition 4.1. *L'estimateur imputé (4) peut s'écrire comme*

$$\widehat{t}_{\mathrm{rf}} = \sum_{k \in S_r} w_{ks} y_k,$$

où les poids w_{ks} sont donnés par

$$w_{ks} = \frac{1}{\pi_k} + \sum_{\ell \in S_{nr}} \frac{W_k^{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{5}$$

Ces poids caractérisent donc l'estimateur $\widehat{t}_{\mathrm{rf}}$. Ils peuvent être bornés de la manière suivante.

Proposition 4.2. *Les poids w_{ks} en (5) peuvent être bornés de la manière suivante:*

$$\frac{1}{\pi_k} \leqslant w_{ks} \leqslant \frac{1}{\pi_k} + \frac{1}{n_0} \sum_{\ell \in S_{nr}} \frac{1}{\pi_\ell}, \qquad k \in S_r.$$

Ces bornes peuvent toutes les deux être atteintes. En effet, la borne inférieure se réalise pour les répondants n'ayant été sélectionné dans aucun échantillon bootstrap ou pour ceux appartenant à une feuille ne contenant aucun nonrépondant (dans chacun des arbres); la borne supérieure se réalise pour les répondants tels que, tous les nonrépondants se trouvent dans la même feuille que le répondant en question, de taille exactement n_0, et ce pour chacun des B arbres. Bien que potentiellement réalisable, il est possible de montrer que la probabilité que ces bornes soient atteintes tend vers 0 lorsque le nombre d'arbre tend vers l'infini.

Afin d'établir les propriétés asymptotiques de l'estimateur $\widehat{t}_{\text{rf}}$, nous utilisons le cadre asymptotique de Isaki and Fuller (1982). Considérons pour cela une suite emboitée infinie de populations $\{U_v\}_v$ de tailles $N_v \to \infty$ et une suite d'échantillons $S_v \subset U_v$ de taille $n_v \to \infty$. Nous supposons que l'estimateur d'Horvitz-Thompson en (1) est L^2 consistant. Nous ferons également l'hypothèse que les données sont Missing At Random (MAR):

$$\mathbb{P}\left(r_k = 1 | \mathbf{x}_k, y_k\right) = \mathbb{P}\left(r_k = 1 | \mathbf{x}_k\right), \qquad k \in U.$$

Nous formulons de plus certaines hypothèses sur le modèle, analogues à celles formulées par Scornet et al. (2015).

Result 4.1. *L'estimateur $\widehat{t}_{rf}$ est L^2 consistant*

$$\lim_{v \to \infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{\text{rf}} - t_{\text{y}}\right)\right)^2\right] = 0.$$

Nous présenterons de plus les résultats d'une étude par simulation visant à évaluer le compartement de l'estimateur $\widehat{t}_{rf}$ en termes de biais et de variance. Ces simulations suggèrent que l'estimateur proposé est particulièrement intéressant lorsque les relations entre variable d'intérêt et variables auxiliaires sont non-linéaires, ainsi que lorsque le nombre de variables auxiliaires est important.

Bibliographie

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

Dagdoug, M., Goga, C., and Haziza, D. (2020). Model-assisted estimation through random forests in finite population sampling. *arXiv preprint arXiv:2002.09736*.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:49–61.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.