Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Approche CHAID pour les enquêtes : un outil pour construire des cellules d’ajustement pour la non-réponse dans un cadre fondé sur le plan

par Jean D. Opsomer et Minsun K. Riddles

Date de diffusion : le 30 juin 2025

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2025
Vol. 51, nº 1, pp. 235-252
Statistique Canada, nº 12-001-X au catalogue

Approche CHAID pour les enquêtes : un outil pour construire des cellules d'ajustement pour la non-réponse dans un cadre fondé sur le plan

Jean D. Opsomer et Minsun K. Riddles[1]

Résumé

Les spécialistes des enquêtes ont de plus en plus adopté les avantages des techniques modernes d'apprentissage automatique, y compris les algorithmes arborescents de classification et de régression, dans la création d'ajustements pour la non-réponse. Ces méthodes, lesquelles ne nécessitent pas de relation fonctionnelle prédéfinie entre les résultats et les prédicteurs, offrent un moyen pratique de procéder à la sélection des variables et d'établir des structures interprétables qui lient la propension à répondre aux variables explicatives. Toutefois, lors de l'application de ces algorithmes aux données d'enquête, il arrive couramment que l'on néglige des facteurs décisifs, comme les poids d'échantillonnage, ainsi que les caractéristiques du plan d'échantillonnage, telles que la stratification et le partitionnement. Pour combler cette lacune, nous proposons une extension de l'approche de la détection automatique d'interactions du khi carré (CHAID) et nous décrivons les propriétés asymptotiques fondées sur le plan de la méthode « CHAID pour les enquêtes » qui en résulte. Pour faciliter son utilisation pratique, nous intégrons une correction de Rao-Scott dans le critère de fractionnement pour tenir compte du plan d'enquête. À l'aide de données de l'American Community Survey, nous illustrons l'utilisation de la méthode et évaluons son rendement en le comparant à celui d'algorithmes pondérés et non pondérés existants.

Mots-clés : Partitionnement récursif; propension à répondre; test du khi carré.

1. Introduction

La prise en compte de la non-réponse des unités est une composante cruciale de l'estimation d'enquête et une étape clé dans la création de poids d'enquête. Les approches les plus couramment utilisées reposent sur le fait de postuler un modèle de propension à répondre et de créer des poids qui sont des inverses des propensions à répondre estimées. Les modèles de propension peuvent être énoncés explicitement ou seulement implicitement, par la spécification de covariables qui sont considérées comme liées à la propension à répondre. Lorsque les modèles sont explicites, ils peuvent être soit paramétriques (le modèle de régression logistique est un choix commun), soit non paramétriques, et les propensions à répondre estimées sont obtenues sous forme de prédictions du modèle de régression. L'approche implicite ne précise pas de modèle et s'appuie plutôt sur des méthodes de calage pour obtenir des poids qui tiennent compte du plan, du mécanisme de réponse et des connaissances au niveau de la population dans un ajustement combiné. Comme l'objectif est presque toujours de créer des poids qui peuvent être appliqués à toutes les variables de l'enquête, toutes ces approches reposent sur l'hypothèse de répartition au hasard des données manquantes. Veuillez consulter Brick (2013) pour obtenir un aperçu des différents cadres permettant d'obtenir des ajustements des poids de la propension à répondre.

La division de l'échantillon en cellules de pondération et l'application d'un ajustement par la propension égale à l'intérieur de chaque cellule représentent une méthode courante pour obtenir des ajustements des

1. Jean D. Opsomer et Minsun K. Riddles, Westat, Inc., 1600 Research Blvd., Rockville (Maryland), 14850, États-Unis. Courriel : jeanopsomer@westat.com.

poids. Cette méthode peut être justifiée dans le cadre de multiples modèles explicites ou comme stratification *a posteriori* dans le cadre de l'approche implicite de calage assisté par un modèle. Le modèle du *groupe de réponses homogènes* (GRH) (voir Särndal, Swensson et Wretman, 1992, chapitre 15.6) suppose que le modèle de propension sous-jacent est constitué de cellules dans lesquelles les unités d'échantillonnage répondent à l'enquête selon une probabilité égale (mais inconnue). Les cellules peuvent être présumées connues, par exemple à l'aide de strates démographiques *a posteriori* sous forme de cellules du GRH, ou peuvent être déterminées en fonction de l'échantillon. Cette dernière approche est intéressante si le nombre de covariables potentielles selon lesquelles on définit les cellules de pondération est important par rapport à la taille de l'échantillon. Comme nous le préciserons davantage dans les sections suivantes, nous supposons dans la présente étude que les réponses suivent le modèle du GRH; les cellules d'ajustement de la pondération seront sélectionnées par l'algorithme de détection automatique d'interactions du khi carré (CHAID), un type populaire de partitionnement récursif. D'autres auteurs se sont appuyés sur le modèle du GRH et ont proposé d'autres méthodes de partitionnement récursif pour créer les cellules d'ajustement; voir, par exemple, Phipps et Toth (2012).

Comme solution de rechange au modèle du GRH, une forme fonctionnelle paramétrique telle qu'une régression logistique peut être utilisée pour le modèle de propension à répondre. Toutefois, même dans ce cas, les ajustements des poids sont souvent appliqués comme des ajustements égaux au sein des cellules (Little, 1986). Cela est mis en œuvre en divisant les propensions prédites par le modèle en cellules et en faisant la moyenne des prédictions au sein de chaque cellule pour obtenir un seul ajustement des poids. Étant donné que ces prévisions moyennes permettent maintenant d'estimer une approximation de la fonction de la réponse réelle, contrairement au modèle du GRH, il faut régler les problèmes de biais d'approximation liés au nombre de cellules et à leur taille lors de la mise en œuvre de cette approche. L'utilisation de quantiles égaux des propensions prédites pour définir les cellules est une approche courante; voir, par exemple, Eltinge et Yansaneh (1997).

L'algorithme CHAID est un algorithme de partitionnement récursif proposé à l'origine par Kass (1980). Il a été utilisé dans une grande variété de domaines comme méthode de classification et d'exploration de données. Dans le contexte des enquêtes, lequel nous intéresse dans la présente étude, il est souvent utilisé pour la construction des cellules de pondération des ajustements pour la non-réponse. L'approche habituelle consiste à adapter un arbre de classification à l'ensemble de données non pondérées à l'aide de la méthode CHAID; les nœuds de l'arbre résultant définissent les cellules de pondération. Par la suite, les ajustements de pondération pour la non-réponse sont calculés comme des inverses des proportions de réponse pondérées dans chaque cellule. Par conséquent, cette méthode de construction de l'ajustement pour la non-réponse est mise en œuvre comme une méthode hybride des approches non pondérées et pondérées. Bien que cette méthode fonctionne généralement bien, il est à craindre que l'omission du plan puisse mener à une détermination inexacte des cellules de pondération. En outre, sa nature hybride la rend difficile à étudier ou à justifier sur le plan théorique. Les objectifs du présent article sont d'introduire la méthode CHAID pour les enquêtes (sCHAID en anglais), une version entièrement basée sur le plan de la méthode CHAID, et d'étudier ses propriétés asymptotiques et pratiques.

La littérature sur le partitionnement récursif dans un contexte fondé sur le plan est limitée. La méthodologie du partitionnement récursif la plus comparable qui tient compte du plan d'échantillonnage est celle proposée par Toth et Eltinge (2011), qui repose sur l'approche de Gordon et Ohlshen (1980) et est mise en œuvre dans le paquet rpms dans R (Toth, 2017). Une autre approche récente est celle de Beaumont, Bosa, Brennan, Charlebois et Chu (2024), qui proposent une méthode de partitionnement récursif fondé sur le plan reposant sur les arbres de classification et de régression de Breiman, Friedman, Olshen et Stone (1984) pour créer des cellules de probabilité de sélection pour la pondération des données d'échantillons non probabilistes. Nous comparerons notre approche à celle de Toth et Eltinge (2011) dans les simulations.

La structure de la suite de l'article est la suivante. À la section 2, nous présentons le modèle de propension à répondre et l'approche d'estimation proposée. La section 3 décrit les propriétés asymptotiques basées sur le plan de l'approche CHAID pour les enquêtes. À la section 4, les résultats d'une étude par simulation comparant la méthode CHAID pour les enquêtes et la méthode CHAID sont présentés.

2. Notation et description de la méthode CHAID pour les enquêtes

Supposons que I_i représente l'indicateur d'appartenance à l'échantillon pour $i \in U$, où U désigne la population de taille N. L'échantillon $s \subset U$ est sélectionné selon le plan d'échantillonnage $p(s)$, qui détermine la répartition de I_i et, en particulier, les probabilités d'inclusion $\pi_i = \mathrm{E}(I_i), \pi_{ij} = \mathrm{E}(I_i I_j), i, j \in U$. Les indicateurs de réponse R_i sont des variables aléatoires de Bernoulli indépendantes où $\mathrm{E}(R_i) = p_i, i \in U$. Nous permettons à p_i de dépendre du plan d'échantillonnage, mais pas de l'échantillon réalisé s, ce qui est un scénario commun pour la modélisation de la propension à répondre (voir, par exemple, Shao et Steel, 1999). Bien que l'indépendance des indicateurs de réponse soit généralement présumée dans l'élaboration des ajustements pour la non-réponse, elle n'est pas toujours justifiée, par exemple lorsqu'il y a un effet notable de l'intervieweur sur la non-réponse au niveau de l'unité dans les enquêtes par visite en personne ou par téléphone. Par souci de simplicité, nous continuerons de limiter notre attention au cas indépendant.

Dans le modèle du GRH,

$$p_i = \sum_{g=1}^{G} I_{\{i \in U_g^*\}} P_g^*, \tag{2.1}$$

où $I_{\mathcal{D}} = 1$ si l'événement $\mathcal{D}$ est vrai et égal à 0 autrement, et les groupes de réponses homogènes qui ne se chevauchent pas U_g^* de taille N_g^* sont définis par les intersections de K variables auxiliaires catégoriques (ou « catégorisées » si elles étaient initialement continues) $X_k, k = 1, \ldots, K$. La partie de l'échantillon s qui se trouve dans U_g^* est désignée par $s_g^* = s \cap U_g^*$.

Si G est modeste par rapport à la taille de l'échantillon n, il serait raisonnable d'utiliser simplement les groupes définis par les intersections des variables auxiliaires comme cellules d'ajustement. Nous nous intéressons au scénario dans lequel le nombre de cellules obtenues de cette façon est trop important pour être pratique, de sorte que choisir un plus petit nombre de cellules est intéressant pour améliorer la stabilité des poids d'enquête. Nous ferons appel à la méthode CHAID pour les enquêtes à cette fin. L'algorithme

sera fondé sur une séquence de « fractionnements de l'échantillon » selon les valeurs des variables catégoriques définies sur les unités de population, donc nous définissons d'abord la notation pour les sous-ensembles pertinents.

Soit $X_{ki}, k = 1,\ldots,K,$ les valeurs des variables catégoriques pour l'unité d'échantillonnage $i,$ prenant les valeurs $1,\ldots,L_k$ (possiblement après la conversion des catégories en nombres entiers). Pour simplifier la notation, nous supposerons que $L_k = L$ pour toutes les $k,$ mais le cas comportant des nombres inégaux de valeurs est facilement généralisé. Pour chaque variable $X_k,$ nous définissons les sous-ensembles $U^*_{kl}, l = 1,\ldots,L$ de taille N^*_{kl} correspondant à l'ensemble des unités où $X_{ki} = l.$ Les U^*_{kl} ne se chevauchent pas pour une k fixe. Les $G = L^K$ groupes de réponses homogènes U^*_g sont obtenus comme des intersections des sous-ensembles U^*_{kl} pour $k = 1,\ldots,K$ et $l = 1,\ldots,L.$ Nous écrivons A^*_{kl} pour l'ensemble des indices $g,$ de sorte que $U^*_g \subseteq U^*_{kl},$ et nous notons que $\sum_{g\in A^*_{kl}} N^*_g = N^*_{kl}.$ En d'autres termes, A^*_{kl} définit une « portion » parmi les groupes de réponses homogènes correspondant à la valeur l de la variable catégorique $X_k.$ De manière analogue, nous écrivons s^*_{kl} pour la partie de s qui se trouve dans $U^*_{kl}.$

Les fractionnements de l'échantillon sont choisis en comparant la propension à répondre moyenne dans les sous-ensembles de l'échantillon définis par les variables auxiliaires. Comme les propensions réelles sont inconnues, cela est mis en œuvre dans la méthode CHAID pour les enquêtes (et la méthode CHAID) par la réalisation de tests statistiques fondés sur les propensions à répondre estimées (voir ci-dessous). Si X_k est une variable ordinale (ou une variable continue catégorisée), nous souhaitons comparer la propension moyenne dans U^*_{kl} où l est inférieure ou égale à une catégorie cible, avec la propension moyenne dans U^*_{kl} où l est supérieure à la catégorie cible. En d'autres termes, un fractionnement à l pour une variable ordinale X_k définira deux sous-ensembles

$$U_{kl} = \cup_{l'\le l} U^*_{kl'} \quad \text{et} \quad U_{lk^c} = \cup_{l'>l} U^*_{kl'}.$$

De même, pour une variable purement catégorique, un fractionnement à l définit deux sous-ensembles

$$U_{kl} = U^*_{kl} \quad \text{et} \quad U_{kl^c} = \cup_{l'\ne l} U^*_{kl'}.$$

Dans ce qui suit, nous ignorerons la distinction entre les types de variables auxiliaires et utiliserons la notation (U_{kl}, U_{kl^c}) pour aborder les sous-ensembles de population évalués. De manière analogue à ce qui précède, nous définissons les ensembles d'indices $A_{kl}, A_{kl^c},$ les tailles de sous-ensembles de population N_{kl}, N_{kl^c} et les échantillons s_{kl}, s_{kl^c} qui correspondent à $U_{kl}, U_{kl^c},$ respectivement.

Pour clarifier cette structure de population, le tableau 2.1 montre un exemple où il y a $K = 2$ variables ordinales ayant chacune $L = 5$ catégories. Dans ce cas, la population U est divisée en $G = 25$ sous-ensembles U^*_g et chaque g correspond à une combinaison unique de valeurs pour $(X_1, X_2),$ mais est arbitraire autrement. On peut définir les sous-ensembles U^*_{1l} comme les colonnes de la grille du tableau 2.1 ou, à titre équivalent, comme un ensemble de 5 indices $A^*_{1l} \subset \{1,\ldots,25\}$ et, d'une manière similaire, on peut définir les sous-ensembles U^*_{2l} comme les rangées de la grille avec l'ensemble d'indices A^*_{2l} correspondant. Par exemple, $A^*_{12} = \{2,7,12,17,22\}$ est l'ensemble d'indices pour $U^*_{12} = \{i : X_{1i} = 2\}.$ Enfin, on peut définir

un fractionnement de l'échantillon pour X_1 à l par les sous-ensembles $U_{1l} = \{i : X_{1i} \le l\}$ et $U_{1l^c} = \{i : X_{1i} > l\}$, définir les ensembles d'indices A_{1l}, A_{1l^c} par les unions du A_{1l}^* correspondant et faire de même pour un fractionnement de l'échantillon sur X_2. Par exemple, le fractionnement pour X_2 à $l = 2$ se traduit par les sous-ensembles U_{22} où $A_{22} = \{1,\ldots,10\}$ et U_{22^c} où $A_{22^c} = \{11,\ldots,25\}$.

Tableau 2.1
Exemple de la structure de la population du groupe de réponses homogènes pour $K = 2, L = 5$

	(k,l) =	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	
$X_2 \le 2$	(2,1)	1	2	3	4	5	U_{22}
	(2,2)	6	7	8	9	10	
$X_2 > 2$	(2,3)	11	12	13	14	15	U_{22}^C
	(2,4)	16	17	18	19	20	
	(2,5)	21	22	23	24	25	

En tenant compte d'un U_{kl} particulier, nous définissons les propensions à répondre moyennes

$$P_{kl} = \frac{\sum_{i \in U_{kl}} p_i}{N_{kl}}, \tag{2.2}$$

qui peuvent également être écrites comme une moyenne pondérée de $P_g^*, g \in A_{kl}$,

$$P_{kl} = \frac{\sum_{g \in A_{kl}} N_g^* P_g^*}{N_{kl}}.$$

Dans l'étude des propriétés asymptotiques de l'approche CHAID pour les enquêtes (voir la section 3), N augmentera à l'infini. Même si les quantités de population telles que P_{kl} dépendent donc de N, nous ne l'écrirons pas explicitement.

Nous définissons l'estimateur de Horvitz-Thompson (HT) de P_g^*, la probabilité de réponse dans U_g^*, de la façon suivante :

$$\tilde{R}_g^* = \frac{\sum_{s_g^*} \frac{1}{\pi_i} R_i}{N_g^*}. \tag{2.3}$$

Nous désignons son espérance conditionnelle par rapport au plan d'échantillonnage comme suit :

$$R_g^* = \frac{\sum_{U_g^*} R_i}{N_g^*}.$$

L'estimateur de la probabilité de réponse moyenne dans un sous-ensemble U_{kl} peut être défini comme étant soit du type HT,

$$\tilde{P}_{kl} = \frac{\sum_{s_{kl}} \frac{1}{\pi_i} R_i}{N_{kl}} = \frac{\sum_{g \in A_{kl}} N_g^* \tilde{R}_g^*}{N_{kl}},$$

soit du type Hájek (HA),

$$\hat{P}_{kl} = \frac{\sum_{s_{kl}} \frac{1}{\pi_i} R_i}{\widetilde{N}_{kl}} = \frac{\sum_{g \in A_{kl}} N_g^* \widetilde{R}_g^*}{\widetilde{N}_{kl}}, \tag{2.4}$$

où $\widetilde{N}_{kl} = \sum_{g \in A_{kl}} \widetilde{N}_g^*$ et $\widetilde{N}_g^* = \sum_{s_g^*} 1/\pi_i$. Les deux sont des fonctions de $\widetilde{R}_g^*$ dans (2.3), ce qui sera utile lorsque nous étudierons les propriétés théoriques des estimateurs, mais que nous n'utilisons pas explicitement lors du calcul des estimations. Nous allons mettre l'accent sur l'estimateur de HA dans (2.4) pour la sélection des fractionnements de l'approche CHAID pour les enquêtes, car il ne nécessite pas de connaître N_{kl} et est généralement considéré comme plus efficace que l'estimateur de HT.

Alors que l'approche CHAID repose sur un test du χ^2 pour les observations indépendantes et identiquement distribuées afin de déterminer si P_{kl} et P_{kl^c} sont différentes, nous remplaçons ce test par un test de Wald où l'on utilise $\hat{P}_{kl}$ et $\hat{P}_{kl^c}$ pour que le plan d'échantillonnage puisse être pris en compte. Dans ce cas-ci, nous ne considérons que le test de Wald dans l'enquête théorique, mais, dans la pratique, il peut être remplacé par une des approximations de Rao et Scott (1981), selon leur mise en œuvre dans plusieurs paquets pour les enquêtes. La variable à tester est

$$\widehat{W}_{kl} = n \frac{(\hat{P}_{kl} - \hat{P}_{kl^c})^2}{\hat{V}_{kl}}, \tag{2.5}$$

où $\hat{V}_{kl}$ est un estimateur de variance fondé sur le plan mis à l'échelle à préciser dans la section suivante. La valeur-*p* (en anglais *p*-value pour probability value) pour le test est

$$\hat{q}_{kl} = \Pr(\chi_1^2 > \widehat{W}_{kl}). \tag{2.6}$$

Dans l'approche CHAID pour les enquêtes, la variable à tester (2.5) sera calculée non seulement sur deux groupes au sein de l'échantillon global, mais également au sein des sous-échantillons obtenus lors d'itérations antérieures de l'algorithme CHAID pour les enquêtes. Disons que s^0 représente un tel sous-échantillon. Dans ce cas, on comprend que les groupes de l'échantillon s_{kl} et leur complément s_{kl^c} ne sont compris que dans s^0, et la variable à tester n'est également calculée que sur s^0. Par souci de simplicité, nous ne préciserons pas quels sous-échantillons sont utilisés et continuerons à nous référer à s_{kl} et à s_{kl^c}.

L'algorithme CHAID construit l'arbre de classification en effectuant un ensemble de tests statistiques à chaque étape pour sélectionner et fractionner un des nœuds de l'arbre disponibles en nœuds plus petits. Il existe de nombreuses variantes de l'algorithme, selon les niveaux de signification utilisés pour déterminer les fractionnements, selon les fractionnements admissibles, selon les critères d'arrêt de l'algorithme, etc. La plupart d'entre elles peuvent également être mises en œuvre dans le contexte de l'algorithme CHAID pour les enquêtes. Toutefois, afin de pouvoir étudier la méthode proposée de manière asymptotique, nous considérons dans ce cas-ci une version simplifiée de l'algorithme. Plus précisément, l'algorithme CHAID pour les enquêtes proposé permet de procéder à la division de l'échantillon en sous-échantillons (« nœuds ») comme suit :

Étape 1 : L'échantillon est constitué d'un seul nœud.

Étape r : Le nombre de nœuds augmente de $r-1$ à r en effectuant un fractionnement binaire sur l'un des nœuds existants; les fractionnements respectent la nature ordinale ou catégorique des variables auxiliaires. Les fractionnements admissibles à l'étape r sont tous les fractionnements qui entraînent d'autres divisions au sein des nœuds, y compris les fractionnements sur les variables déjà utilisées aux étapes précédentes. Les valeurs-p des tests de Wald $\widehat{W}_{kl}$ sur tous les fractionnements admissibles sont calculées, et seuls ceux dont les valeurs-p sont inférieures à une valeur prédéterminée α sont pris en compte par la suite. Parmi ceux-ci, celui ayant la plus petite valeurs-p est sélectionné pour un fractionnement. Le fractionnement consiste à classer les unités d'échantillonnage des nœuds en deux sous-nœuds, selon qu'elles appartiennent à U_{kl} ou à U_{kl^c} pour le test sélectionné. Si aucune des valeurs-p n'est inférieure à α, l'algorithme s'arrête.

Étape R : L'algorithme s'arrête lorsque R nœuds ont été créés, à moins qu'il ne se soit arrêté plus tôt en raison du manque de fractionnements admissibles.

Le résultat de l'algorithme CHAID pour les enquêtes est un ensemble d'un maximum de R nœuds, qui sont déterminés de manière unique par la séquence des fractionnements. À leur tour, ces fractionnements sont déterminés par les valeurs des variables auxiliaires qui correspondent à $\min_{kl} \hat{q}_{kl}$ (ou, de manière équivalente, à $\max_{kl} \widehat{W}_{kl}$) à chaque étape, où il est entendu que les fractionnements aux étapes ultérieures sont conditionnels à ceux effectués précédemment. Nous utiliserons la notation $\widehat{\mathcal{T}}_s$ pour désigner le résultat de l'algorithme CHAID pour les enquêtes pour l'échantillon s, exprimable sous l'une ou l'autre de ces formes, et nous élaborerons ses propriétés théoriques à la section suivante.

3. Résultats théoriques

Nous étudierons les propriétés de l'algorithme CHAID pour les enquêtes dans un cadre conjoint de conception et de modélisation, en utilisant l'approche inversée (Shao et Steel, 1999). Selon cette approche, les répondants sont d'abord tirés au niveau de la population comme des réalisations de N variables aléatoires de Bernoulli indépendantes avec une espérance $p_i, i=1,\ldots,N,$ comme cela est défini dans (2.1). Ensuite, l'échantillon s est tiré à partir de U selon le plan d'échantillonnage indiqué $p(s),$ indépendamment de l'état réalisé de la réponse de la population. La population finie est intégrée dans une séquence indexée par $N,$ et nous supposons que $N \to \infty.$ Le plan d'échantillonnage $p(s)$ dépend de la population et change donc également selon N.

Dans les hypothèses suivantes, nous énonçons des conditions suffisantes pour démontrer les résultats théoriques de la méthode CHAID pour les enquêtes proposée.

A.1. *Le nombre de groupes de réponses homogènes* G *est fixe et il existe* $\lambda_1 > 0$ *de sorte que* $N_g^* / N > \lambda_1$ *pour* $g=1,\ldots,G.$ *Il existe* $0 < \lambda_2 < \lambda_3 < 1$ *de sorte que* $\lambda_2 < P_g^* < \lambda_3$ *pour* $g=1,\ldots,G.$

A.2. *Pour toute combinaison des groupes* A_{kl}, A_{kl^c} *qui peuvent figurer parmi ceux évalués par l'algorithme CHAID pour les enquêtes, il n'y a pas deux valeurs* $(P_{kl} - P_{kl^c})^2 / V_{kl}$ *égales sauf si* $P_{kl} - P_{kl^c} = 0$. *Le nombre de fractionnements admissibles dans l'ensemble de la population est d'au moins* R.

A.3. *Le plan d'échantillonnage* $p(s)$ *est tel qu'il existe un nombre* N_0 *de sorte que pour tout* $N > N_0$, $n_g^* \geq 1$ *pour* $g = 1,\ldots,G$ *avec une probabilité de 1. La taille de l'échantillon* n *est non aléatoire et satisfait* $n / N \to \lambda_4$, *où* $\lambda_4 \in (0,1)$.

A.4. *Pour toute variable non aléatoire bornée* z_i, *le vecteur des estimateurs de HT* $\tilde{\mathbf{Z}}^* = (\tilde{z}_1^*, \ldots, \tilde{z}_G^*)^T$, *où* $\tilde{z}_g^* = \sum_{s_g^*} \frac{z_i}{\pi_i} \Big/ N_g^*$, *présente la distribution asymptotique suivante :*

$$\sqrt{n}\left(\tilde{\mathbf{Z}}^* - \mathrm{E}(\tilde{\mathbf{Z}}^*)\right) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{V}_z^*)$$

par rapport à la séquence des plans d'échantillonnage, pour une matrice $\mathbf{V}_z^*$.

A.5. *Les probabilités d'inclusion satisfont* $\pi_i \geq \lambda_5 > 0$, $\pi_{ij} \geq \lambda_6 > 0$ *pour tous les couples* $i, j \in U$ *pour certaines constantes* λ_5, λ_6. *En outre,*

$$\lim_{N\to\infty} n \max_{i,j\in U;\, i\neq j} |\Delta_{ij}| < \infty$$

avec $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, *et*

$$\lim_{N\to\infty} \max_{i,j,i',j'\in U;\, i\neq j\neq i'\neq j'} E\left((I_i I_j - \pi_{ij})(I_{i'} I_{j'} - \pi_{i'j'})\right) = 0.$$

L'hypothèse 1 assure que le modèle du GRH dans (2.1) reste valide et fixe dans la séquence des populations et que la probabilité de réponse est bornée loin de 0 et de 1 partout. La première hypothèse est faite pour simplifier la théorie et pourrait être assouplie en permettant à G de croître selon N à un rythme convenablement lent. La dernière hypothèse est faite pour éviter les composantes de variance dégénérées. L'hypothèse 2 est formulée pour éviter les égalités entre les sélections de fractionnement, qui pourraient être traitées au prix d'une notation et de résultats plus compliqués, et pour s'assurer que la probabilité que l'algorithme CHAID pour les enquêtes se termine avant R étapes en raison du manque de fractionnements admissibles passe à zéro. L'hypothèse 3 précise les conditions de régularité sur le plan d'échantillonnage, ce qui empêche l'apparition de domaines vides pour un N suffisamment grand et permet de s'assurer que la fraction d'échantillonnage ne devient pas négligeable. L'hypothèse 4 stipule que le plan d'échantillonnage comporte un théorème de la limite centrale pour les estimateurs de HT. Il s'agit d'une hypothèse courante dans les investigations asymptotiques sur les estimateurs parce que certains plans d'échantillonnage ont des conditions différentes pour assurer la normalité de leurs estimateurs associés, tandis que pour d'autres, la normalité asymptotique est utilisée pour l'inférence dans la pratique, mais n'a pas été officiellement établie dans la littérature. Enfin, les limites et les conditions sur les taux pour les probabilités d'inclusion de différents ordres dans l'hypothèse 5 sont établies pour assurer l'existence de l'estimateur de HT et de l'estimateur non biaisé de sa variance selon le plan d'échantillonnage, qui est de l'ordre $O(n^{-1})$ et peut être estimée de manière constante.

Lemme 1. *Dans les hypothèses 1 à 5, le vecteur* $\widetilde{\mathbf{R}}^* = (\widetilde{R}_1^*,\ldots,\widetilde{R}_G^*)^T$ *présente la distribution asymptotique suivante en ce qui concerne le mécanisme de réponse et la séquence des plans d'échantillonnage :*

$$\sqrt{n}\,(\widetilde{\mathbf{R}}^* - \mathbf{P}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{V}_R^*), \tag{3.1}$$

où $\mathbf{P}^* = (P_1^*,\ldots,P_G^*)^T$ *et* $\mathbf{V}_R^*$, *la matrice comportant l'élément* gg', *est égale à*

$$V_{R,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} \Delta_{ij} \frac{p_i}{\pi_i} \frac{p_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N_g^{*2}} \sum_{U_g^*} \frac{1}{\pi_i} p_i (1-p_i).$$

L'estimateur

$$\widetilde{V}_{R,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{s_g^*} \sum_{s_{g'}^*} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N_g^{*2}} \widetilde{N}_g^* \widetilde{R}_g^* (1-\widetilde{R}_g^*) \tag{3.2}$$

est constant pour $V_{R,gg'}^*$ *en ce qui concerne le mécanisme de réponse et la séquence des plans d'échantillonnage.*

Démonstration du lemme 1 : Nous appliquons le théorème 1.3.6 de Fuller (2009) pour obtenir la distribution asymptotique selon la combinaison du mécanisme de réponse et du plan d'échantillonnage. Nous considérons d'abord la moyenne de la population $R_g^* = \sum_{U_g^*} R_i / N_g^*$, l'espérance conditionnelle de $\widetilde{R}_g^*$ par rapport au plan d'échantillonnage. Parce que R_g^* est une moyenne de N_g^* variables de Bernoulli indépendantes et identiquement distribuées, elle satisfait à une condition de Liapounov (par exemple Fuller, 1996, théorème 5.3.2) et est distribuée de manière asymptotiquement normale. En supposant que $\mathcal{R}_N = (R_1,\ldots,R_N)^T$, nous savons que $\mathrm{E}(\widetilde{R}_g^* \mid \mathcal{R}_N) = R_g^*$ et que $\widetilde{R}_g^*$ présente une distribution normale asymptotique conditionnelle, selon l'hypothèse 4, par rapport à la séquence des plans d'échantillonnage, qui se vérifie pour tout $\mathcal{R}_N$. Nous pouvons donc combiner les distributions normales dans la distribution asymptotique globale donnée dans l'énoncé du lemme. La variance asymptotique est obtenue par des calculs standard des moments conditionnels.

Pour montrer que l'estimateur proposé pour les éléments de $\mathbf{V}_R^*$ est cohérent, nous considérons d'abord le premier terme,

$$\widetilde{V}_{R1,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{s_g^*} \sum_{s_{g'}^*} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j}. \tag{3.3}$$

En posant la condition sur la R_i réalisée dans la population, nous trouvons

$$\mathrm{E}(\widetilde{V}_{R1,gg'}^* \mid \mathcal{R}_N) = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} \Delta_{ij} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j}$$

et

$$\mathrm{Var}(\widetilde{V}_{R1,gg'}^* \mid \mathcal{R}_N) = \frac{n^2}{N_g^{*2} N_{g'}^{*2}} \sum_{U_g^*} \sum_{U_g^*} \sum_{U_{g'}^*} \sum_{U_{g'}^*} \Delta_{ij}\, \Delta_{i'j'} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j} \frac{R_{i'}}{\pi_{i'}} \frac{R_{j'}}{\pi_{j'}} \times \mathrm{E}\Big((I_i I_j - \pi_{ij})(I_{i'} I_{j'} - \pi_{i'j'})\Big).$$

Par conséquent,

$$\mathrm{E}(\widetilde{V}^*_{R1,gg'}) \;=\; \frac{n}{N^*_g N^*_{g'}} \sum_{U^*_g} \sum_{U^*_{g'}} (\pi_{ij} - \pi_i \pi_j) \frac{p_i}{\pi_i} \frac{p_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N^{*2}_g} \sum_{U^*_g} \left(\frac{1}{\pi_i} - 1 \right) p_i \,(1 - p_i).$$

Pour $\mathrm{Var}(\widetilde{V}^*_{R1,gg'})$, nous notons que $\mathrm{Var}(\widetilde{V}^*_{R1,gg'} \mid \mathcal{R}_N) \to 0$ uniformément en appliquant les bornes sur les probabilités d'inclusion des différents ordres dans l'hypothèse 5 et en utilisant $R_i \le 1$. Par calcul direct, nous trouvons également

$$\mathrm{Var}(\mathrm{E}(\widetilde{V}^*_{R1,gg'} \mid \mathcal{R}_N)) \;=\; \frac{n^2}{N^{*2}_g N^{*2}_{g'}} \sum_{U^*_g} \sum_{U^*_g} \sum_{U^*_{g'}} \sum_{U^*_{g'}} \Delta_{ij}\, \Delta_{i'j'} \frac{\mathrm{Cov}(R_i R_j, R_{i'} R_{j'})}{\pi_i \pi_j \pi_{i'} \pi_{j'}},$$

qui converge à 0 encore par les bornes de l'hypothèse 5 et le fait que $\mathrm{Cov}(R_i R_j, R_k R_l) = 0$ à moins que (i, j) ne chevauche (i', j'). Par conséquent, $\mathrm{Var}(\widetilde{V}^*_{R1,gg'}) \to 0$ et $\widetilde{V}^*_{R1,gg'}$ converge à (3.3).

Pour le deuxième terme de $\widetilde{V}^*_{R,gg'}$, nous pouvons montrer directement que

$$\frac{n}{N^{*2}_g} \widetilde{N}^*_g \widetilde{R}^*_g (1 - \widetilde{R}^*_g) \;=\; \frac{n}{N^*_g} P^*_g (1 - P^*_g) \;+\; o_p(1) \tag{3.4}$$

par linéarisation. En combinant cela avec (3.3), on obtient le résultat suivant.

Lemme 2. *Dans les hypothèses 1 à 5, la variable à tester*

$$\widehat{W}_{kl} \;=\; n \frac{(\hat{P}_{kl} - \hat{P}_{kl^c})^2}{\hat{V}_{kl}}$$

où

$$\begin{aligned}
\hat{V}_{kl} \;=\;& \frac{n}{N^2_{kl}} \sum \sum_{s_{kl}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl}}{\pi_i} \frac{R_j - P_{kl}}{\pi_j} \\
&+ \frac{n}{N^2_{kl^c}} \sum \sum_{s_{kl^c}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl^c}}{\pi_i} \frac{R_j - P_{kl^c}}{\pi_j} \\
&- \frac{2n}{N_{kl} N_{kl^c}} \sum \sum_{s_{kl}} \sum_{s_{kl^c}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl}}{\pi_i} \frac{R_j - P_{kl^c}}{\pi_j} \\
&+ \frac{n}{N^2_{kl}} \sum_{g \in A_{kl}} \widetilde{N}^*_g \widetilde{R}^*_g (1 - \widetilde{R}^*_g) \;+\; \frac{n}{N^2_{kl^c}} \sum_{g \in A_{kl^c}} \widetilde{N}^*_g \widetilde{R}^*_g (1 - \widetilde{R}^*_g)
\end{aligned}$$

est distribuée asymptotiquement selon la distribution χ^2_1 *non centrale, avec le paramètre de non-centralité* $n\lambda_{kl} = n(P_{kl} - P_{kl^c})^2 / V_{kl}$ *et*

$$\begin{aligned}
V_{kl} \;=\;& \frac{n}{N^2_{kl}} \sum \sum_{U_{kl}} \Delta_{ij} \frac{p_i - P_{kl}}{\pi_i} \frac{p_j - P_{kl}}{\pi_j} \;+\; \frac{n}{N^2_{kl^c}} \sum \sum_{U_{kl^c}} \Delta_{ij} \frac{p_i - P_{kl^c}}{\pi_i} \frac{p_j - P_{kl^c}}{\pi_j} \\
&- \frac{2n}{N_{kl} N_{kl^c}} \sum \sum_{U_{kl}} \sum_{U_{kl^c}} \Delta_{ij} \frac{p_i - P_{kl}}{\pi_i} \frac{p_j - P_{kl^c}}{\pi_j} \\
&+ \frac{n}{N^2_{kl}} \sum_{U_{kl}} \frac{1}{\pi_i} p_i (1 - p_i) \;+\; \frac{n}{N^2_{kl^c}} \sum_{U_{kl^c}} \frac{1}{\pi_i} p_i (1 - p_i).
\end{aligned}$$

Démonstration du lemme 2 : En utilisant la linéarisation, nous obtenons l'approximation suivante

$$\hat{P}_{kl} - \hat{P}_{kl^c} - (P_{kl} - P_{kl^c}) = \frac{1}{N_{kl}} \sum_{g=1}^{G} (N_g^* \tilde{R}_g^* - P_{kl} \tilde{N}_g^*) (I_{\{g \in A_{kl}\}} - I_{\{g \in A_{kl^c}\}}) + O_p(n^{-1}). \tag{3.5}$$

La normalité asymptotique de $\hat{P}_{kl} - \hat{P}_{kl^c}$ découle directement de l'hypothèse 4 et du lemme 1. À partir de l'approximation obtenue dans (3.5), les calculs simples des moments montrent que la distribution asymptotique de $\hat{P}_{kl} - \hat{P}_{kl^c}$ a une espérance $P_{kl} - P_{kl^c}$ et une variance V_{kl}. Par conséquent, en appliquant le théorème 5.2.4 de Fuller (1996), nous concluons que $n(\hat{P}_{kl} - \hat{P}_{kl^c})^2 / V_{kl}$ a une distribution χ_1^2 asymptotique non centrale, avec un paramètre de non-centralité égal à $n(P_{kl} - P_{kl^c})^2 / V_{kl}$. La distribution de $\widehat{W}_{kl}$ découle du corollaire 5.2.6.1 de Fuller (1996) une fois que nous avons remplacé V_{kl} par un estimateur cohérent. En utilisant la même approche que dans la démonstration du lemme 1, la cohérence de $\hat{V}_{kl}$ pour V_{kl} peut à nouveau être démontrée.

Théorème 1. *Dans les hypothèses 1 à 5, lors de l'évaluation de l'ensemble des fractionnements admissibles à chaque étape de l'algorithme CHAID pour les enquêtes décrit à la section 2, les deux énoncés suivants sont valables :*

(i) Pour toute valeur-p limite prédéterminée $\alpha > 0$, *tous les fractionnements pour lesquels* $P_{kl} - P_{kl^c} \neq 0$ *peuvent être examinés avec une probabilité tendant vers 1.*

(ii) En sélectionnant le fractionnement ayant la plus petite valeur-p $\hat{q}_{kl}$, *le fractionnement unique correspondant à* $\max (P_{kl} - P_{kl^c})^2 / V_{kl}$ *sera sélectionné avec une probabilité tendant vers 1.*

Démonstration du théorème 1 : Pour démontrer (i), nous considérons la variable à tester $\widehat{W}_{kl}$ et la valeur-*p* associée $\hat{q}_{kl}$. Selon les propriétés de la distribution χ^2 non centrale, nous savons que la distribution asymptotique de $\widehat{W}_{kl}$ a une moyenne $\lambda_{kl} + 1 = O(n)$ et une variance $2\lambda_{kl} + 4 = O(n)$, à moins que $P_{kl} - P_{kl^c} = 0$. Par conséquent, $\widehat{W}_{kl} / n$ converge en probabilité vers une constante et $\hat{q}_{kl}$ converge en probabilité vers 0 chaque fois que $P_{kl} - P_{kl^c} \neq 0$, de sorte que $\hat{q}_{kl} < \alpha$ avec une probabilité tendant vers 1.

Pour (ii), nous considérons deux fractionnements possibles avec les variables à tester $\widehat{W}_{kl}, \widehat{W}_{k'l'}$ et leur valeur-*p* associée $\hat{q}_{kl}, \hat{q}_{k'l'}$, avec $P_{kl} - P_{kl^c} \neq 0$ et $P_{k'l'} - P_{k'l'^c} \neq 0$. Le test ayant la plus petite valeur-*p* doit être sélectionné pour le fractionnement, que nous pouvons écrire comme un indicateur $I_{\{\hat{q}_{kl} - \hat{q}_{k'l'} < 0\}}$; si cet indicateur est de 1, alors le fractionnement sur $X_k = l$ est sélectionné et si cet indicateur est de 0, le fractionnement sur $X_{k'} = l'$ est sélectionné. Encore, selon les moments de la distribution χ^2 non centrale, nous avons

$$\begin{aligned} I_{\{\hat{q}_{kl} - \hat{q}_{k'l'} < 0\}} &= I_{\{\widehat{W}_{kl}/n - \widehat{W}_{k'l'}/n > 0\}} \\ &= I_{\{\lambda_{kl} - \lambda_{k'l'} > 0\}} + O_p(n^{-1/2}). \end{aligned}$$

Ce résultat est valable pour toutes les comparaisons par paires évaluées simultanément, de sorte que nous concluons que le fractionnement ayant la valeur la plus élevée de $\lambda_{kl} = n(P_{kl} - P_{kl^c})^2 / V_{kl}$ est sélectionné avec une probabilité tendant vers 1.

Le théorème 1 décrit le comportement asymptotique des statistiques du test de Wald et la sélection de fractionnement qui en résulte à chaque étape. Lorsque ces résultats sont pris en compte pour la séquence de fractionnements évaluée lors de l'exécution de l'algorithme CHAID pour les enquêtes, il est possible de décrire le comportement asymptotique de l'arbre résultant. Cela est énoncé dans le corollaire suivant, qui montre que l'algorithme a une limite de probabilité bien définie dans la population.

Corollaire 1. *L'arbre de l'échantillon $\widehat{\mathcal{T}}_s$ obtenu par l'algorithme CHAID pour les enquêtes décrit à la section 2 est constitué de la séquence de longueur $\leq R$ des fractionnements de nœuds appliquée à l'échantillon s. L'arbre de l'échantillon $\widehat{\mathcal{T}}_s$ converge vers un arbre de population $\mathcal{T}_U$ avec une probabilité tendant vers 1. L'arbre $\mathcal{T}_U$ est constitué de la séquence de longueur R des fractionnements effectués sur l'ensemble des valeurs $X_k = l$ qui maximisent*

$$\frac{(P_{kl} - P_{kl^c})^2}{V_{kl}}$$

sur tous les fractionnements admissibles à chaque étape. L'arbre $\mathcal{T}_U$ est unique selon les hypothèses énoncées.

L'estimateur de variance défini dans le lemme 2 est cohérent, mais ses deux derniers termes nécessitent l'utilisation des estimateurs de propension $\widetilde{R}_g^*$ pour le niveau le plus détaillé des GRH. Si G est de grande taille, ces estimateurs peuvent être très variables ou même donner des cellules vides dans la pratique. Cela contraste avec les termes restants dans $\widehat{V}_{kl}$, qui peuvent tous être calculés directement pour les deux côtés du fractionnement des nœuds étant évalués. Par conséquent, nous proposons deux autres approches pour obtenir un dénominateur pour les variables test.

Premièrement, les termes concernant $\widetilde{R}_g^*$ pourraient simplement être omis. Cela simplifierait grandement les calculs, car les termes restants sont faciles à obtenir au moyen des logiciels pour les enquêtes standard qui mettent en œuvre des tests du χ^2, par exemple ceux fournis par la procédure SURVEYFREQ dans SAS et le paquet survey dans R. Cela reviendrait en quelque sorte à ne calculer que le terme de l'estimateur de variance au niveau de l'unité primaire d'échantillonnage (UPE) dans le contexte de l'échantillonnage à plusieurs degrés. Puisque dans ce contexte, ces termes ne devraient représenter qu'une petite fraction de l'estimateur de la variance globale lorsque la fraction d'échantillonnage est petite. Deuxièmement, les termes pourraient être remplacés en effectuant la moyenne des deux côtés du fractionnement des nœuds, c'est-à-dire à l'aide de

$$\frac{n}{N_{kl}} \widehat{P}_{kl}(1 - \widehat{P}_{kl}) + \frac{n}{N_{kl^c}} \widehat{P}_{kl^c}(1 - \widehat{P}_{kl^c}).$$

Cette dernière option peut être préférable si la fraction d'échantillonnage n'est pas négligeable. Nous évaluerons les deux options dans les simulations ci-dessous.

4. Simulation

Pour étudier le rendement de l'algorithme CHAID pour les enquêtes, nous avons réalisé une étude par simulation comparant la méthode CHAID pour les enquêtes proposée à la méthode CHAID traditionnelle et au paquet rpms (Toth, 2017). Tous les calculs ont été effectués à l'aide du logiciel statistique R (R Core Team, 2021).

Les données au niveau du ménage des échantillons de microdonnées à grande diffusion pour l'American Community Survey (ACS) de 2017 à 2021 ont été utilisées comme base de sondage. La base de sondage a été stratifiée par division de recensement. Les zones de microdonnées à grande diffusion (ZMGD) étaient les UPE et les ménages étaient les unités secondaires d'échantillonnage, excluant les logements collectifs. À partir de chacune des neuf divisions de recensement, nous prélevons un échantillon aléatoire simple de 10 UPE, et un échantillon aléatoire simple de 100 ménages est sélectionné dans chaque UPE échantillonnée, ce qui donne des échantillons en grappes de ménages ayant des probabilités inégales. Le tableau 4.1 présente le nombre de ZMGD et le nombre moyen de ménages par ZMGD selon la division de recensement dans la base de sondage.

Tableau 4.1
Nombre de ZMGD et nombre moyen de ménages par ZMGD selon la division de recensement, échantillon de microdonnées de l'ACS de 2017 à 2021

Division de recensement	Nombre de ZMGD	Nombre de ZMGD échantillonnées	Nombre moyen de ménages par ZMGD	Nombre de ménages échantillonnés par ZMGD
Nouvelle-Angleterre	109	10	2 946	100
Atlantique centre	310	10	2 790	100
Centre nord-est	339	10	3 026	100
Centre nord-ouest	159	10	3 008	100
Atlantique sud	455	10	2 950	100
Centre sud-est	138	10	2 984	100
Centre sud-ouest	294	10	2 599	100
Rocheuses	180	10	2 791	100
Pacifique	367	10	2 642	100

Note : ZMGD = zone de microdonnées à grande diffusion; ACS = American Community Survey.

Deux ensembles de 1 000 échantillons de 9 000 ménages ont été tirés pour la simulation, ce qui correspond à un scénario de non-réponse faible et élevée. Le poids de base des ménages sélectionnés a été fixé au produit du poids au niveau des ménages de l'ACS et de la réciproque de la probabilité de sélection. L'état de réponse, R_i, a été généré pour chaque ménage $(i = 1, 2, \ldots, 9\,000)$ à partir d'une distribution de Bernoulli avec une probabilité p_i, où

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{8} \beta_j x_{j,i}.$$

Les huit covariables, $\mathbf{x} = (x_1, \ldots, x_8)^T$, étaient la région du recensement (trois variables indicatrices pour les régions du Midwest, du Sud et de l'Ouest), le type de bâtiment (maison unifamiliale non attenante ou non), le statut d'occupation (propriétaire ou non), la présence d'enfants dans le ménage, le statut de couverture par l'assurance-maladie et la valeur de la propriété. Afin de reproduire un mécanisme de réponse réaliste, les paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_6)^T$ ont été obtenus en adaptant un modèle pour la réponse initiale à l'ensemble de données des échantillons de microdonnées à grande diffusion; l'indicateur de réponse a été fixé à 1 si le mode de réponse de l'ACS était le Web ou le courrier et à 0 si le mode de réponse était le téléphone ou la visite en personne (ce qui indique que le répondant n'a pas répondu aux tentatives antérieures de communiquer avec lui par les autres moyens). Dans la simulation, le ménage i était considéré comme un répondant si l'état de réponse r_i réalisé était de 1 et comme un non-répondant autrement. Le taux de réponse moyen pour les deux ensembles de 1 000 échantillons était d'environ 74,5 % et 31,7 %, respectivement. Le premier scénario correspond à l'ajustement du modèle logistique à l'état de réponse à l'ACS observé décrit ci-dessus, et le scénario pour lequel le taux de réponse était de 31,7 % a été établi en modifiant l'ordonnée à l'origine β_0.

Lors de la mise en œuvre des différentes méthodes de partitionnement récursif, l'indicateur de réponse généré était la variable dépendante, et 12 covariables potentielles ont été prises en compte. Ces dernières comprenaient les huit covariables du modèle de réponse ci-dessus et quatre autres covariables qui n'ont pas été utilisées : état du service téléphonique, type de famille, nombre de personnes dans la famille et accès à Internet. Pour la méthode CHAID traditionnelle, nous avons mis en œuvre deux versions : une reposant sur les 12 covariables seulement et une qui ajoute la variable du plan d'échantillonnage (ZMGD) et le poids de base dans le modèle. La première version est appelée « CHAID », tandis que la seconde est appelée « CHAID et information sur le plan ». La seconde version peut se comparer davantage à la méthode CHAID pour les enquêtes, puisqu'elle comprend les renseignements du plan dans un cadre basé sur un modèle (Little et Vartivarian, 2003).

Dans chaque méthode, à l'exception de la valeur-*p* fixée à 0,05, nous n'avons pas ajusté d'autres paramètres de l'algorithme pour contrôler le comportement des arbres résultants. La variable khi carré de Rao-Scott du deuxième ordre (Satterthwaite) pour les tableaux à deux dimensions a été utilisée dans la méthode CHAID pour les enquêtes (Rao et Scott, 1981). Cette variable est la valeur par défaut de la fonction « svychisq » du paquet survey (Lumley, 2024) dans R. Nous avons également étudié la possibilité d'utiliser l'estimateur de variance $\widehat{V}_{kl}$ du lemme 2, en appliquant la deuxième option de simplification proposée à la section 3 pour tenir compte des deux derniers termes de variance. Les résultats obtenus en ajoutant les deux derniers termes étaient pratiquement impossibles à distinguer et ne sont pas présentés dans la présente étude.

Le tableau 4.2 présente l'erreur quadratique moyenne de prédiction pour l'état de réponse, la proportion d'arbres construits avec les huit prédicteurs utilisés pour générer l'état de réponse et la proportion d'arbres construits avec au moins une des quatre autres covariables qui n'ont pas été utilisées dans le mécanisme de

réponse. Le tableau 4.3 présente le biais relatif empirique et la racine carrée de l'erreur quadratique moyenne relative empirique pour les estimations du revenu moyen des ménages, une variable mesurée par l'ACS. Les estimations reposaient sur les poids ajustés pour la non-réponse, qui ont été calculés en divisant le poids de base par les propensions à répondre estimées obtenues pour chacun des nœuds terminaux de l'arbre dans le cadre du modèle du GRH.

Tableau 4.2
Erreur quadratique moyenne de prédiction et proportions des arbres construits avec les huit covariables et des arbres construits avec d'autres covariables à partir de 1 000 échantillons simulés

Taux de réponse	Algorithme	Erreur quadratique moyenne de prédiction (non pondérée)	Proportion des arbres construits avec les huit covariables	Proportion des arbres construits avec d'autres covariables
Élevé	CHAID	0,0021	0,99	0,11
74,5 %	CHAID et information sur le plan	0,0022	0,99	0,12
	rpms	0,0043	0,95	0,07
	CHAID pour les enquêtes	0,0032	0,99	0,06
Faible	CHAID	0,0026	0,99	0,12
31,7 %	CHAID et information sur le plan	0,0025	0,99	0,13
	rpms	0,0044	0,96	0,07
	CHAID pour les enquêtes	0,0031	0,98	0,07

Note : CHAID = Chi-square Automatic Interaction Detector.

Tableau 4.3
Biais relatif empirique et racine carrée de l'erreur quadratique moyenne relative empirique pour le revenu des ménages selon l'algorithme arborescent à partir de 1 000 échantillons simulés

Taux de réponse	Algorithme	Biais relatif empirique (%)	Racine carrée de l'erreur quadratique moyenne relative empirique (%)
Élevé	Aucun ajustement	-3,70	3,98
74,5 %	CHAID	-0,20	1,68
	CHAID et information sur le plan	-0,19	1,70
	rpms	-0,33	1,59
	CHAID pour les enquêtes	-0,21	1,50
Faible	Aucun ajustement	-6,62	7,57
31,7 %	CHAID	-1,12	2,75
	CHAID et information sur le plan	-1,10	2,77
	rpms	-1,51	2,98
	CHAID pour les enquêtes	-0,98	2,24

Note : CHAID = Chi-square Automatic Interaction Detector.

Les résultats du tableau 4.2 montrent que les quatre méthodes ont bien cerné les variables dans le modèle de réponse, même si la méthode rpms a manqué au moins une variable un peu plus souvent. Les deux versions de la méthode CHAID traditionnelle présentaient la fraction la plus élevée d'arbres contenant des variables parasites, ce qui n'est pas surprenant étant donné qu'elles ne tiennent pas compte de la réduction

de la taille effective de l'échantillon en raison de la pondération et de la mise en grappes dans le processus de construction des arbres. Par conséquent, elles ont eu tendance à entraîner un ajustement excessif plus fréquemment que les méthodes qui tiennent compte du plan d'échantillonnage. Lorsque les arbres estimés obtenus par les quatre méthodes ont été utilisés pour tenir compte de la non-réponse, les résultats du tableau 4.3 montrent qu'elles ont toutes été efficaces pour éliminer la majorité du biais des estimateurs résultants; il n'y a eu que des différences modestes entre eux, mais un léger avantage en précision existait pour l'estimateur ajusté par la méthode CHAID pour les enquêtes.

5. Discussion

Nous avons proposé la méthode CHAID pour les enquêtes comme une extension de la méthode CHAID qui convient mieux à la formation de cellules d'ajustement pour la réponse dans un contexte fondé sur le plan, en intégrant à la fois les caractéristiques de pondération et de variance du plan, telles que la mise en grappes et la stratification, dans son critère de partitionnement. Parallèlement, elle est toujours fondée sur l'approche CHAID générale, qui comprend l'utilisation de tests du χ^2 et de valeurs-p de sorte que les utilisateurs actuels de la méthode CHAID seront en mesure de s'adapter facilement à cet algorithme de partitionnement récursif modifié. Nous sommes en train d'élaborer un paquet dans R facile à utiliser qui s'intègre au paquet pour les enquêtes et aux paquets de partitionnement existants dans R.

Nous avons décrit la méthode CHAID pour les enquêtes dans le contexte de la construction d'ajustements pour la non-réponse. Toutefois, les algorithmes de partitionnement récursif peuvent également être utilisés pour adapter les arbres de classification et de régression de manière plus générale. L'algorithme CHAID pour les enquêtes peut également être utilisé en y apportant des modifications mineures pour les arbres de classification et de régression pour les données d'enquête. L'élaboration de la théorie à cet effet suivrait l'approche de Toth et Eltinge (2011).

Bibliographie

Beaumont, J.-F., Bosa, K. Brennan, A., Charlebois, J. et Chu, K. (2024). Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada. *Techniques d'enquête*, 50(1), 87-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00004-fra.pdf.

Breiman, L., Friedman, J.H., Olshen, R. et Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, Californie: Wadsworth.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Offcial Statistics*, 29, 329-353.

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23(1), 37-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3103-fra.pdf.

Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2 ed.). New York: John Wiley & Sons, Inc.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Gordon, L., et Ohlshen, R. (1980). Nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10, 611-627.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society, Series C*, 29, 119-127.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

Little, R.J.A., et Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 9(22), 1589-1599.

Lumley, T. (2024). survey: Analysis of complex survey samples. R package version 4.4.

Phipps, P., et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772-794.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienne, Autriche: R Foundation for Statistical Computing.

Rao, J.N.K., et Scott, A. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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

Toth, D. (2017). rpms: An R package for modeling survey data with regression trees. US Bureau of Labor Statistics.

Toth, D., et Eltinge, J.L. (2011). Building consistent regression trees from complex survey data. *Journal of the American Statistical Association*, 106, 1626-1636.