Journées de méthodologie statistique de l'Insee 2022

Arbres et 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. Échantillonnage, forêts aléatoires, estimation par modélisation assistée, données manquantes, imputation.

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

1 Introduction

Depuis de nombreuses années, les modèles prédictifs jouent un rôle prépondérant dans la théorie des sondages. En effet, deux paradigmes peuvent être observés en échantillonnage : l'approche "sous le modèle" et l'approche "sous le plan", voir notamment Brewer and Gregoire (2009) pour un article de revue récent sur ces concepts. Comme son nom l'indique, les modèles prédictifs sont au cœur de la théorie lorsque l'inférence est réalisée sous le modèle. En réalité, même lorsque l'approche choisie est l'approche sous le plan, il est fréquent d'avoir recours aux modèles prédictifs ; l'approche est dans ce cas communément qualifiée d'être sous le plan, assistée d'un modèle, voir notamment les travaux fondateurs de Cassel et al. (1976), Särndal (1980), Robinson and Särndal (1983) et popularisée par Särndal et al. (1992). Dans ce cas de figure, bien que les stratégies utilisées restent évaluées vis-à-vis de leurs propriétés sous le plan, les modèles prédictifs sont utilisés pour construire des estimateurs pouvant s'avérer plus efficaces que les estimateurs usuels. Enfin, quelque soit le paradigme adopté pour l'inférence, les modèles prédictifs interviennent aussi dans le contexte de la non-réponse ; contexte dans lequel certains des éléments sélectionnés dans l'échantillon refusent de divulguer certaine(s) information(s) demandée(s) par le statisticien d'enquête. Dans ce cas de figure, si certaines variables auxiliaires ont été complètement observées, il est pratique courante d'avoir recours à l'imputation, un procédé consistant à remplacer dans les estimateurs usuels les valeurs manquantes par des valeurs prédites au moyen d'un modèle prédictif spécifié par le statisticien. Le lecteur intéressé peut se référer à Haziza (2009) ou Chen and Haziza (2019) pour une revue de la littérature sur les données manquantes en sondage.

Naturellement, les modèles prédictifs sont utilisés dans de nombreux domaines et leur étude est par conséquent devenue un domaine propre très en vue des statistiques : l'apprentissage statistique. L'attractivité croissante de ce domaine a donné lieu à la découverte ainsi qu'à l'étude de

nombreux modèles prédictifs. Certains d'entre eux s'avèrent être particulièrement performants dans un grand nombre de scénarios. En particulier, certains sont reconnus pour demeurer efficaces lorsque le nombre de covariables disponible est important ; c'est notamment le cas des forêts aléatoires Breiman (2001). Ces dernières sont des modèles prédictifs dits d'ensemble, c'est-à-dire, basés sur un grand nombre de modèles sous-jacents. En l'occurrence, dans le cas des forêts aléatoires, les modèles sous-jacents sont les arbres de régression. Les performances empiriques encourageantes des forêts aléatoires ont entraîné leur utilisation dans une grande variété d'applications, allant de la médecine (Fraiwan et al., 2012), aux séries temporelles (Kane et al., 2014), en passant par l'agriculture (Grimm et al., 2008), les données manquantes (Stekhoven and Buhlmann, 2011), l'analyse génomique (Qi, 2012), pour n'en citer que quelques-uns. Ce sont des algorithmes toutefois relativement complexes ; par conséquent, leur analyse mathématique est encore assez récente et partiellement incomplète. En effet, bien souvent, les modèles étudiés dans la littérature sont des modèles simplifiés qui sont quelques fois relativement éloignés de ceux utilisés en pratique. Toutefois, certains articles récents (e.g. Scornet et al. (2015), Klusowski (2021)) se sont intéressés à des algorithmes semblables à ceux utilisés en pratique et ont permis d'établir la consistance faible des forêts aléatoires vers la fonction de régression. Le lecteur intéressé par une revue récente de ces résultats peut notamment consulter Biau and Scornet (2016).

En théorie des sondages, l'étude des arbres de régression a été initiée par Toth and Eltinge (2011), un article dans lequel les auteurs exhibent des conditions sous lesquels un algorithme particulier d'arbre de régression (voir Exemple 3.2) entraîné dans un échantillon provenant d'un sondage est faiblement consistant pour la fonction de régression. Quelques années plus tard, McConville and Toth ont étudié les propriétés asymptotiques sous le plan d'un estimateur par modélisation assistée basé sur l'algorithme précédemment introduit. Enfin, plus récemment, Dagdoug et al. (2021) ont généralisé les résultats de McConville and Toth (2019) en étudiant une classe d'estimateurs par modélisation assistée basée sur une classe de forêts aléatoires. Dans le contexte de sondage en présence de valeurs manquantes, Dagdoug et al. (2022) étudient les propriétés des estimateurs imputés lorsque ceux-ci sont construits à partir d'arbres de régression ou de forêts aléatoires.

Cet article propose une revue de la théorie développée concernant les arbres de régression et les forêts aléatoires en théorie des sondages. La Section 2 définit les notations, les motivations et le cadre de travail de cet article. En Section 3, nous nous plaçons dans une population finie, constituée d'observations indépendantes et identiquement distribuées (i.i.d.) et définissons les concepts d'arbres de régression et de forêts aléatoires. En Section 4, nous faisons l'hypothèse que la variable d'intérêt est totalement observée (donc, sans non-réponse) et nous rappelons les propriétés des estimateurs par modélisation assistée construits à partir d'arbres ou de forêts. La Section 5 s'intéresse quant à elle aux propriétés de l'estimateur imputé basé sur ces modèles. Enfin, nous concluons cet article avec quelques remarques finales en Section 6.

2 Contexte et notations

Soit $U := \{1, 2, ..., N\}$ une population finie de taille N, avec $N \in \mathbb{N}^*$ fixé, connu, et Y une variable d'intérêt. Dans l'intégralité de cet article, nous nous intéressons à l'estimation du total de la variable d'intérêt Y au niveau de la population U défini par

$$t_y := \sum_{k \in U} y_k, \tag{1}$$

où y_k désigne la valeur de la variable d'intérêt Y pour l'élément k de la population. Pour clarifier l'exposition de nos arguments, nous ferons l'hypothèse que la variable Y est à valeurs bornées, c'est-à-dire inclue dans un intervalle $[C_{1,Y}; C_{2,Y}]$, avec $C_{1,Y} < C_{2,Y}$ deux réels. Pour estimer

t_y, un échantillon aléatoire $S \subset U$ de taille n est sélectionné à l'aide d'un plan de sondage p. Nous noterons $\{\pi_k\}_{k\in U}$ et $\{\pi_{k\ell}\}_{k\neq l\in U}$, les N-uplets des probabilités d'inclusion du premier et du second ordre, c'est-à-dire $\pi_k := \mathbb{P}\{k \in S\}$, $\pi_{k\ell} := \mathbb{P}\{k, \ell \in S\}$ pour $(k, \ell) \in U^2$. Dans la suite, toutes les probabilités d'inclusions seront supposées être strictement positives.

Sans information supplémentaire, il est courant d'utiliser l'estimateur d'Horvitz-Thompson (Horvitz and Thompson, 1952) pour estimer t_y :

$$\widehat{t}_\pi := \sum_{k\in S} \frac{y_k}{\pi_k}. \tag{2}$$

L'estimateur $\widehat{t}_\pi$ est un estimateur sans biais, au sens où

$$\mathbb{E}_p\left[\widehat{t}_\pi - t_y\right] = 0,$$

avec pour variance

$$\mathbb{V}_p\left(\widehat{t}_\pi\right) = \sum_{k\in U}\sum_{l\in U} \Delta_{k\ell}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}, \tag{3}$$

où $\Delta_{k\ell} := \pi_{k\ell} - \pi_k\pi_\ell$ et $\mathbb{V}_p$ désigne l'opérateur de variance vis-à-vis de la distribution induite par le plan de sondage. Toutefois, dans certains cas, certaines informations auxiliaires sont disponibles. À cet effet, à chacun des éléments de la population est aussi associé un vecteur de p covariables $X_1, X_2, ..., X_p$ supposées être dans l'hypercube unité $[0; 1]^p$. Nous noterons $\mathbf{x}_k := [x_{k1}, x_{k2}, ..., x_{kp}]^\top$ le vecteur de covariables pour l'élément k. Les covariables et la variable d'intérêt sont de plus supposées être liées par le modèle de superpopulation suivant :

$$\xi : \qquad y_k = m(\mathbf{x}_k) + \epsilon_k, \qquad k \in U, \tag{4}$$

où $m : [0; 1]^p \to [-C_{1,Y}; C_{2,Y}]$ est une fonction inconnue et $\{\epsilon_k\}_{k\in U}$ un N-uplet de bruit blanc i.i.d. supportés dans un compact de $\mathbb{R}$ tel que $\mathbb{E}[\epsilon_k|\mathbf{x}_k] := 0$ et $\mathbb{E}\left[\epsilon_k^2|\mathbf{x}_k\right] := \sigma^2$. Dans cet article, la fonction m sera appelée fonction de régression. Nous utiliserons la notation $[T] := \{1, 2, ..., T\}$ pour dénoter la liste des entiers naturels strictement positifs allant jusqu'à T.

3 Arbres de régression et forêts aléatoires

Les arbres et les forêts aléatoires sont des algorithmes ayant été conçus pour estimer la fonction de régression inconnue m du modèle (4) et faire des prédictions ; ce sont des méthodes non paramétriques de prédiction. Plus précisément, nous appellerons "méthode de prédiction" (au niveau de la population) toute fonction $\widetilde{m}(\cdot, D_U) := \widetilde{m}(\cdot)$ où $D_U := \{(\mathbf{x}_k, y_k)\}_{k\in U}$. Dans cette section, par souci d'uniformisation, nous faisons dans un premier temps l'hypothèse (irréaliste, en pratique) que D_U est totalement observé et définissons les arbres et les forêts sur ces données ; les extensions de ces définitions aux cadres de travail concrets que nous considèrerons par la suite seront mentionnés dans les sections correspondantes.

3.1 Arbres de régression

Un arbre de régression est une méthode de prédiction pouvant être vu comme un algorithme composé de deux éléments : un algorithme de partitionnement et une règle de prédiction. Notons D_N l'ensemble des N-uplets de vecteurs de $[0; 1]^p \times [-C_Y; C_Y]$.

Un *algorithme de partitionnement* est un algorithme qui, à données fixées, permet de définir une partition de l'espace des covariables, c'est-à-dire une fonction déterministe $P : D_N \longrightarrow \mathcal{P}([0; 1]^p)$ où $\mathcal{P}([0; 1]^p)$ désigne l'ensemble des partitions de l'hypercube unité de $\mathbb{R}^p$, voir Nobel (1996) pour davantage de détails. Généralement, les partitions sont créées par splits successifs

ayant pour objectif l'optimisation d'un certain critère. Les éléments de la partition seront appelés les feuilles de l'arbre.

Une *règle de prédiction* est un algorithme qui, à partition $\mathcal{P} := \{\widetilde{\mathcal{A}}_1, \widetilde{\mathcal{A}}_2, ..., \widetilde{\mathcal{A}}_T\}$ et données fixées D_U, retourne une prédiction. Dans le cas des arbres de régression, la règle de prédiction traditionnellement utilisée retourne la moyenne de l'ensemble $\{y_k;\ k \in U \text{ tel que } \mathbf{x}_k \in \widetilde{\mathcal{A}}(\mathbf{x})\}$, où $\widetilde{\mathcal{A}}(\mathbf{x})$ désigne la feuille de l'arbre contenant le point $\mathbf{x}$. Plus précisément, la prédiction $\widetilde{m}_{tree}(\cdot, P, D_U) := \widetilde{m}_{tree}(\cdot)$ construite à partir de l'arbre est définie par

$$\widetilde{m}_{tree}(\mathbf{x}) = \sum_{k \in U} \frac{\mathbb{1}_{\mathbf{x}_k \in \widetilde{A}(\mathbf{x})}}{\sum_{l \in U} \mathbb{1}_{\mathbf{x}_\ell \in \widetilde{A}(\mathbf{x})}} y_k, \tag{5}$$

où, $\mathbb{1}_{\mathbf{x}_k \in \widetilde{A}(\mathbf{x})} = 1$ si $\mathbf{x}_k \in \widetilde{A}(\mathbf{x})$, et 0 sinon. Dans cet article, sauf mention contraire, le terme arbre ou arbre de régression désignera un arbre de régression quelconque, construit à partir d'un algorithme de partitionnement quelconque. La Figure 1 illustre un exemple d'arbre de régression construit à l'aide de deux covariables, auquel est associé la partition de $\mathbb{R}^2$ correspondante.

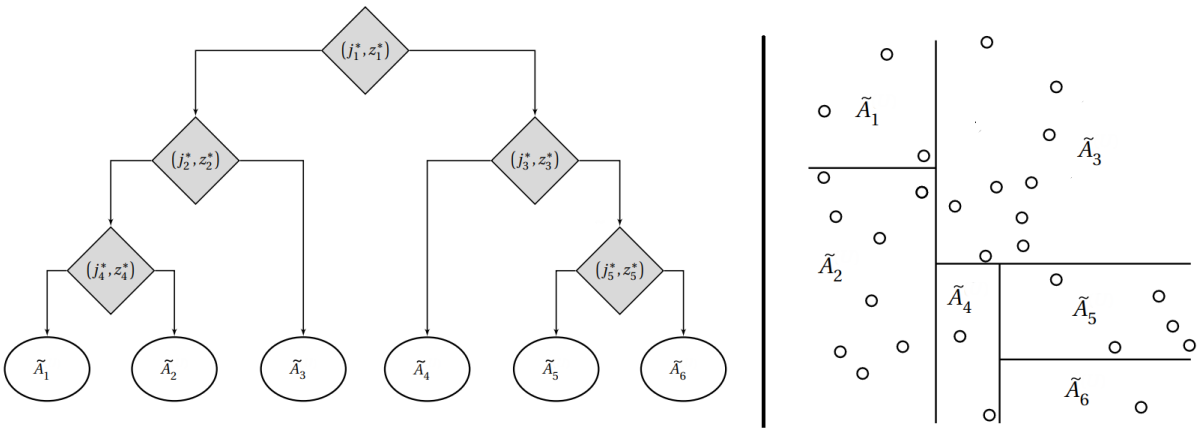

FIGURE 1 – Arbre de régression sur $\mathbb{R}^2$ (gauche) et sa partition correspondante (droite).

Ci-dessous, nous décrivons l'algorithme de partitionnement CART, couramment utilisé en pratique.

Exemple 3.1. *Algorithme de partitionnement CART (Breiman, 1984).*
Dans l'algorithme de partitionnement CART, la partition est obtenue par splits successifs. Plus précisément, soit A une feuille de cardinalité $\#(A)$ considérée pour le split suivant ; $\mathcal{C}_A$ l'ensemble de tous les splits possibles dans la feuille A, ce qui correspond à toutes les paires $(j, z) = (variable, position)$. Définissons

$$mse(A) := \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} (y_k - \bar{y}_A)^2 \qquad et \qquad \bar{y}_A := \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} y_k.$$

La procédure de split est réalisée en recherchant le meilleur split (j^*, z^*), c'est-à-dire celui maximisant le critère suivant

$$L(j, z) = mse(A) - mse(A_L) - mse(A_R) \tag{6}$$

où $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$. De manière équivalente, maximiser (6) revient aussi à minimiser

$$L(j, z) = mse(A_L) + mse(A_R).$$

Ce critère cherche donc le split qui permettra d'obtenir des feuilles filles les plus homogènes possibles, au sens de l'erreur quadratique moyenne. Les splits sont toujours réalisés au milieu de deux points. La procédure continue tant qu'un critère d'arrêt n'est pas atteint. Les critères d'arrêts habituels sont obtenus en spécifiant un nombre minimal d'éléments (n_0) dans les feuilles terminales, ou encore une profondeur maximale (K) pour l'arbre.

Exemple 3.2. *Règle de partitionnement proposé par McConville and Toth (2019).*
L'algorithme proposé par McConville and Toth peut être décrit par les étapes suivantes, réalisées dans chacune des feuilles existantes.

1. *Considérer $n_0 := n^{11/20}$, le nombre minimal d'unités dans chaque feuille terminale, et choisir un paramètre $\alpha \in]0; 0.5[$, un seuil de confiance.*
2. *Si la feuille choisie $\mathcal{A}$ contient moins de $2 \times n_0$ éléments, alors $\mathcal{A}$ est une feuille terminale. Dans ce cas, retourner à l'étape 1. pour la feuille suivante.*
3. *Parmi les p covariables disponibles, choisir celle qui a la statistique de test avec la p-valeur la plus faible dans le test d'hypothèse H_0 : "$\exists C \in \mathbb{R}$ tel que $\mathbb{E}[Y|X_j \in \mathcal{A}] = C$" pour $j = 1, ..., p$. Si aucune de ces statistiques de tests n'est significative (vis-à-vis du seuil α fixé à l'étape 1.), alors $\mathcal{A}$ est une feuille terminale. Dans ce cas, retourner à l'étape 1. pour la feuille suivante.*
4. *Effectuer le split à une position $z^* \in \arg\max_z L(j, z)$, avec L défini de la même manière que pour le critère CART. Ce critère n'est optimisé que sur les positions amenant à des feuilles filles contenants au moins n_0 éléments dans chaque feuille fille.*

3.2 Forêts aléatoires

En pratique, les arbres de régression sont particulièrement appréciés car ils ont l'avantage d'être simple à comprendre et à interpréter. Cependant, leurs capacités prédictives sont assez limitées. En effet, par construction, un arbre de régression appartient à l'ensemble des fonctions en escalier (c'est-à-dire, l'ensemble des fonctions constantes par morceaux). De par la règle de prédiction considérée, il suit que le nombre de morceaux est nécessairement inférieur ou égal à N ; on a donc $\widetilde{m}_{tree}$ qui appartient à l'ensemble des fonctions en escaliers de $\mathbb{R}^p$ dans $\mathbb{R}$ avec au maximum N morceaux. Si la fonction de régression m du modèle (4) appartient aussi à (ou est proche d'appartenir à) cet ensemble, alors il est possible que $\widetilde{m}_{tree}$ soit un bon estimateur de m. En revanche, si m est une fonction lisse, disons continue, alors $\widetilde{m}_{tree}$ risque d'être particulièrement éloigné de m. Toutefois, il est possible de montrer que toute fonction continue est limite uniforme d'une suite de fonctions en escaliers. La construction d'une telle suite repose sur deux éléments : 1) le nombre de morceaux doit tendre vers l'infini ; 2) le diamètre de chacun des morceaux doit tendre vers 0. Il est donc possible de considérer N comme étant un indicateur de la complexité maximale qu'un arbre peut atteindre. On observe donc que, asymptotiquement (pour N grand), il est possible qu'un arbre de régression soit un bon estimateur d'une fonction de régression continue ; pour de faibles tailles d'échantillons, en revanche, cette estimation peut être assez éloignée. Il est possible de montrer que si $\{f_b\}_{b\in[B]}$ est un B-uplet à valeurs dans l'ensemble des fonctions en escaliers ayant au maximum N morceaux, alors la fonction "moyenne"

$$f_{ave} := \frac{1}{B}\sum_{b\in B} f_b,$$

appartient à l'ensemble des fonctions en escaliers ayant au maximum $N \times B$ morceaux.

Une forêt aléatoire utilise cette idée pour estimer m par une moyenne de B fonctions en escaliers différentes. Cela implique que la complexité de f_{ave} peut être beaucoup plus importante que celle des f_b. Naturellement, pour que le gain de complexité soit effectif, il est nécessaire que les fonctions f_b soient différentes les unes des autres. Les forêts aléatoires reposent sur ce

principe ; une forêt aléatoire est un estimateur de m, construit sur une moyenne d'arbres de régression (et donc de fonctions en escaliers ; voir notamment la Figure 2). En observant que les règles de partitionnement décrites dans les Exemples 3.1 et 3.2 sont déterministes, il est clair que, à données fixées, utiliser le même algorithme plusieurs fois pour construire plusieurs arbres reviendrait simplement à construire plusieurs fois le même arbre. Pour remédier à ce problème, il a été proposé (Breiman, 1996, 2001) d'introduire une source aléatoire dans les algorithmes de partitionnement et/ou règles de prédictions. Nous continuons donc notre discussion sur la notion de forêt aléatoire en considérant le concept de *méthode de prédiction stochastique*. Ci-dessous, nous motivons l'utilisation de forêts aléatoires en illustrant les différences d'estimation de m qui peuvent intervenir entre un arbre de régression $\widetilde{m}_{tree}$ et une forêt aléatoire $\widetilde{m}_{rf}^{(B)}$; dans les deux cas, l'algorithme de partitionnement utilisé est l'algorithme CART décrit en Exemple 3.1, avec $n_0 = 10$. Nous avons généré 100 observations d'une covariable X_1 de loi uniforme $\mathcal{U}([0;1])$ et défini une variable d'intérêt $Y = m(X_1) + \mathcal{N}(0; 0.2)$, avec $m : x \mapsto 4 + 2x^2$.

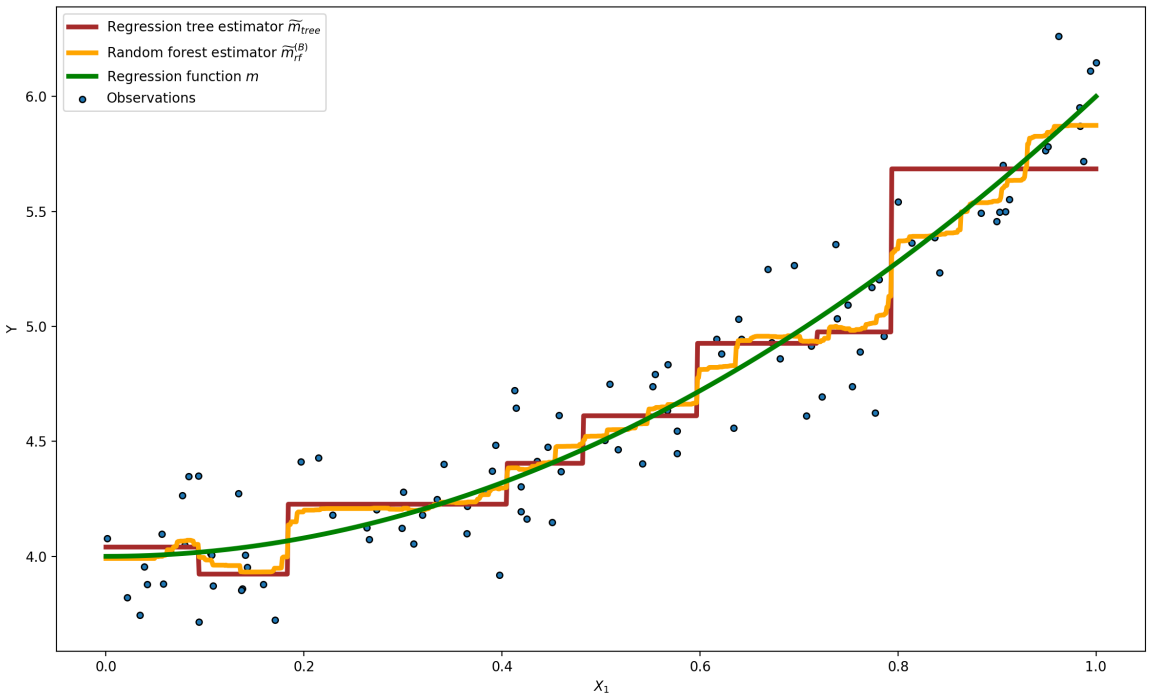

FIGURE 2 – Estimation d'une fonction de régression par arbre et forêt.

Soit Θ défini sur un espace mesurable $(E, \mathcal{E})$. Une méthode de prédiction stochastique $\widetilde{m}$ est une fonction mesurable $\widetilde{m}$ telle que $\widetilde{m} : \mathbb{R}^p \times E \to \mathbb{R}$. En d'autres termes, la méthode de prédiction $\widetilde{m}$ peut utiliser une variable aléatoire pour faire ses prédictions. Il s'ensuit que la méthode de prédiction $\widetilde{m}$ est aléatoire vis-à-vis de Θ, et par conséquent, une source de variation additionnelle est présente.

Exemple 3.3. *Soient $q \in]0;1[$ et Θ une variable aléatoire de loi $\mathcal{B}(q)$; définissons $\widetilde{m}(\mathbf{x}, \Theta) := \Theta||\mathbf{x}||_2$, où $||\cdot||_2$ dénote la norme euclidienne. Alors, $\widetilde{m}$ est une méthode de prédiction stochastique, au sens où, pour deux réalisations de Θ, $\widetilde{m}$ peut prendre des valeurs différentes. Une source de variation supplémentaire apparaît, i.e. $\mathbb{V}_\Theta(\widetilde{m}(\mathbf{x}, \Theta)) = q(1-q)||\mathbf{x}||_2^2 > 0$.*

Considérant ces concepts, il est désormais possible de définir une forêt aléatoire comme étant la méthode de prédiction moyenne construit à partir de B arbres de régression stochastiques. Plus précisément, soit $\{\Theta^{(b)}\}_{b\in[B]}$ une suite de variables aléatoires i.i.d. de loi $\mathbb{P}_\Theta$ et

$\{\widetilde{m}_{tree}(\cdot, \Theta^{(b)})\}_{b\in[B]}$ une suite d'arbres de régression stochastiques ; une forêt aléatoire est alors définie par

$$\widetilde{m}_{rf}(\,\cdot\,, \{\Theta^{(b)}\}_{b\in[B]}) := \frac{1}{B}\sum_{b\in[B]} \widetilde{m}_{tree}(\,\cdot\,, \Theta^{(b)}). \tag{7}$$

Par souci de notation, nous utiliserons $\widetilde{m}_{rf}^{(B)}$ pour nommer l'estimateur (7).

Exemple 3.4. *Algorithme original de Breiman (2001).*

1. *Selectionner B échantillons bootstrap (ré-échantillonage d'individus de taille N avec remise) dans U denotés $\{U(\Theta_b)\}_{b\in[B]}$.*
2. *Dans chaque réplication bootstrap, $U(\Theta_b)$, construire un arbre de régression stochastique $\widetilde{m}(\,\cdot\,, \Theta_b)$ en utilisant le critère CART défini en Exemple 3.1, où le critère est optimisé seulement sur p_0 covariables parmi les p disponibles. Les p_0 covariables sont choisies de manière uniformément aléatoire (sans remise) parmi les p disponibles, selon Θ_b, à chaque split.*

Exemple 3.5. *Forêts aléatoires uniformes (Biau et al., 2008; Scornet, 2016).*
Tous les arbres de la forêt se comportent de manière similaire ; par conséquent, nous décrivons ci-dessous le comportent d'un arbre générique. Pour commencer, considérons comme feuille initiale l'hypercube unité $[0;1]^p$. Ensuite, récursivement, l'algorithme fait ses splits de la manière suivante. La procédure suivante est répétée K fois, avec $K \in \mathbb{N}$, un hyper-paramètre choisi en amont par l'utilisateur.

1. *Une feuille G est choisie uniformément au hasard parmi les feuilles existantes.*
2. *Une variable X_j est choisie uniformément au hasard parmi les p covariables $X_1, X_2, ..., X_p$.*
3. *Un split est réalisé dans la feuille G sur l'axe X_j à une localisation choisie uniformément au hasard.*

Remarque 3.1. *Initialement, le terme forêt aléatoire fait référence à l'algorithme proposé par Breiman (2001), décrit dans l'Exemple 3.4. Toutefois, la définition donnée dans cette section décrit une classe de forêts aléatoires, plutôt qu'un algorithme particulier. En effet, à chaque règle de partitionnement et manière d'utiliser la variable de randomisation Θ il est possible de définir un algorithme de "forêt aléatoire". Par conséquent, la définition donnée ci-dessus est très générale et englobe de nombreux algorithmes (y compris l'algorithme original de Breiman (2001)). Il est aussi important de noter que les arbres de régression (non stochastiques) tels que définis ci-avant font aussi partie de cette classe ; en effet, en prenant $B = 1$ et une règle de partitionnement n'utilisant pas de variable de randomisation, on obtient effectivement un arbre de régression.*

4 Estimation par modélisation assistée à l'aide d'arbres et forêts

4.1 Définition de l'estimateur par forêt aléatoire

Dans cette section, nous considérons un cadre de travail dans lequel la variable d'intérêt est supposée complètement observée au niveau de l'échantillon (par conséquent, sans valeur manquante). Nous faisons de plus l'hypothèse que l'information auxiliaire est connue sur l'ensemble de la population. Nous avons donc à notre disposition l'information contenue dans l'ensemble

$$D_{ma} := \{(\mathbf{x}_k, y_k); k \in S\} \bigcup \{\mathbf{x}_k; k \in U\backslash S\}. \tag{8}$$

L'estimateur d'Horvitz-Thompson (2) est un estimateur qui requière seulement la connaissance du n-uplet $\{y_k\}_{k\in S}$ et des poids de sondage. Par conséquent, utiliser un tel estimateur dans

un contexte où D_{ma} est connu reviendrait à ne pas utiliser l'intégralité de l'information disponible. Notre objectif est donc de construire un estimateur, aussi efficace que possible, en utilisant l'intégralité de l'information contenue dans D_{ma}. L'approche dite par modélisation assistée permet de construire de tels estimateurs. Cette approche a été étudiée pour un grand nombre de modèles, allant des méthodes paramétriques (Robinson and Särndal, 1983), aux méthodes non paramétriques tels que les polynômes locaux (Breidt and Opsomer, 2000), B-splines (Goga, 2005) et (Goga and Ruiz-Gazen, 2014), splines pénalisés (Breidt et al., 2005; McConville and Breidt, 2013), réseaux de neurones (Montanari and Ranalli, 2005), modèles additifs généralisés (Opsomer et al., 2007), et arbres de régression (McConville and Toth, 2019).

Le point de départ pour la construction d'un tel estimateur est l'estimateur par différence (Cassel et al., 1976). Dagdoug et al. (2021) définissent l'estimateur par différence basé sur des forêts aléatoires de la manière suivante

$$\widehat{t}_{dif}^{(B)} := \sum_{k\in U} \widetilde{m}_{rf}^{(B)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widetilde{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}, \tag{9}$$

où $\widetilde{m}_{rf}^{(B)}$ est défini en (7). Naturellement, en pratique, l'estimateur $\widehat{t}_{dif}^{(B)}$ n'est pas réalisable. En effet, celui-ci est construit à partir de la méthode de prédiction $\widetilde{m}_{rf}$, lui-même construit à partir de D_U, et par conséquent à partir de données inconnues. Dagdoug et al. (2021) ont proposé d'estimer la méthode de prédiction inconnue $\widetilde{m}_{rf}$ par $\widehat{m}_{rf1}^{(B)}$ à partir de l'information contenue dans D_{ma} :

$$\widehat{m}_{rf1}^{(B)}(\mathbf{x}) := \frac{1}{B}\sum_{b\in[B]}\sum_{k\in S(\Theta_b)} \frac{\pi_k^{-1}\mathbb{1}_{\mathbf{x}_k\in\widehat{A}^{(b)}(\mathbf{x})}}{\sum_{\ell\in S(\Theta_b)}\pi_\ell^{-1}\mathbb{1}_{\mathbf{x}_\ell\in\widehat{A}^{(b)}(\mathbf{x})}} y_k, \tag{10}$$

où $S(\Theta_b)$ désigne le ré-échantillon qui a servi à construire le b-ème arbre et $\widehat{A}^{(b)}(\mathbf{x})$ la feuille du b-ème arbre contenant le point $\mathbf{x}$. Les sommes sur la population sont donc remplacées par des sommes sur l'échantillon, et une pondération est appliquée. Les poids de sondages sont incorporés au numérateur et au dénominateur, permettant ainsi de mieux prendre en compte des plans de sondages à probabilités inégales. Si le plan de sondage considéré induit des probabilités d'inclusions égales pour tous les éléments de la population, alors la pondération appliquée en (10) s'annule et l'estimateur $\widehat{m}_{rf}^{(B)}$ de $\widetilde{m}_{rf}^{(B)}$ est simplement un estimateur construit en remplaçant les sommes sur la population par des sommes sur l'échantillon.

Nous pouvons désormais définir un estimateur par modélisation assistée par forêt aléatoire de la manière suivante :

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k)}{\pi_k}. \tag{11}$$

Remarque 4.1. *Comme mentionné en Remarque 3.1, la définition de forêt aléatoire utilisée en Section 3 englobe une large classe d'algorithmes. Nous pouvons par conséquent noter que, au vu des définitions de $\widehat{m}_{rf1}^{(B)}$ et $\widehat{t}_{rf1}^{(B)}$, l'équation $\widehat{t}_{rf1}^{(B)}$ définit une classe d'estimateurs plutôt qu'un estimateur particulier. Plus précisément, dénotons par $\mathcal{F}_{rf}(D_{ma}, B)$ l'ensemble des fonctions de type forêts aléatoires pondérées avec B arbres, entraînées sur $\{(\mathbf{x}_k, y_k); k \in S\}$. Dans cet article, $\widehat{t}_{rf1}^{(B)}$ représente en réalité un élément quelconque de l'ensemble*

$$\mathcal{T}_{rf}(D_{ma}, B) := \left\{ \widehat{t} = \sum_{k\in U} f(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - f(\mathbf{x}_k)}{\pi_k};\ f \in \mathcal{F}_{rf}(D_{ma}, B) \right\}.$$

Observons que $\mathcal{T}_{rf}(D_{ma}, 1)$ revient à l'espace des estimateurs par modélisation assistée basés sur des arbres de régression (strochastiques ou non) ; ainsi, l'ensemble $\mathcal{T}_{rf}(D_{ma}, 1)$ contient donc

l'estimateur proposé par McConville and Toth (2019). Les résultats présentés dans cette section étant indépendants du nombre d'arbre B, nous énoncerons donc ces résultats pour un élément $\widehat{t}_{rf1}^{(B)}$ quelconque de

$$\mathcal{T}_{rf}\left(D_{ma}\right) := \bigcup_{B\in\mathbb{N}^*} \mathcal{T}_{rf}\left(D_{ma}, B\right).$$

4.2 Quelques propriétés de $\widehat{t}_{rf1}^{(B)}$

Proposition 4.1. *Considérons un estimateur $\widehat{t}_{rf1}^{(B)}$ par forêt aléatoire.*

1. *L'estimateur $\widehat{t}_{rf1}^{(B)}$ peut être vu comme une moyenne d'estimateurs par modélisation assistée :*

$$\widehat{t}_{rf1}^{(B)} = \frac{1}{B}\sum_{b\in[B]} \widehat{t}_{tree1}^{(b)},$$

où $\widehat{t}_{tree1}^{(b)}$ désigne l'estimateur par modélisation assistée basé sur le b-ème arbre de la forêt, i.e.

$$\widehat{t}_{tree1}^{(b)} = \sum_{k\in U} \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k)}{\pi_k},$$

et $\widehat{m}_{tree1}^{(b)}$ est un estimateur de (5).

2. *L'estimateur $\widehat{t}_{rf1}^{(B)}$ peut être écrit*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k) + \frac{1}{B}\sum_{b\in[B]}\sum_{k\in O_b^{(S)}} \frac{y_k - \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k)}{\pi_k}, \tag{12}$$

où $O_b^{(S)} := S - S(\Theta_b)$ dénote les éléments dits "out-of-bag" de l'arbre b.

3. *Si $\widehat{m}_{rf1}^{(B)}$ n'utilise pas de mécanisme de ré-échantillonnage, alors $\widehat{t}_{rf1}^{(B)}$ possède la propriété de projection :*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k).$$

Le point 1. ci-dessus révèle que l'estimateur $\widehat{t}_{rf1}^{(B)}$ est en réalité une moyenne de B estimateurs par modélisation assistée. Plus généralement, il est possible de montrer qu'une moyenne d'estimateurs par modélisation assistée reste un estimateur par modélisation assistée. Les estimateurs par forêts aléatoires ont de plus la propriété suivante : si $\widehat{t}_{rf1}^{(B)}$ et $\widehat{t}_{rf1'}^{(B)}$ désignent deux estimateurs par forêts avec chacun B arbres et construits à partir du même algorithme, alors leur moyenne $\left(\widehat{t}_{rf1}^{(B)} + \widehat{t}_{rf1}^{(B)}\right)/2$ est un estimateur par forêt aléatoire construit sur $2B$ arbres. Cette propriété n'est plus exacte si l'on effectue la moyenne de deux forêts avec des nombres d'arbres différents. Le point 2. montre que l'estimateur par forêts aléatoires calcule ses résidus uniquement sur les éléments non sélectionnés pour construire le modèle. Ceci est un point positif inattendu et propre aux estimateurs par modélisation assistée construits sur des algorithmes de type "bagging", voir Breiman (1996). Par conséquent, le second terme à droite de (21) peut-être vu comme une protection contre de mauvaises prédictions du modèle et contre le sur-apprentissage. En particulier, cela implique que l'efficacité des forêts n'incluant pas de mécanisme de ré-échantillonnage repose entièrement sur le modèle de prédiction (conséquence du point 3.).

Notons qu'il est aussi possible d'écrire $\widehat{m}_{rf1}^{(B)}$ comme une somme pondérée des valeurs de la variable d'intérêt :

$$\widehat{m}_{rf1}^{(B)}(\mathbf{x}) = \sum_{k\in S} \widehat{W}_{k1}^{(B)}(\mathbf{x}) y_k, \tag{13}$$

où

$$\widehat{W}_{k1}^{(B)}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \frac{\pi_k^{-1}\psi_k^{(b)}\mathbb{1}_{\mathbf{x}_k \in \widehat{A}(\mathbf{x})}}{\sum_{\ell \in S}\pi_\ell^{-1}\psi_\ell^{(b)}\mathbb{1}_{\mathbf{x}_\ell \in \widehat{A}(\mathbf{x})}}, \qquad k \in S, \tag{14}$$

avec $\psi_\ell^{(b)} = 1$ si $\ell \in S(\Theta_b)$, 0 sinon.

Proposition 4.2. *Considérons un estimateur $\widehat{t}_{rf1}^{(B)}$ par forêt aléatoire.*

1. *L'estimateur $\widehat{t}_{rf1}^{(B)}$ peut être vu comme une somme pondérée des valeurs de Y :*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k \in S} w_{k1}^{(B)} y_k,$$

où

$$w_{k1}^{(B)} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell \in U}\widehat{W}_{k1}^{(B)}(\mathbf{x}_\ell)\left(1 - \frac{I_\ell}{\pi_\ell}\right)\right\}, \quad k \in S. \tag{15}$$

2. *Quel que soit le plan de sondage considéré, on a $\sum_{k\in S} w_{k1}^{(B)} = N$, pour tout $s \in \mathcal{S}$.*
3. *On a*

$$w_{k1}^{(B)} = 1/\pi_k$$

pour les éléments k jamais sélectionnés dans les sous-échantillons, i.e. $k \in \bigcap_{b=1}^{B} O_b^{(S)}$.
4. *Si le mécanisme de ré-échantillonnage choisi est de type bootstrap (avec remplacement), la probabilité qu'un élément ne soit jamais sélectionné tend vers 0 quand B tend vers l'infini.*
5. *Les poids $\{w_{k1}^{(B)}\}_{k\in S}$ sont indépendants de la variable d'intérêt si et seulement si la règle de partitionnement utilisée par les arbres de la forêt est indépendante de la variable d'intérêt.*

L'estimateur $\widehat{t}_{rf1}^{(B)}$ est donc un estimateur pouvant s'écrire comme une somme pondérée des valeurs de la variable d'intérêt, au sens où il s'écrit sous la forme d'une somme pondérée de la variable d'intérêt. En revanche, ces poids $\{w_{k1}\}_{k\in S}$ peuvent dépendre de la variable d'intérêt si le mécanisme utilisé pour construire la partition est lui même dépendant de la variable d'intérêt (souvent le cas, en pratique). Par conséquent, l'application de ce système de pondération à d'autres variables d'intérêts doit être faite prudemment. En revanche, lorsque le mécanisme de split ne dépend pas de la variable d'intérêt, l'estimateur $\widehat{t}_{rf1}^{(B)}$ appartient donc à la classe des estimateurs linéaires. La proposition précédente révèle aussi que la somme des poids est toujours égale à la taille de la population, et qu'il est possible que certains de ces poids soient en réalité égaux aux poids de sondage initiaux. Cependant, pour de larges forêts, ce phénomène ne se réalise que très rarement. Même lorsque ce scénario se produit, il est à noter que ces éléments sont tout de même utilisés dans la construction de l'estimateur : ils contribuent au terme de correction dans la forme 2. Enfin, nous pouvons noter que, lorsqu'aucun mécanisme de rééchantillonage n'est utilisé dans l'algorithme, alors les poids (15) sont toujours positifs (voir notamment Dagdoug et al. (2021)).

4.3 Propriétés asymptotiques

Pour obtenir les propriétés asymptotiques, nous supposons le cadre asymptotique de Isaki and Fuller (1982). Considérons pour cela une suite emboitée infinie de populations $\{U_v\}_{v\to\infty}$ de tailles $N_v \to \infty$ et d'échantillons $S_v \subset U_v$ de taille $n_v \to \infty$. Les résultats que nous décrivons dans la suite requièrent certaines hypothèses de régularité concernant le plan de sondage, la variable d'intérêt et l'algorithme de forêt sur lequel l'estimateur est construit, voir Dagdoug et al.

(2021) pour plus de précisions. La plupart de ces hypothèses sont communément utilisées dans la littérature et vérifiées en pratique, voir par exemple Breidt and Opsomer (2000) et McConville and Toth (2019) pour plus de détails. Nous considérerons désormais des suites d'estimateurs $\{\widehat{t}_{rf1,v}\}_{v\in\mathbb{N}}$ tels que, pour tout v, $\widehat{t}_{rf1,v} \in \mathcal{T}^*_{rf}(D_{ma,v})$, où $\mathcal{T}^*_{rf}(D_{ma,v})$ dénote la restriction de $\mathcal{T}_{rf}(D_{ma,v})$ aux estimateurs construits sur des forêts avec un mécanisme de ré-échantillonnage sans remplacement et avec au minimum n_{0v} éléments dans chaque feuille. L'étude asymptotique des estimateurs ne remplissant pas ces conditions n'est pas incluse dans Dagdoug et al. (2021). Pour simplifier les notations, l'indice v sera supprimé dans la notation des estimateurs $\widehat{t}_{rf1,v}$.

Résultat 4.1. *Il existe des constantes $C_1 > 0$ et $C_2 > 0$ telles que :*

$$\mathbb{E}_p\left[\left|\frac{1}{N_v}\left(\widehat{t}^{(B)}_{rf1} - t_y\right)\right|\right] \leqslant \frac{C_1}{\sqrt{N_v}} + \frac{C_2}{n_{0v}}, \quad \textit{presque sûrement (p.s.).} \tag{16}$$

Il est donc possible de borner l'erreur L^1 de chaque estimateur dans la classe $\mathcal{T}^*_{rf}(D_{ma})$. De plus, si n_{0v} tend vers l'infini, alors cette borne décroit vers 0. Par conséquent, dans ce cas de figure, les estimateurs de $\mathcal{T}^*_{rf}(D_{ma})$ sont asymptotiquement sans biais et consistant pour t_y. Dans le reste de la section, nous supposerons donc que n_{0v} tend vers l'infini lorsque n tend vers l'infini. Pour obtenir certains des résultats suivants, nous aurons en réalité besoin de considérer n_{0v} tel que $\sqrt{n_v}/n_{0v}$ converge vers 0.

L'équivalence suivante permet de guider notre suggestion au regard de l'estimateur de variance et de déterminer la distribution asymptotique de $\widehat{t}_{rf}$.

Résultat 4.2. *L'estimateur $\widehat{t}_{rf}$ est équivalent à l'estimateur par différence généralisée $\widehat{t}^{(B)}_{dif}$:*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}^{(B)}_{rf1} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}^{(B)}_{dif} - t_y\right) + o_{\mathbb{P}}(1),$$

où $\widehat{t}^{(B)}_{dif}$ est donné en (9).

Ce résultat nous permet de déduire la variance asymptotique de $\widehat{t}^{(B)}_{rf1}$, égale à

$$\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}^{(B)}_{rf1}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}(\pi_{kl} - \pi_k\pi_\ell)\frac{y_k - \widetilde{m}^{(B)}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell - \widetilde{m}^{(B)}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{17}$$

En pratique, cette variance ne peut pas être calculée et nous proposons donc de l'estimer par

$$\widehat{V}^{(B)}_{rf1} = \frac{1}{N_v^2}\sum_{k\in S_v}\sum_{\ell\in S_v}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k - \widehat{m}^{(B)}_{rf1}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell - \widehat{m}^{(B)}_{rf1}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{18}$$

Il s'agit d'un estimateur consistant est asymptotiquement sans biais pour la variance asymptotique de $\widehat{t}^{(B)}_{rf1}$, comme le garantit le résultat suivant.

Résultat 4.3. *L'estimateur de variance $\widehat{V}_{rf1}$ est convergent pour $\mathbb{AV}_p\left(\widehat{t}_{rf1}\right)$, c'est-à-dire,*

$$\lim_{v\to\infty}\mathbb{E}_p\left(\frac{n_v}{N_v^2}\left|\widehat{V}^{(B)}_{rf1} - \mathbb{AV}_p(\widehat{t}^{(B)}_{rf1})\right|\right) = 0.$$

Afin de pouvoir déterminer des intervalles de confiances asymptotiques, il est nécessaire de déterminer la distribution asymptotique de l'estimateur proposé qui est obtenue sous l'hypothèse supplémentaire que l'estimateur par différence généralisée $\widehat{t}^{(B)}_{dif}$ suit une distribution normale.

Résultat 4.4. *Si*

$$\frac{N_v^{-1}\left(\widehat{t}_{dif}^{(B)} - t_y\right)}{\sqrt{\mathbb{AV}_p(\widehat{t}_{rf1})}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}(0,1)\,,$$

alors

$$\frac{N_v^{-1}\left(\widehat{t}_{rf1}^{(B)} - t_y\right)}{\sqrt{\widehat{V}_{rf1}}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}(0,1)\,.$$

Comme illustré en Exemple 3.3, les modèles stochastiques possèdent une source de variation supplémentaire, introduite par les variables de randomisation. L'estimateur de variance (18) ne prend pas en compte ces variations additionnelles (induite par Θ, voir 3.3). Toutefois, il est possible de montrer qu'il existe une constante positive C telle que

$$\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf1}^{(B)}}{N_v}\right) \leqslant \frac{C}{B}.$$

Par conséquent, si B est choisi suffisamment grand, alors les variations provenant de la randomisation introduite sont négligeables et ne nécessitent pas d'être estimées.

Dagdoug et al. (2021) ont noté que les estimateurs de variance construits sous la forme (18), traditionnellement utilisés dans la littérature, peut s'avérer sous-estimer très fortement la variance si le modèle de prédiction utilisé est trop complexe. Plus précisément, si le modèle utilisé $\widehat{m}_{rf1}^{(B)}$ fait du sur-apprentissage (voir Hastie et al. (2011) pour détails), alors les résidus $\{y_k - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k)\}_{k\in S}$ seront sous-estimés, et l'estimateur $\widehat{V}_{rf1}$ sous-estimera ainsi la variance de l'estimateur par modélisation assistée basé sur $\widehat{m}_{rf1}^{(B)}$. Le cas extrême étant lorsque $\widehat{m}_{rf1}^{(B)}$ interpole parfaitement chacune des observations (i.e. $\widehat{m}_{rf1}^{(B)}(\mathbf{x}_k) = y_k, k \in S$), auquel cas l'estimateur de variance serait égal à 0, n'impliquant bien entendu par que la vraie variance de l'estimateur soit 0. Dagdoug et al. (2021) ont donc proposé un estimateur utilisant la validation croisée pour remédier à ce problème ; les simulations semblent indiquer que l'estimateur proposé ne souffre plus de ce défaut. Cette observation est généralisable à de nombreuses méthodes de prédictions.

Notons de plus que l'intégralité des résultats asymptotiques présentés dans cette section restent valables si l'on considère un cadre de travail dit "en grande dimension", dans lequel le nombre de covariables est autorisé à tendre vers l'infini. En particulier, contrairement au cas des modèles linéaires étudiés en grande dimension (voir Ta et al. (2020), Chauvet and Goga (2022), Dagdoug et al. (2020b)), aucune condition sur la vitesse de divergence du nombre de covariables n'est requise. Ce phénomène semble être expliquer par deux raisons : 1) les résultats asymptotiques présentés ici sont des propriétés évaluées sous le plan de sondage et non sous le modèle ; 2) la "structure asymptotique" d'un estimateur linéaire (e.g. GREG) versus un estimateur de type arbre est différente en essence. En effet, un estimateur linéaire peut être vu comme un estimateur calé (voir Deville and Särndal (1992) pour détails) sur les p covariables. Par conséquent, lorsque $p = p_v$ est autorisé à tendre vers l'infini, il s'agit en réalité d'imposer un nombre de contraintes de calage toujours plus grand. Ce n'est pas le cas des estimateurs construits sur des arbres de régression : ceux-ci peuvent aussi être vus comme des estimateurs calés, mais sur les T covariables formées par les indicatrices d'appartenances aux T feuilles de l'arbre, et non sur les p covariables. Typiquement, T est fonction de n_0 plutôt que de p, et, par conséquent, le nombre de contraintes de calages imposés à un estimateur construit sur un arbre reste fixe lorsque p tend vers l'infini.

Enfin, en pratique, la majorité des packages permettant l'utilisation de forêts aléatoires ne permettent pas de faire des prédictions pondérées au sens de (10). Dagdoug et al. (2020b) ont

conduit une étude empirique dans laquelle les forêts aléatoires utilisées n'incluent pas les probabilités d'inclusion. Dans cette étude, il est mis en évidence qu'une telle pratique, sans ajustement de l'algorithme utilisé, peut résulter en une augmentation du biais de l'estimateur par modélisation assistée (défini ci-dessous) dans certains cas. Plus précisément, si les poids de sondages ne sont pas incorporés dans les prédictions du modèle considéré, alors il est particulièrement important (dans le cas de plans de sondage informatifs) d'inclure les variables ayant servies à la construction du plan de sondage (variables dites de design) dans le modèle, sous la forme (éventuellement) de covariables supplémentaires. Naturellement, cette pratique est généralisable pour l'intégralité des modèles de prédiction. Dans ce cas, les estimateurs par modélisation conservent généralement leurs bons comportements. En revanche, dans le cas des forêts aléatoires, il est important de noter que certains algorithmes utilisent les variables de randomisation afin de faire des sélections aléatoires de variables, voir notamment l'algorithme de Breiman décrit en 3.1. Dans ce cas, si le nombre de covariables choisies à chaque split est faible et le nombre de covariables est important, alors, avec grande probabilité, les variables de design ne seront pas considérées. Dans de tels scénarios, il est possible que l'estimateur par modélisation assistée basé sur ce modèle soit fortement biaisé. Pour remédier à ce problème, nous suggérons de choisir avec probabilité 1 chacune des variables de design, et de ne réaliser la sélection aléatoire que sur variables n'ayant pas servies à la construction du plan de sondage. Une telle pratique est notamment réalisable à l'aide du `R` package `ranger`, voir Wright and Ziegler (2015).

5 Imputation par forêts aléatoires

5.1 Définition de l'estimateur imputé par forêts aléatoires

Nous allons désormais considérer un cadre de travail dans lequel la variable d'intérêt est supposée partiellement observée car certains éléments sélectionnés dans l'échantillon ont refusé de répondre. Nous supposerons de plus que les covariables sont totalement observés au niveau de l'échantillon (et non au niveau de la population). Plus précisément, soit $\{r_k\}_{k\in U}$ le N-uplet des indicatrices de non-réponse, c'est-à-dire $r_k := 1$ si y_k est observé, et $r_k := 0$ sinon, pour tout élément k de la population. Soient $S_r := \{k \in S; r_k = 1\}$ et $S_m := \{k \in S; r_k = 0\}$ les échantillons des répondants et des non-répondants pour la variable d'intérêt, respectivement. À notre disposition, nous disposons de l'information contenue dans

$$D_{imp} := \{(\mathbf{x}_k, y_k); k \in S_r\} \bigcup \{\mathbf{x}_k; k \in S_m\}.$$

L'estimateur d'Horvitz-Thompson (2) n'est désormais plus utilisable, car celui-ci dépend des valeurs $\{y_k\}_{k\in S_m}$, qui sont inconnues. Dans un tel cadre de travail, il est pratique courante d'avoir recours à l'imputation pour remplacer les valeurs non observées par des valeurs prédites. Pour que l'estimation de la fonction de régression m soit possible, il est nécessaire de faire l'hypothèse que le mécanisme de non-réponse est ignorable ou MAR (Rubin, 1976), c'est-à-dire, pour tout $k \in S$, nous supposons que

$$\mathbb{P}\left(r_k|\mathbf{x}_k, y_k\right) = \mathbb{P}\left(r_k|\mathbf{x}_k\right).$$

Pour une méthode de prédiction $\widehat{m}$ quelconque, alors, nous définissons l'estimateur imputé de t_y comme suit

$$\widehat{t}_{\widehat{m}} := \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\widehat{m}(\mathbf{x}_k)}{\pi_k} = \sum_{k\in S} \frac{y_k^*}{\pi_k}, \tag{19}$$

où $y_k^* := y_k$ pour $k \in S_r$ et $y_k^* := \widehat{m}(\mathbf{x}_k)$ pour $k \in S_m$. Les propriétés théoriques de l'estimateur imputé ont été étudiés pour certaines méthodes de prédictions ; notamment : la méthode du plus proche voisin Chen and Shao (2000, 2001); Yang and Kim (2019), la méthode des scores Little (1986); Haziza and Beaumont (2007), le "predictive mean matching" Yang and Kim (2017),

méthodes à noyaux Zhong and Chen (2014), pour n'en citer que quelques-uns. Pour davantage d'informations sur le sujet, le lecteur intéressé peut se référer à Chen and Haziza (2019) ; une large investigation empirique du comportement de ces méthodes est menée dans Dagdoug et al. (2020a). Le cas de l'imputation par arbres de régression et forêts aléatoires est traité dans Dagdoug et al. (2022), et leurs résultats principaux sont présentés dans cette section. En particulier, comme dans la section précédente, nous nous concentrons principalement sur le cas plus général des forêts aléatoires.

Remarque 5.1. *Il est possible de voir que la méthode des scores mentionnée ci-dessus, ayant pour objectif de créer des classes puis d'imputer la moyenne dans chaque classe, est en réalité un cas particulier d'arbre de régression. Par conséquent, les résultats mentionnés dans cette section contribuent aussi à une analyse de cette méthodologie.*

Soit $\widehat{m}_{rf2}^{(B)}$ un estimateur de m obtenu selon une forêt aléatoire (avec règle de partitionnement quelconque), non pondérée, construite sur $\{(\mathbf{x}_k, y_k); k \in S_r\}$:

$$\widehat{m}_{rf2}^{(B)}(\mathbf{x}) = \sum_{k \in S_r} \widehat{W}_{k2}^{(B)}(\mathbf{x}) y_k,$$

où

$$\widehat{W}_{k2}^{(B)}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\psi_k^{(b)} \mathbb{1}_{\mathbf{x}_k \in \widehat{A}(\mathbf{x})}}{\sum_{\ell \in S_r} \psi_\ell^{(b)} \mathbb{1}_{\mathbf{x}_\ell \in \widehat{A}(\mathbf{x})}}, \qquad k \in S_r,$$

La classe des estimateurs imputés par forêts aléatoires $\widehat{t}_{rf2}^{(B)}$ est définie par l'ensemble des éléments de la forme

$$\widehat{t}_{rf2}^{(B)} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k}. \tag{20}$$

Ici, contrairement à la section précédente, les prédictions de l'algorithme sont non pondérées. Cela influera notamment sur les propriétés de l'estimateur, voir notamment la Remarque 5.2 ci-dessous. Il est toutefois possible de voir de fortes connexions entre le cadre de travail dit par modélisation assistée, présenté en Section 4, et le cadre avec valeurs manquantes présenté ici. Tout d'abord, il existe une symétrie entre les données disponibles dans chacun des scénarios (entre D_{ma} et D_{imp}). Traditionnellement, les estimateurs imputés sont définis de la manière précédente, i.e. (24). Il est toutefois aussi possible de définir $\widehat{t}_{rf2}^{(B)}$ d'une manière relativement similaire à la définition donnée en (11) ; plus précisément, il est possible de définir alternativement $\widehat{t}_{rf2}^{(B)}$ de la manière suivante :

$$\widehat{t}_{rf2}^{(B)} := \sum_{k \in S} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k} + \sum_{k \in S_r} \frac{y_k - \widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k}.$$

Par conséquent, beaucoup des propriétés observées pour l'estimateur par modélisation assistée $\widehat{t}_{rf1}^{(B)}$ seront aussi observées pour l'estimateur imputé $\widehat{t}_{rf2}^{(B)}$. Les propositions suivantes illustrent cette symétrie.

5.2 Quelques propriétés de $\widehat{t}_{rf2}^{(B)}$

Proposition 5.1. *Considérons un estimateur imputé par forêt aléatoire $\widehat{t}_{rf2}^{(B)}$.*

1. *L'estimateur $\widehat{t}_{rf2}^{(B)}$ peut être vu comme une moyenne d'estimateurs imputés :*

$$\widehat{t}_{rf2}^{(B)} = \frac{1}{B} \sum_{b \in [B]} \widehat{t}_{tree2}^{(b)},$$

où $\widehat{t}_{tree2}^{(b)}$ désigne l'estimateur imputé basé sur le b-ème arbre de la forêt.

2. *Si le plan de sondage est à probabilité d'inclusions égales, alors l'estimateur $\widehat{t}_{rf2}^{(B)}$ peut être écrit*

$$\widehat{t}_{rf2}^{(B)} = \sum_{k\in S} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k} + \frac{1}{B}\sum_{b\in[B]}\sum_{k\in O_b(S_r)} \frac{y_k - \widehat{m}_{tree2}^{(b)}(\mathbf{x}_k)}{\pi_k}, \tag{21}$$

où $O_b(S_r) := S_r - S_r(\Theta_b)$, et $\widehat{m}_{tree2}^{(b)}$ dénote le b-ième arbre de la forêt $\widehat{m}_{rf2}^{(B)}$.

3. *Si le plan de sondage est à probabilité d'inclusions égales et si $\widehat{m}_{rf2}^{(B)}$ n'utilise pas de mécanisme de ré-échantillonnage, alors $\widehat{t}_{rf2}^{(B)}$ possède la propriété de projection :*

$$\widehat{t}_{rf2}^{(B)} = \sum_{k\in S} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k}.$$

Remarque 5.2. *En comparant les Proposition 4.1 et Proposition 5.1, on observe que les propriétés de $\widehat{t}_{rf1}^{(B)}$ et $\widehat{t}_{rf2}^{(B)}$ sont très proches. Toutefois, nous pouvons noter que les points 2 et 3 de la Proposition 5.1 ne sont valables que si le plan de sondage considéré induit des probabilités d'inclusions égales ; cette hypothèse n'est nécessaire que pour l'estimateur imputé. Cette condition supplémentaire provient du fait que les prédictions de $\widehat{m}_{rf2}^{(B)}$ ne sont pas pondérés par les poids de sondage, alors que celles de $\widehat{m}_{rf1}^{(B)}$ le sont.*

Proposition 5.2. *Considérons un estimateur imputé par forêt aléatoire $\widehat{t}_{rf2}^{(B)}$.*

1. *L'estimateur $\widehat{t}_{rf2}^{(B)}$ peut être écrit comme une somme pondérée des valeurs de la variable d'intérêts :*

$$\widehat{t}_{rf2} = \sum_{k\in S_r} w_{k2}^{(B)} y_k,$$

où

$$w_{k2}^{(B)} = \frac{1}{\pi_k} + \sum_{\ell\in S_m} \frac{\widehat{W}_{k2}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell} = \frac{1}{\pi_k} + \frac{1}{B}\sum_{b\in[B]} \psi_k^{(b)} \frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{N_b\left(\mathbf{x}_k, S_r(\Theta_b)\right)}, \qquad k \in S_r, \tag{22}$$

où $\widehat{N}_b\left(\mathbf{x}_k, S_m\right)$ dénote le nombre d'éléments pondérés de S_m appartenant à la feuille contenant $\mathbf{x}_k$ et $N_b\left(\mathbf{x}_k, S_r(\Theta_b)\right)$ désigne le nombre d'éléments de $S_r(\Theta_b)$ appartenant à la feuille contenant le point $\mathbf{x}_k$.

2. *Dans le cas d'un arbre non stochastique, si le plan de sondage est à probabilité d'inclusions égales, alors on a*

$$w_{k2}^{(B)} = \frac{1}{\pi_k} \times \left(1 + \frac{1}{B}\sum_{b\in[B]} \psi_k^{(b)} \frac{N\left(\mathbf{x}_k, S_m\right)}{N\left(\mathbf{x}_k, S_r(\Theta_b)\right)}\right).$$

3. *Si les poids initiaux sont calibrés sur la constante $\sum_{k\in S} \frac{1}{\pi_k} = N$, alors $\sum_{k\in S} w_{k2}^{(B)} = N$.*

4. *On a*

$$w_{k2}^{(B)} = \frac{1}{\pi_k}$$

pour les éléments $k \in \bigcap_{b=1}^{B} O_b^{(S_r)}$.

5. *Si il y a au minimum n_0 éléments dans les feuilles de chaque arbre, alors les poids sont bornés de la façon suivante*

$$d_k \leqslant w_{k2}^{(B)} \leqslant d_k \left(1 + \frac{n_m}{n_0}\right), \quad p.s. \qquad k \in S_r. \tag{23}$$

Ces bornes peuvent être atteintes.

6. *Les poids $\{w_{k2}^{(B)}\}_{k \in S_r}$ sont indépendants de la variable d'intérêt si et seulement si la règle de partitionnement utilisée par les arbres de la forêt est indépendante de la variable d'intérêt.*

Comme dans le cas de l'estimateur par modélisation assistée, l'estimateur imputé par forêt peut être écrit sous la forme d'une somme pondérée des valeurs de la variable d'intérêt. Dans le cas de l'estimateur imputé, ces poids révèlent beaucoup d'information sur le comportement de l'estimateur. Tout d'abord, on observe que, les poids d'imputation sont toujours plus grands ou égaux aux poids initiaux et sont calibrés sur ces derniers. Si l'on considère les poids d'un estimateur basé sur un arbre non-stochastique (e.g. CART, méthode des scores) avec probabilités d'inclusions égales, alors on a

$$w_{k2}^{(1)} = d_k \times \left(1 + \frac{N\left(\mathbf{x}_k, S_m\right)}{N\left(\mathbf{x}_k, S_r\right)}\right) = d_k \times \left\{1 + R_{mr}\left(\mathbf{x}_k\right)\right\}, \qquad k \in S_r.$$

On observe donc que, si la plupart des éléments ayant des caractéristiques similaires à un élément $k \in S_r$ n'ont pas répondus, alors un poids important sera attribué à cet élément car le ratio $R_{mr}\left(\mathbf{x}_k\right)$ sera important. Dans le cas contraire, si quasiment tous les éléments ayant des caractéristiques similaires à l'individu $k \in S_r$ ont répondus, les poids d'imputation seront très proches des poids initiaux. Il s'agit d'un comportement souhaité : lorsque nous avons peu d'exemplaires de certains éléments, les poids attribués sont larges, dans le cas contraire, ils sont proches des poids originaux. En particulier, un élément répondant a un poids d'imputation égal à son poids initial si et seulement si tous les éléments dans sa feuille sont des répondants. La même interprétation est valable dans le cas de probabilités d'inclusions inégales. En revanche, si l'on considère des arbres stochastiques et forêts aléatoires, ces propriétés sont perdues pour B faible. En effet, pour les éléments n'ayant pas été sélectionnés (ils peuvent être nombreux pour B faible), les poids d'imputation sont égaux aux poids initiaux. Indépendamment de ça, la somme des poids d'imputation reste égale à la somme des poids initiaux : un phénomène de compensation est donc nécessairement introduit et ces poids peuvent être relativement instables. Lorsque B est large, cette instabilité disparait et le comportement des poids de la forêt se rapproche très fortement du comportement des poids d'un arbre.

5.3 Propriétés asymptotiques

Contrairement aux résultats asymptotiques donnés en Section 4, nous allons ici donner des résultats qui seront valables, non pas pour une large classe d'estimateurs, mais seulement pour des estimateurs particuliers (précisés dans les résultats). Cette différence provient du fait que les estimateurs sont ici évalués sous la loi jointe induit par le modèle de superpopulation, le modèle de nonréponse ainsi que le plan de sondage, là où ils étaient évalués uniquement vis-à-vis du plan de sondage dans le cadre de travail par modélisation assistée.

Pour étudier les propriétés asymptotiques des estimateurs imputés par forêts, il est intéressant de considérer la méthode de prédiction infinie $\widehat{m}_{rf2}^{(\infty)}$ défini par

$$\widehat{m}_{rf2}^{(\infty)} := \lim_{B \to \infty} \widehat{m}_{rf2}^{(B)} = \lim_{B \to \infty} \frac{1}{B} \sum_{b \in [B]} \widehat{m}_{tree2}^{(b)}$$

ainsi que l'estimateur infini

$$\widehat{t}_{rf2}^{(\infty)} := \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k} = \mathbb{E}_\Theta\left[\widehat{t}_{rf2}^{(B)}\right]. \tag{24}$$

Il s'agit d'un estimateur d'intérêt purement théorique au sens où, il n'est pas possible de l'utiliser en pratique. Toutefois, cet estimateur infini est particulièrement intéressant pour trois raisons : 1) il est plus simple à étudier théoriquement que l'estimateur par forêt finie ; 2) il est plus efficient que l'estimateur par forêt finie, comme le révèle la proposition ci-dessous ; 3) il est approchable à précision donnée par l'estimateur par forêt finie.

Proposition 5.3. *Il existe $C > 0$ tel que*

$$0 \leqslant \mathbb{E}\left[\left(\frac{\widehat{t}_{rf2}^{(B)} - t_y}{N_v}\right)^2\right] - \mathbb{E}\left[\left(\frac{\widehat{t}_{rf2}^{(\infty)} - t_y}{N_v}\right)^2\right] \leqslant \frac{C}{B}. \tag{25}$$

Enfin, pour tout $\epsilon > 0$,

$$\mathbb{P}_\Theta\left(|\widehat{t}_{rf2}^{(B)} - \widehat{t}_{rf2}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2n_m^2\left(\frac{C_{2,Y} - C_{1,Y}}{\min_{k\in U}\pi_k}\right)^2}\right).$$

Cette proposition montre donc qu'il semble être intéressant de construire de larges forêts, au sens où l'estimateur par forêt infinie est plus efficient que l'estimateur par forêt finie. Nous restreignons désormais notre analyse au cas où, pour imputation dans S_v, nous choisissons B_v tel que, si $v_1 < v_2$ alors le nombre d'arbres utilisés pour imputer dans B_{v_1} est strictement inférieur au nombre d'arbres utilisés pour imputer dans B_{v_2}. Ceci permet l'utilisation du résultat (25) impliquant, dès lors que l'estimateur des forêts infinies est consistant dans L^2, la consistance L^2 de l'estimateur des forêts finies. De plus, nous nous restreignons au cas où $\widehat{t}_{rf2}^{(B)}$ est un estimateur imputé basé sur l'algorithme de forêts aléatoires au sens original de Breiman.

Résultat 5.1. *Sous certaines conditions usuelles, l'estimateur $\widehat{t}_{brf}$ converge en moyenne quadratique,*

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{rf2}^{(B)} - t_y\right)\right)^2\right] = 0.$$

Enfin, concernant l'estimation de la variance, de la même manière que pour les estimateurs par modélisation assistée, on montre que les variations dues aux variables de randomisation décroissent lorsque B diverge :

$$\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf2}^{(B)}}{N_v}\right) \leqslant \frac{C}{B_v}.$$

Il est par conséquent suffisant d'estimer la variance de $\widehat{t}_{rf2}^{(B)}$ vis-à-vis de la distribution jointe induite par le plan, le modèle et le mécanisme de non-réponse. Dans la plupart des cas, l'estimateur de variance "naïf"

$$\widehat{V}_{naive} := \sum_{k\in S}\sum_{\ell\in S} \Delta_{k\ell}\frac{y_k^*}{\pi_k}\frac{y_\ell^*}{\pi_\ell} \tag{26}$$

est un estimateur sévèrement biaisé ; l'utilisation d'estimateurs de variance spécifiques est donc nécessaire. Deux approches sont traditionnellement utilisées : l'approche "deux phases" Särndal

(1992), et l'approche "inversée" Shao and Steel (1999). Le lecteur intéressé peut se référer à Haziza and Vallée (2020) pour une revue des concepts et outils relatifs à l'estimation de variance des estimateurs imputés. Dans Dagdoug et al. (2022), deux estimateurs correspondants sont suggérés. Pour l'approche deux phases,

$$\widehat{V}_{sar} := \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix}, \tag{27}$$

où

$$\widehat{V}_{sam} := \widehat{V}_{naive} + \sum_{k\in S_m} d_k^2 \left(1-\pi_k\right)\widehat{\sigma}^2 \qquad \widehat{V}_{nr} := \sum_{k\in S} \gamma_k^2 \widehat{\sigma}^2, \qquad \widehat{V}_{mix} := \sum_{k\in S} \gamma_k \left(d_k - 1\right)\widehat{\sigma}^2,$$

avec $\widehat{\sigma}$ est un estimateur de la variance des résidus du modèle et $\gamma_k := r_k w_{k2} - d_k$ pour $k \in S$. Enfin, pour l'approche renversé, en supposant que la fraction de sondage n_v/N_v est négligeable,

$$\widehat{V}_{rev} := \sum_{k\in S}\sum_{\ell\in S} \Delta_{k\ell} \frac{\widehat{\xi}_k^{(B)}}{\pi_k} \frac{\widehat{\xi}_\ell^{(B)}}{\pi_\ell}, \tag{28}$$

où

$$\widehat{\xi}_k^{(B)} := \widehat{m}_{rf2}^{(B)}(\mathbf{x}_k) + r_k \cdot \frac{1}{B}\sum_{b\in[B]} \frac{N_b\left(\mathbf{x}_k, S\right)}{N_b\left(\mathbf{x}_k, S_r\right)} \cdot \left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)\right), \qquad k \in S.$$

6 Conclusion

L'avènement de la recherche dans les domaines de l'apprentissage statistique et du "machine learning" a permis l'émergence de très nombreux modèles prédictifs. Ceux-ci se révèlent particulièrement attractifs pour une utilisation en théorie des sondages, que ce soit pour l'amélioration des plans de sondage (via l'adaptive sampling, par exemple), la construction d'estimateurs plus performants, ou encore le traitement de la non-réponse. Ces modèles et algorithmes prédictifs soulèvent toutefois un dilemme : est-il intéressant, et prudent, d'utiliser des algorithmes parfois extrêmement complexes, dont l'efficacité supposée repose uniquement sur des indices empiriques, sans réelles garanties théoriques ? Il est fréquemment objecté aux modèles d'apprentissage le fait qu'ils se comportent comme des "boites noires", produisant ainsi des résultats que nous ne savons réellement expliquer ni interpréter. Des récentes études empiriques (e.g. Dagdoug et al. (2020a); Larbi et al. (2022)) semblent indiquer la supériorité de modèles très complexes (e.g. Cubist, Quinlan et al. (1992), XGBoost Chen and Guestrin (2016), BART Chipman et al. (2010)) dans de grands nombres de scénarios. Toutefois, à ce jour, il reste encore de nombreux points d'interrogation au regard des mécanismes permettant à ces modèles d'être si efficaces. En théorie des sondages, à notre connaissance, aucune étude théorique n'a permis d'établir les propriétés de ces algorithmes et des estimateurs qui en découlent. Ainsi, utiliser ce genre d'algorithmes en théorie des sondages, c'est aussi prendre le risque de les utiliser dans un scénario qui pourrait leur être mal adapté, e.g. un scénario dans lequel le biais de l'estimateur induit est important, dans lequel l'estimateur de variance traditionnelle sous-estime la variance, etc... Ces scénarios proviennent fréquemment d'interactions, parfois inattendues, entre la théorie des sondages (domaine dans lequel l'algorithme est utilisé) et l'apprentissage statistique (domaine pour lequel l'algorithme a été proposé). Détecter ces phénomènes indésirables appelle par conséquent à la multiplication d'études approfondies (empiriques et théoriques) sur ces méthodologies appliquées en théorie des sondages. Concernant les arbres de régression et forêts aléatoires, les travaux présentés dans cet article ont permis d'établir et de comprendre certains des mécanismes intervenant dans ces algorithmes et estimateurs. Il est néanmoins important de noter que certaines questions relatives à ces méthodologies restent ouvertes. De plus, l'utilisation de ces algorithmes complexes force les utilisateurs à choisir un certain nombre de paramètres, choix parfois très compliqué à faire

en pratique. Dans le cas des forêts aléatoires, si l'algorithme utilisé est celui de Breiman, il est, au minimum, nécessaire de choisir le nombre d'arbres à inclure dans la forêt, le nombre minimal d'éléments dans les feuilles de chaque arbre, ainsi que le nombre de covariables à considérer à chaque split ; le lecteur intéressé peut se référer à Dagdoug et al. (2021) pour une discussion sur le choix de ces paramètres. En revanche, pour des algorithmes plus complexes, comme le boosting, Cubist, les réseaux de neurones ou BART, choisir les hyper-paramètres des modèles peut s'avérer très compliqué. Plus généralement, choisir quel estimateur utiliser en pratique est une question délicate. Dans de nombreux domaines des statistiques, des procédures et méthodologies ont été mises en place pour répondre à cette problématique ; par exemple, la validation croisée, les critères d'information (AIC, BIC, etc...) pour l'apprentissage statistique, la méthodologie de Box et Jenkins en séries temporelles, pour n'en citer que quelques-uns. Ce genre de méthodologie semble être un axe de recherche majeur pour l'avenir et est actuellement en cours d'investigation.

Références

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9 :2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2) :197–227.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92 :831–846.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28 :1023–1053.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2) :123–140.

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

Brewer, K. and Gregoire, T. G. (2009). Introduction to survey sampling. In *Handbook of Statistics*, volume 29, pages 9–37. Elsevier.

Cassel, C., Särndal, C., and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63 :615–620.

Chauvet, G. and Goga, C. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *Journal of Statistical Planning and Inference*, 217 :177–187.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics*, 16(2) :113.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453) :260–269.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys : A critical review. *International Statistical Review*, 87 :S192–S218.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.

Chipman, H., George, E., and McCulloch, R. (2010). BART : Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1) :266–298.

Dagdoug, M., Goga, C., and Haziza, D. (2020a). Imputation procedures in surveys using nonparametric and machine learning methods : an empirical comparison. *arXiv preprint arXiv :2007.06298*.

Dagdoug, M., Goga, C., and Haziza, D. (2020b). Model-assisted estimation in high-dimensional settings for survey data. *arXiv preprint arXiv :2012.07385*.

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.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Regression tree and random forest imputation in surveys with application to data integration. unpublished.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 :376–382.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1) :10–19.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche non paramétrique par splines de régression. *The Canadian Journal of Statistics*, 33 :163–180.

Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society : Series B*, 76 :113–140.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on barro colorado island — digital soil mapping using random forests analysis. *Geoderma*, 146(1-2) :102–113.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning : Data Mining, Inference and Prediction*. Springer, New York.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1) :25–43.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data : a critical review. *Japanese Journal of Statistics and Data Science*, 3(2) :583–623.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685.

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

Kane, M., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, 15(1).

Klusowski, J. M. (2021). Universal consistency of decision trees in high dimensions.

Larbi, K., Haziza, D., and Dagdoug, M. (2022). Treatment of unit nonresponse in surveys through machine learning methods : an empirical comparison. unpublished.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, pages 139–157.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Regression*, 25 :745–763.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46 :389–413.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *Journal of the American Statistical Association*, 100 :1429–1442.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3) :1084–1105.

Opsomer, J. D., Breidt, F. J., Moisen, G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, (478) :400–409.

Qi, Y. (2012). *Random forests for bioinformatics*, pages 307–323. Springer.

Quinlan, J. et al. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific.

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā Series B*, 45 :240–248.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

Särndal, C.-E. (1980). On the π-inverse weighting best linear unbiased weighting in probability sampling. *Biometrika*, 67 :639–650.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18 :241–252.

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

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146 :72–83.

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

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445) :254–265.

Stekhoven, D. J. and Buhlmann, P. (2011). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112–118.

Ta, T., Shao, J., Li, Q., and Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*, 30(3) :1135.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106 :1626–1636.

Wright, M. and Ziegler, A. (2015). ranger : A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv :1508.04409*.

Yang, S. and Kim, J. K. (2017). Predictive mean matching imputation in survey sampling. *arXiv preprint arXiv :1703.10256*.

Yang, S. and Kim, J. K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In *The Econometrics of Complex Survey Data*. Emerald Publishing Limited.

Zhong, P.-S. and Chen, S. (2014). Jackknife empirical likelihood inference with regression imputation and survey data. *Journal of Multivariate Analysis*, 129 :193–205.