Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Développer des arbres de régression qui utilisent les covariables de la base de sondage pour explorer le fardeau de réponse afin d'établir le plan d'enquête

par Yeng Xiong, Laura T. Bechtel, Colt S. Viehdorfer et Diane K. Willimack

Date de diffusion : le 29 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Développer des arbres de régression qui utilisent les covariables de la base de sondage pour explorer le fardeau de réponse afin d'établir le plan d'enquête

Yeng Xiong, Laura T. Bechtel, Colt S. Viehdorfer et Diane K. Willimack [1]

Résumé

La Direction des études économiques du U.S. Census Bureau élabore des procédures coordonnées de plan et de sélection des échantillons pour son Annual Integrated Economic Survey (AIES, Enquête économique annuelle intégrée). L'échantillon unifié remplacera la pratique actuelle de la Direction qui consiste à élaborer de façon indépendante des bases de sondage et des procédures d'échantillonnage pour une série d'enquêtes annuelles distinctes, ce qui optimise les caractéristiques du plan d'échantillonnage au prix d'un fardeau de réponse accru. Les attributs de taille des populations d'entreprises (p. ex. les revenus et l'emploi) sont considérablement biaisés. Un pourcentage élevé d'entreprises exercent au sein de plusieurs secteurs. De nombreuses entreprises sont donc échantillonnées dans le cadre de multiples enquêtes; ce qui accroît le fardeau de réponse, en particulier pour les entreprises de taille moyenne.

Même si cette composante de fardeau de réponse est réduite en sélectionnant un seul échantillon coordonné, elle n'est pas entièrement supprimée. Le fardeau de réponse dépend de plusieurs facteurs, notamment (1) de la longueur et la complexité du questionnaire, (2) de l'accessibilité des données, (3) du nombre attendu de mesures répétées et (4) de la fréquence de la collecte. Le plan de sondage peut avoir des répercussions profondes sur les troisième et quatrième facteurs. Pour contribuer aux décisions relatives au plan de sondage intégré, nous utilisons des arbres de régression afin de relever les covariables de la base de sondage associées au fardeau de réponse. En utilisant une base et des données de réponse historiques provenant de quatre enquêtes échantillonnées indépendamment, nous mettons à l'essai divers algorithmes, puis dressons des arbres de régression qui expliquent les relations entre les niveaux attendus de fardeau de réponse (tels qu'ils sont mesurés par le taux de réponse) et les covariables de base communes à plusieurs enquêtes. Nous validons les constats initiaux par une validation croisée, en examinant les résultats au fil du temps. Enfin, nous faisons des recommandations sur la façon d'intégrer nos résultats robustes au plan de sondage coordonné.

Mots-clés : fardeau de réponse; enquêtes auprès des établissements; arbres de régression; échantillonnage coordonné.

1. Introduction

Conformément à une recommandation de la National Academy of Sciences (National Academies of Sciences, Engineering, and Medicine, 2018), la Direction économique du U.S. Census Bureau est en train de créer un système intégré (coordonné) de sélection d'échantillons, nommé Annual Integrated Economic Survey (AIES), conçu pour six des enquêtes annuelles de la Direction. La portée de chaque enquête tend à être axée sur une industrie de l'économie en particulier, comme la fabrication ou le commerce de détail, mais le contenu de ces enquêtes se chevauche souvent. La base de sondage de chaque enquête est élaborée et tenue à jour indépendamment, ce qui optimise les caractéristiques du plan de sondage au prix d'un fardeau de réponse accru. Étant donné qu'un pourcentage élevé d'entreprises exercent leurs activités dans plus d'une industrie, elles peuvent être sélectionnées pour recevoir plusieurs enquêtes. Le plan de l'AIES remplacera ce processus inefficace par un échantillon unifié qui tient compte du fardeau de réponse. La recherche présentée ici, qui étudie la relation entre le fardeau de réponse (tel qu'il est mesuré par le taux de réponse) et les covariables disponibles dans la base de sondage pour les enquêtes annuelles, constitue une

[1]Yeng Xiong, U.S. Census Bureau, 4 600 Silver Hill Rd, Washington, DC, É.-U., 20233 (yeng.xiong@census.gov); Laura T. Bechtel, U.S. Census Bureau, (laura.bechtel@census.gov); Colt S. Viehdorfer, U.S. Census Bureau (colt.s.viehdorfer@census.gov); Diane K. Willimack, U.S. Census Bureau (diane.k.willimack@census.gov)
Tous les points de vue exprimés sont ceux des auteurs et ne reflètent pas nécessairement ceux du U.S. Census Bureau (Bureau du recensement des États-Unis). Le U.S. Census Bureau a vérifié si ce produit de données respectait les règles relatives à la divulgation non autorisée de renseignements confidentiels et a approuvé les pratiques de prévention de la divulgation appliquées. (Numéro d'approbation : CBDRB-FY21-ESMD005-002).

petite partie de la transformation de l'AIES et la première étape de l'élaboration d'une mesure du fardeau aux fins du remaniement de l'échantillon.

Dans cette recherche, nous nous intéressons à quatre des six enquêtes annuelles dont les procédures d'échantillonnage sont en train d'être intégrées : l'Annual Survey of Manufactures (ASM, Enquête annuelle sur les manufactures), l'Annual Retail Trade Survey (ARTS, Enquête annuelle sur le commerce de détail), l'Annual Wholesale Trade Survey (AWTS, Enquête annuelle sur le commerce de gros) et l'Annual Capital Expenditures Survey (ACES, Enquête annuelle sur les dépenses en immobilisations). Les entreprises échantillonnées sont dans l'obligation légale de répondre aux sondages. L'ASM tire son échantillon des établissements de fabrication stratifiés selon l'industrie et leur admissibilité au questionnaire. Son échantillon est sélectionné avec une probabilité proportionnelle aux recettes. L'ARTS, l'AWTS et l'ACES sont des échantillons aléatoires simples stratifiés au niveau de l'entreprise. La population de l'ARTS est propre aux entreprises de commerce détail et l'AWTS au commerce de gros. L'ARTS et l'AWTS sont stratifiées par industrie et par taille. Les entreprises échantillonnées pour l'ACES couvrent toutes les industries non agricoles et sont d'abord divisées en bases de sondage des employeurs et des non-employeurs. Les unités des employeurs sont ensuite stratifiées selon l'industrie et la taille, tandis que les unités des non-employeurs sont classées dans l'une des quatre strates suivantes : entreprises individuelles, sociétés et sociétés en nom collectif, les entreprises ayant un numéro d'identification d'employeur enregistré mais qui n'ont pas d'employés ni d'employés salariés, et les entreprises qui n'ont pas d'employés ou de salariés, mais qui peuvent en avoir d'ici la période de collecte des données. L' ASM[2], l' ARTS[3] et l' AWTS[4] sont des enquêtes longitudinales ayant un cycle d'échantillonnage d'au moins cinq ans, et l' ACES[5] est une enquête transversale.

2. Fardeau de réponse

On considère que le fardeau de réponse contribue à la non-réponse et nuit à la qualité des données. La législation américaine exige que les répondants potentiels soient informés du fardeau de réponse prévu, défini en fonction du temps nécessaire estimé pour remplir un sondage (y compris la lecture des instructions et la collecte de données). Toutefois, la littérature scientifique montre que le fardeau de réponse a d'autres attributs que le temps passé à répondre. Willeboordse (1997) classe le fardeau en quatre dimensions ayant chacune deux aspects possibles : objectif vs subjectif, brut vs net, maximaliste vs minimaliste, et imposé vs accepté. Les dimensions fournissent un contexte facilitant l'examen et l'évaluation du fardeau. Le temps, une mesure courante du fardeau, relève de l'objectif, car il mesure le fardeau de réponse au moyen de quantités calculables, tandis que le subjectif fait référence au fardeau perçu, qui relève de l'expérience ressentie, par exemple la mesure dans laquelle les répondants ont trouvé l'enquête facile ou difficile. Les données brutes tiennent compte du coût du fardeau, tandis que les données nettes tiennent compte des avantages compensatoires de la réponse à l'enquête. Le fardeau maximaliste tient compte des mesures supplémentaires prises au moment de remplir un questionnaire, comme la formation d'un dossier, les tâches de préparation et les nouveaux contacts après la collecte; le fardeau minimaliste tient seulement compte du fait que le questionnaire a été rempli. Le fardeau imposé suppose que toutes les unités de l'échantillon subissent le fardeau. Le fardeau accepté suppose que les répondants ont consenti au fardeau et l'ont assumé. Dans notre recherche, le fardeau est défini comme étant objectif, brut, maximaliste et imposé.

En s'appuyant sur cette définition du fardeau, la littérature scientifique détermine plusieurs facteurs associés au fardeau objectif ou au fardeau perçu (Dale et Haraldsen, 2007; Giesen et coll., 2018). Ces facteurs sont classés selon qu'ils tiennent compte de la motivation, qui peut être associée aux avantages perçus de la réponse à l'enquête, des procédures de collecte des données, qui sont habituellement considérées comme associées à l'aspect du coût, ou de la complexité organisationnelle, que l'on suppose également corrélée positivement au fardeau. Une covariable de la base de sondage, covariable accessible pendant l'échantillonnage, est ensuite reliée, si possible, à l'un des facteurs. Les facteurs de motivation susceptibles d'avoir une incidence sur le fardeau de réponse peuvent comprendre la question

[2] https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html

[3] https://www.census.gov/programs-surveys/arts/technical-documentation/methodology.html

[4] https://www.census.gov/programs-surveys/awts/technical-documentation/methodology.html

[5] https://www.census.gov/programs-surveys/aces/technical-documentation/methodology.html

de savoir si le répondant fait généralement confiance au U.S. Census Bureau et s'il trouve utiles les statistiques économiques produites. Toutefois, on ne connaît pas les covariables présentes sur la base de sondage qui peuvent être associées empiriquement à ces facteurs de motivation. De fait, la détermination des facteurs de motivation exige généralement un suivi supplémentaire, qui peut être utile ou ne pas l'être jusqu'à la sélection de l'échantillon suivant. Il a été montré que les caractéristiques du questionnaire lui-même sont associées au fardeau de réponse (Haraldsen et Jones, 2007). Par conséquent, il peut être utile d'avoir de l'information sur le questionnaire (p. ex. nombre de questions, disponibilité des données dans les dossiers, lisibilité, mise en page, mode de collecte, etc.). Toutefois, cette information ne figure pas dans la base de sondage; il s'agit de métadonnées générées à partir du questionnaire, qu'il peut être difficile de définir. Considérons le nombre de questions de l'enquête. Une enquête peut modifier le nombre de questions présentées à une unité en fonction de sa réponse à certaines questions. Autrement dit, certaines unités peuvent avoir moins de questions que d'autres unités répondant à la même enquête. Le choix du moment et les formalités de déclaration sont également des facteurs y contribuant. Selon la période de l'année où un questionnaire est remis, les entreprises peuvent avoir d'autres responsabilités en matière de déclaration susceptibles de les empêcher de répondre. Si elles doivent répondre à plusieurs enquêtes ou si le personnel en mesure de le faire n'est pas nombreux, elles peuvent ne pas répondre. La présente analyse ne comprend pas ces types de métadonnées, mais une future analyse pourrait les prendre en compte.

Bien que les bases de sondage manquent habituellement de covariables indiquant la motivation et la collecte des données, elles ont plusieurs covariables qu'on considère comme étant associées à la complexité organisationnelle d'une unité, qui joue un rôle important dans le processus de réponse aux enquêtes-entreprises (Willimack et Nichols, 2010; Bavdaz, 2010) et contribue au fardeau de réponse. Par exemple, à mesure de la croissance des entreprises, leur structure peut se complexifier par l'ajout d'emplacements, l'exercice de leurs activités dans plusieurs industries, et la division en plusieurs services, voire par l'adoption ou la redéfinition de leur cadre juridique. Dans ce cas, les données sont distribuées dans toute l'organisation, ce qui alourdit le fardeau, car les répondants pourraient devoir consulter plusieurs sources pour répondre aux questions de l'enquête. Pourtant, ils élaborent des formalités de déclaration pour réduire au minimum leur propre fardeau de réponse. Parce que les relations entre ces variables sont hypothétiques, tout comme leur association avec le fardeau de réponse, ces facteurs et les covariables connexes de la base d'échantillonnage sont définis dans le tableau 2-1.

Tableau 2-1
Covariables de la base de sondage utilisées comme variables explicatives et facteurs correspondants

Facteur	Covariable de la base de sondage	Description de la covariable	Disponible pour les enquêtes
Structure organisationnelle	TYPE	Indicateur des unités ayant un établissement (SU) ou plusieurs (MU)	ASM, ARTS, AWTS, ACES
	STRATE	Les facteurs de l'unité de la strate d'échantillonnage sont dans[6]	ACES
	LFO	Facteurs de la forme juridique de l'organisation	ACES
Diversité d'industrie	NO_NAICS	Nombre d'industries dans lesquelles l'entreprise a des activités	ASM, ACES
Taille	NO_ESTAB	Nombre d'établissements de l'entreprise	ASM, AWTS
	EMP_TYPE	Facteur employeur et non-employeur	ACES
	PAYROLL	Masse salariale annuelle de l'entreprise (en dollars)	ASM, ACES
	PAYROLL_ESTAB	Masse salariale annuelle pour l'établissement (en dollars)	ASM
	RECEIPT	Recette annuelle de l'entreprise (en dollars)	ARTS, AWTS
	CERT	Indicateur des unités échantillonnées avec certitude (c.-à-d. probabilité de sélection = 1)	ASM, ARTS, AWTS, ACES
Formalités de déclaration	INPREVSAMP	Indicateur d'unité échantillonnée auparavant	ACES

[6] Les unités des employeurs aux fins de l'ACES sont stratifiées selon l'industrie et la taille; les unités des non-employeurs aux fins de l'ACES sont stratifiées selon la forme juridique d'organisation (FJO).

3. Méthodologie

La réponse est la variable dépendante et sert de variable substitutive du fardeau de réponse. Cela suppose que les répondants ont accepté le fardeau et que le fardeau des non-répondants est trop lourd pour qu'ils répondent. Pour examiner la relation entre le fardeau de réponse et les covariables de la base de sondage du tableau 2-1, un arbre de régression est formé à partir d'une série de fractionnements qui produisent des branches et des nœuds par l'ajout des covariables successives les plus associées à la réponse. L'arbre cesse de croître une fois qu'un paramètre de complexité est atteint ou qu'aucune amélioration ne peut être apportée pour réduire au minimum l'erreur entre la réponse et le taux de réponse (Breiman et coll., 1983). Les nœuds finaux d'un arbre, appelés nœuds terminaux, représentent des unités ayant des profils de réponse homogènes. On suppose que les nœuds ayant des taux de réponse inférieurs représentent un fardeau relatif plus élevé.

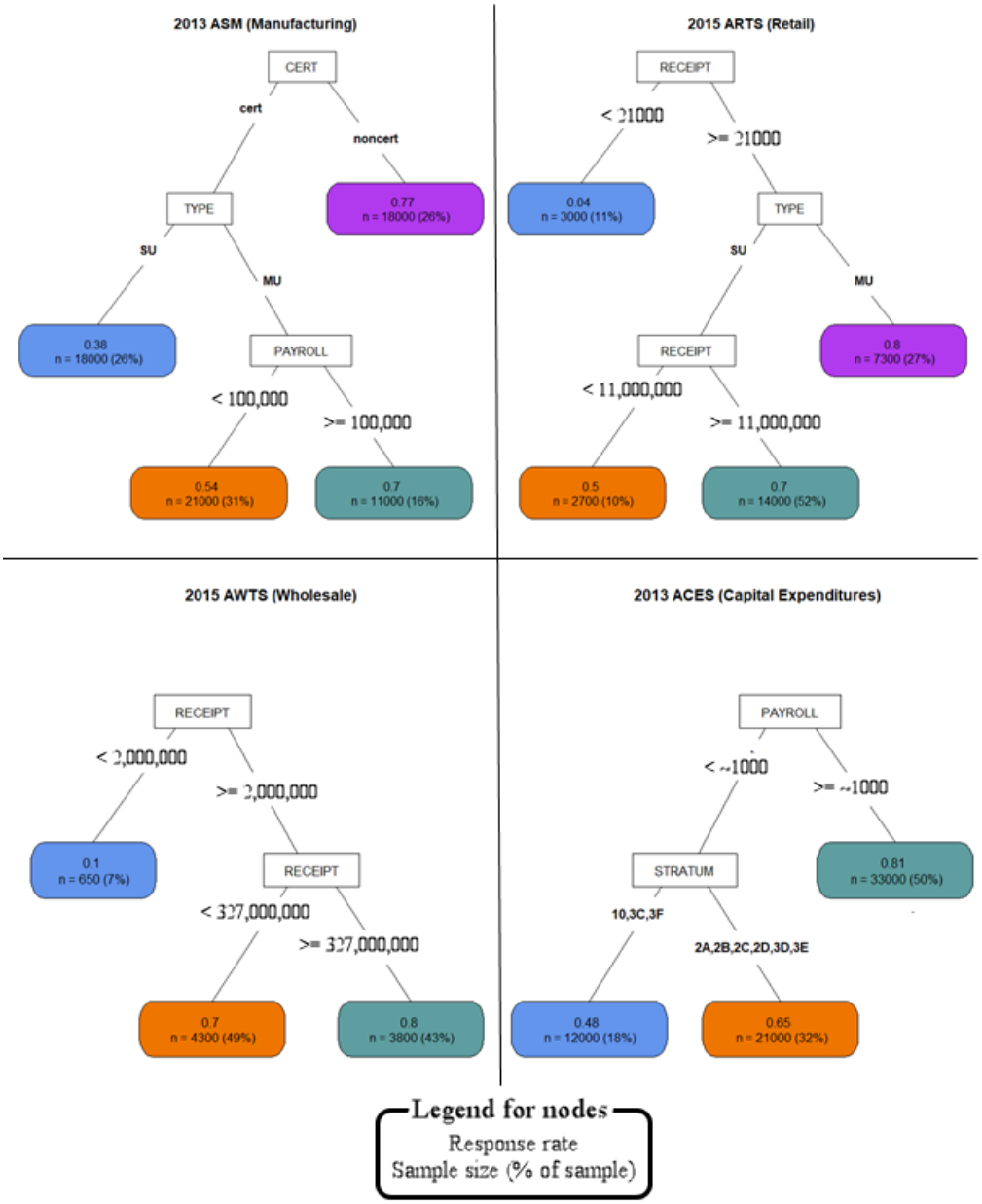
Le progiciel R *rpart* sert à développer les arbres de régression (Therneau et Atkinson, 2019). La croissance des arbres est illimitée pour toutes les enquêtes, et les arbres ne sont pas pondérés puisque le fardeau de réponse concerne seulement les unités échantillonnées. Aux fins de l'ASM et l'ACES, les données de 2013 à 2019 sont utilisées dans l'analyse. Les données de l'ASM sont recueillies dans le cadre du recensement économique des années se terminant par 2 ou 7. C'est le cas des données de l'ASM de 2017 présentées ici. Les données de 2015 à 2019 sont disponibles pour l'ARTS et l'AWTS. Dans cette recherche, les arbres de régression analysent le fardeau de réponse dans une seule enquête et non sur plusieurs enquêtes. Pour nos enquêtes longitudinales ASM, ARTS et AWTS, un arbre de régression de base est construit au moyen des données de la première année – qui est la dernière année d'un cycle d'échantillonnage – et évalué les années suivantes quand un nouvel échantillon est sélectionné (l'ASM de 2019 est le début d'un nouvel échantillon). Dans le cas de l'ACES, pour laquelle un nouvel échantillon est sélectionné chaque année, les données de 2013 sont utilisées aux fins de la construction de l'arbre de régression de base et les données de 2014-2019 servent à l'évaluation (l'ACES de 2018 a utilisé les données d'employeur de l'ACES de 2017).

4. Résultats

Les arbres de régression de base sont tracés à la figure 4-1. Les nœuds colorés en bas d'un arbre de régression correspondent aux niveaux de fardeau attendus des unités ayant les caractéristiques définies par les covariables. Pour l'ASM, le premier fractionnement porte sur la covariable de la certitude. Les unités avec certitude (unités qui doivent être incluses dans l'échantillon) ont plus de fardeau que les unités sans certitude. La taille et la structure des unités avec certitude peuvent être différentes. Par conséquent, les unités avec certitude sont divisées aussi par le type de covariables et la masse salariale. Le faible taux de réponse observé parmi les unités uniques avec certitude implique qu'elles sont celles ayant le plus lourd fardeau. Les unités multiples avec certitude ont un fardeau « moyen », les unités plus petites – pour ce qui est de leur masse salariale – ayant un fardeau plus lourd que les unités plus grandes. Les unités dont le fardeau est le plus léger sont les unités sans certitude.

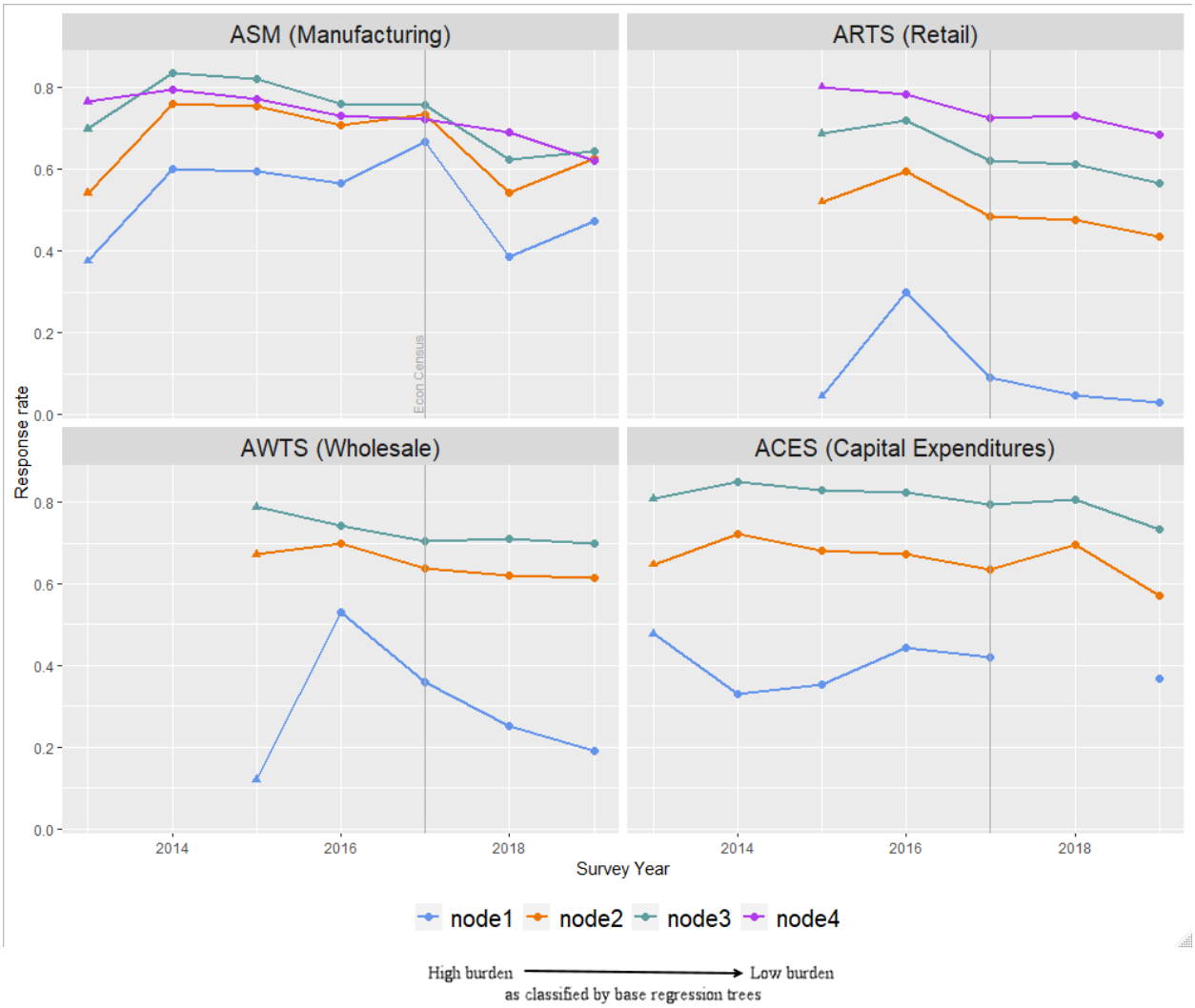
L'arbre de régression ARTS présente un scénario similaire, dans lequel la taille et la structure organisationnelle sont les principales covariables expliquant le fardeau. Les plus petites unités ont le fardeau le plus élevé, suivies des plus grandes unités uniques, qui sont de nouveau divisées par taille. Les unités multiples les plus grandes ont le fardeau le moins grand. L'arbre de l'AWTS est dominé par la taille. Les plus petites unités continuent d'avoir un fardeau plus élevé que les plus grandes.

Figure 4.-1
Arbres de régression de base caractérisant différents niveaux de fardeau, par enquête
(Sources des données : ASM 2013, ARTS 2015, AWTS 2015, ACES 2013)



Le premier fractionnement de l'arbre de l'ACES est fondé sur la taille, les unités plus petites ayant le plus grand fardeau. Les unités plus petites sont ensuite subdivisées en deux groupes. Le premier groupe comprend les petites avec certitude (strate 10) et les entreprises en démarrage (strates 3C,3F). Cela semble un couple inhabituel, mais après réflexion, il semble logique que ces entreprises soient les plus accablées par le fardeau. Les unités plus petites n'ont habituellement pas les ressources et le personnel nécessaires pour traiter une demande d'enquête, et le fait de la certitude alourdit ce fardeau. C'est pourquoi elles sont placées dans la même classe que les entreprises en démarrage. Le deuxième groupe comprend les unités avec certitude plus petites (strate 2A-2D) et les plus petites unités de non-employeurs comme les sociétés et les sociétés en nom collectif (strate 3D) et les entreprises individuelles (strate 3E). Les unités dont le fardeau est le plus léger sont les grandes unités.

Figure 4.-2
Taux de réponse des groupes ayant différents niveaux de fardeau déterminés par les arbres de régression dans le temps, par enquête
(Sources des données : ASM 2013-2019, ARTS 2015-2019, AWTS 2015-2019, ACES 2013-2019)



Au moyen des nœuds terminaux identifiés par les arbres de base de la figure 4-1, la figure 4-2 illustre les taux de réponse pour ces regroupements au fil du temps. Bien que les valeurs numériques des taux de réponse changent, l'ordre de classement des taux, qui est codé en couleur pour indiquer un niveau de fardeau relatif déterminé par son arbre de régression de base, est généralement uniforme dans le temps. Cette tendance se vérifie dans toutes les enquêtes. La seule exception à cette règle est le nœud violet de l'ASM, qui représente les unités de fabrication sans certitude. On s'attend à ce que le fardeau de réponse soit différent entre les unités avec certitude et celle sans certitude, et cela pourrait en être une preuve de plus. À l'avenir, il pourrait être nécessaire d'analyser les unités de sans certitude séparément des unités avec certitude.

5. Conclusions

Dans notre analyse, nous démontrons que l'analyse de l'arbre de régression permet de discerner différents niveaux de réponse, qui servent de substituts des niveaux de fardeau, et qui peuvent continuer à se vérifier dans le temps. Nos résultats empiriques montrent aussi un lien entre le fardeau de réponse et la taille, prévalent dans toutes les enquêtes.

Ce n'est pas surprenant : les grandes unités ont souvent du personnel chargé de remplir les questionnaires des organismes du gouvernement fédéral, alors que les petites unités ne disposent pas de cette ressource. Toutefois, bien qu'elle n'ait pas été détectée empiriquement dans la présente analyse, la relation entre le fardeau de réponse et la taille n'est pas toujours aussi linéaire. Les petites unités sont moins susceptibles d'être échantillonnées pour des enquêtes multiples et la déclaration les concernant est habituellement plus directe. L'intégration de métadonnées supplémentaires – comme celles dont il est question à la section 2 – qui traduisent les procédures de collecte des données pourrait aider à mieux décrire la relation complexe entre le fardeau de réponse et la taille. Nous avons constaté une autre limite : l'utilisation du taux de réponse comme indicateur du fardeau de l'enquête. Le taux de réponse élevé des grandes unités est gonflé sur le plan de la procédure, car elles ont généralement des formalités de suivi plus rigoureuses que les petites unités. Le fait de compléter la réponse par des renseignements sur le comportement du répondant (p. ex. connaître le temps de réponse ou savoir si un délai supplémentaire a été demandé) peut aider à mieux saisir le fardeau. Dans de futurs travaux, nous chercherons à élaborer une procédure, comme une mesure normalisée du fardeau fondée sur les résultats de cette analyse, pouvant servir à tenir compte du fardeau de réponse dans l'AIES au moment de coordonner l'échantillonnage dans les enquêtes annuelles de la Direction économique.

Remerciements
Les auteurs remercient Alfred Dave Tuttle, James Hunt, Ian Thomas, Valerie Mastalski, Steven Roman, Susan Pozzanghera, Jeremy Knutson, Joseph Barth, et Magdalena Ramos, pour leur lecture attentive et leurs commentaires constructifs sur les versions antérieures de l'article.

Bibliographie

Bavdaz, M. (2010), « The multidimensional integral business survey response model », *Survey Methodology*, 36, p. 81-93.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone. (1983), *Classification and Regression Trees*, Belmont, CA : Wadsworth.

Dale, T., et G. Haraldsen (éd.). (2007), *Handbook for Monitoring and Evaluating Business Response Burdens*, Luxembourg : Eurostat.

Giesen, D., M. Vella, C. F. Brady, P. Brown, D. Ravindra, et A. Vaasen-Otten. (2018), « Response Burden Management for Establishment Surveys at Four National Statistical Institutes », *Journal of Official Statistics*, 34, p. 397-418.

Haraldsen, G., et J. Jones. (2007), « Paper and Web Questionnaires Seen from the Business Respondent Perspective », *Proceedings of the Third International Conference on Establishment Surveys, American Statistical Association*, p. 1040-1047.

National Academies of Sciences, Engineering, and Medicine. (2018), *Reengineering the Census Bureau's Annual Economic Surveys*, Washington, DC : The National Academies Press, étude de consensus disponible à l'adresse https://www.nap.edu/read/25098/chapter/1.

Therneau, T. M., et E. J. Atkinson. (2019), « An Introduction to Recursive Partitioning using the RPART Routines », rapport inédit disponible à l'adresse http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Willeboordse, A. (1997), « Minimizing Response Burden », dans A. Willeboordes (éd.) *Handbook on Design and Implementation of Business Surveys*, Luxembourg : Eurostat, p. 111-118.

Willimack, D. K., et E. Nichols. (2010), « A Hybrid Response Process Model for Business Surveys », *Journal of Official Statistics*, 26, p. 3-24.