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

Techniques d'enquête

Commentaires à propos de l'article « Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage »

par M. Giovanna Ranalli

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.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2025

L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2025
Vol. 51, nº 1, pp. 141-150
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage »

M. Giovanna Ranalli[1]

Résumé

Le présent article porte sur certaines avancées réalisées au chapitre des plans d'enquête et de l'estimation et s'inspire de l'évaluation exhaustive des professeurs Jon Rao et Sharon Lohr sur les tendances actuelles dans ce domaine. Il traite de trois sujets en particulier, à savoir l'échantillonnage équilibré, le calage et l'estimation sur petits domaines. Les méthodes d'échantillonnage équilibré probabilistes, telles que la méthode du cube et l'échantillonnage équilibré pénalisé, sont explorées, en mettant l'accent sur la résolution des défis émergents, notamment les extensions aux modèles mixtes linéaires, aux modèles de régression non paramétriques et aux plans d'échantillonnage spatialement équilibrés. Le calage est discuté à l'aide d'un cadre modulaire qui intègre des techniques de régression modernes et met en évidence des utilisations novatrices du calage assisté par un modèle pour la vérification des données et l'inférence causale. L'estimation sur petits domaines dans le contexte de la modélisation par variables latentes et de l'intégration des données est étudiée en mettant l'accent sur son rôle quand la ou les variables d'intérêt ne peuvent être mesurées directement ou sans erreur. Les applications de l'intégration des données probabilistes et non probabilistes et de l'analyse causale à l'échelle locale sont également examinées.

Mots-clés : Calage assisté par un modèle; échantillonnage équilibré; estimation sur petits domaines; inférence causale; intégration des données; modélisation par variables latentes.

1. Introduction

Je tiens à féliciter les professeurs Jon Rao et Sharon Lohr, qui ont accompli la tâche difficile de traiter des tendances actuelles en matière de plans d'enquête et d'estimation dans le numéro qui marquait le 50e anniversaire de la revue *Techniques d'enquête*. Leur évaluation portait sur la collecte de données ainsi que sur les plans d'enquête et l'estimation et mettait plus particulièrement l'accent sur l'application des méthodes d'apprentissage automatique, l'estimation sur petits domaines et l'intégration des données. Il n'est jamais facile de résumer des tendances et des orientations futures, et, selon moi, cette tâche est encore plus exigeante aujourd'hui en raison des progrès rapides en ce qui concerne la théorie et la méthodologie des enquêtes par sondage encouragés par l'arrivée de nouvelles technologies et la diversification des sources de données au cours des dernières années. Je tiens également à les remercier pour les nombreux stimuli que j'ai eus en les lisant. Il m'est impossible d'en décrire tous les éléments dans cet article; je privilégie donc trois aspects : l'échantillonnage équilibré, le calage et l'estimation sur petits domaines.

2. Échantillonage équilibré

Phil Kott (Kott, 2007), se référant à « Model-assisted survey sampling » de Särndal, Swensson et Wretman (1992), a déclaré : « Le Livre jaune a sorti les modèles de l'ombre dans l'estimation des enquêtes

1. M. Giovanna Ranalli, Dept. of Political Science, University of Perugia, Italie. Courriel : maria.ranalli@unipg.it.

par sondage ». Je dirais aujourd'hui que la méthode du cube élaborée par Deville et Tillé (2004) a rendu possible un échantillonnage probabiliste équilibré et a sorti les modèles de l'ombre dans les plans d'échantillonnage. Dans sa formulation de base, l'échantillonnage équilibré peut être vu comme un calage effectué à la phase de planification et se trouve ainsi étroitement lié aux modèles de régression linéaire.

Parmi les multiples travaux traitant de l'échantillonnage équilibré, je trouve particulièrement pertinente l'étude de Breidt et Chauvet (2012), compte tenu des nouveaux défis à relever au cours des prochaines années. Ces auteurs proposent un échantillonnage équilibré pénalisé, qui assouplit les contraintes d'équilibrage pour certaines variables. Plus particulièrement, en étant fondé sur une notation similaire à celle de Rao et Lohr (2025), l'échantillonnage équilibré pénalisé permet de sélectionner un échantillon dans lequel le total de la population estimé $\hat{\mathbf{X}}$ de $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ est égale au total de la population connu $\mathbf{X}$, tandis que, pour le total de la population d'un autre ensemble de variables auxiliaires $\mathbf{z} = (z_1, z_2, \ldots, z_p)$, on accepte que $|\hat{\mathbf{Z}} - \mathbf{Z}| \leq \boldsymbol{\tau}$ corresponde à un vecteur $\boldsymbol{\tau}$ de tolérances prédéfini. Les deux vecteurs de variables auxiliaires $\mathbf{x}$ et $\mathbf{z}$ doivent être présents ou calculables à partir de variables présentes dans la base de sondage. En procédant ainsi, on peut réduire les erreurs de $\hat{\mathbf{Z}}$ si p est très grand et étendre l'échantillonnage équilibré aux modèles linéaires mixtes. Cette caractéristique est particulièrement pertinente en présence de bases de sondage riches; elle permet en outre d'adopter un plan d'échantillonnage qui prévoit une estimation sur petits domaines quand $\mathbf{z}$ représente le vecteur des indicateurs de domaine. À cela s'ajoute le fait qu'elle intègre, dans le plan d'échantillonnage, les écarts par rapport à la linéarité de relation entre les variables de l'enquête et une ou plusieurs variables auxiliaires. Cela se fait au moyen de modèles de régression non paramétrique qui peuvent être ajustés à l'aide de modèles mixtes, comme ceux dans lesquels les effets aléatoires sont les coefficients de p fonctions de base collectés en $\mathbf{z}$. Dans ce cas-ci, J est relativement petit, mais p représente généralement une grande valeur. Cela englobe non seulement des modèles à fonction spline pénalisée, mais aussi un lissage des valeurs bivariées de faible rang au moyen de fonctions de base radiale et un krigeage de faible rang (Ruppert, Wand et Carroll, 2003); on obtient donc des échantillons spatialement équilibrés en faisant appel à une perspective différente de celles déjà suggérées dans la littérature.

Les plans d'échantillonnage spatialement équilibrés ont été principalement utilisés dans les enquêtes écologiques et forestières, mais leur application devrait être davantage explorée dans d'autres contextes. Ils pourraient être particulièrement pertinents compte tenu de l'accessibilité croissante de données géographiques pour les unités, comme les municipalités et les ménages. Il serait par ailleurs utile d'étudier la faisabilité d'élaborer un échantillon spatialement équilibré dans les cas où l'*espace* n'est pas défini par des coordonnées géographiques, mais plutôt par l'espace multidimensionnel formé par d'autres variables auxiliaires. L'utilité pratique de l'échantillonnage spatialement équilibré dans les applications peut toutefois être limitée par l'absence d'algorithmes appropriés pour calculer la variance des estimations, car un échantillonnage équilibré entraîne généralement une réduction des valeurs pour de nombreuses probabilités d'inclusion de second ordre. Une orientation prometteuse de la recherche dans ce domaine réside dans l'extension de ce que suggèrent Breidt, Opsomer et Sanchez-Borrego (2016) pour relever le défi de produire

des estimations de variance stables pour les plans d'échantillonnage hautement stratifiés. Ces auteurs proposent notamment d'inclure des plans comportant une unité par strate et d'autres affichant un très petit nombre d'éléments par strate (par exemple deux ou trois). Dans ces cas, l'estimateur de variance classique est l'estimateur à strates regroupées, qui repose sur la création de pseudo-strates plus grandes et sur un calcul de la somme des différences quadratiques entre les totaux de strates estimés dans toutes les pseudo-strates. Breidt et coll. (2016) proposent l'utilisation d'une solution non paramétrique, qui consiste à remplacer les pseudo-strates par des strates voisines pondérées par la méthode des noyaux et établit l'estimation de la variance au moyen des écarts par rapport à une fonction moyenne ajustée. Cette approche convient particulièrement bien aux situations comportant un ordre naturel des strates, par exemple par une mesure x_i de celles-ci. Bien que l'article cible surtout le cas d'une variable univariée dans l'élaboration d'une valeur de pondération des noyaux, on peut aussi utiliser cette méthode dans des scénarios où $\mathbf{x}_i$ est un vecteur de coordonnées spatiales.

3. Calage

Si Särndal et coll. (1992) ont de nouveau fait la lumière sur les modèles de régression linéaire dans l'estimation des échantillons de sondages, Wu et Sitter (2001) ont, pour leur part, publié un article qui a très visiblement ouvert la voie à l'extension des techniques de régression non linéaires modernes dans l'estimation par calage. Rao et Lohr (2025) ont traité du calage assisté par un modèle et considèrent que dans ce cas, une fonction non linéaire $m(\cdot)$ des variables auxiliaires $\mathbf{x}$ décrit mieux la variable étudiée y qu'un simple modèle de régression linéaire et remplace les contraintes de référence sur la population totale $\mathbf{X}$ par celles qui s'appliquent à la population totale $\hat{m}(\mathbf{x}_i)$. De cette façon, les poids calibrés finaux $w_i = d_i g_i$ peuvent entraîner une amélioration considérable de l'efficacité, mais en même temps éliminer les avantages pertinents de la cohérence avec la population totale connue $\mathbf{X}$ et d'un système de poids polyvalent applicable à toutes les variables y. On peut facilement corriger ces deux inconvénients si l'on ne considère pas le calage assisté par un modèle comme une méthode remplaçant le calage classique, mais cela ne constitue qu'un moyen d'intégrer, dans l'approche de calage, des modèles plus complexes par l'application d'une ou de plusieurs contraintes supplémentaires (une pour chaque prédiction du modèle). Le calage, en fait, peut se faire dans une approche modulaire consistant à élaborer un ensemble final de poids en incluant les variables suivantes dans le vecteur des variables utilisées pour délimiter les contraintes de calage : (i) les variables prévues aux fins de la description du processus de réponse, (ii) les variables qui doivent être soumises au calage à des fins de cohérence et (iii) les variables incluses pour améliorer l'efficacité. Bien entendu, on peut s'attendre à ce que certaines variables servent les trois objectifs et, bien sûr, les prédictions du modèle $\hat{m}_1(\mathbf{x}_i), \hat{m}_2(\mathbf{x}_i), \ldots, \hat{m}_P(\mathbf{x}_i)$ obtenues pour un sous-ensemble de P variables clés de l'enquête sont les candidats naturels pour le groupe (iii). La perspective modulaire peut à l'occasion (et utilement) répartir le problème de calage en différentes étapes, de sorte que le groupe (i) est utilisé pour un premier calage qui produit un ensemble de poids ajustés pour la non-réponse, qui sont ensuite utilisés comme poids de départ

dans une deuxième étape qui effectue le calage sur les groupes (ii) et (iii). Dans la pratique, le nombre de contraintes s'accroît rapidement et la variabilité de l'ensemble final de poids peut être préoccupante.

Selon Rao et Lohr (2025), Montanari et Ranalli (2009) proposent d'assouplir certaines des contraintes de calage pour le groupe (iii), afin de respecter les restrictions d'étendue. Une autre approche consiste à tirer parti d'une caractéristique supplémentaire du calage jusqu'ici relativement ignorée, à savoir son utilisation dans la vérification des données. Dans ce cas, supposons que w_i représente l'ensemble des poids obtenus à la fin du calage (simultané ou consécutif) des groupes (i) et (ii). Ensuite, on désigne w_i comme poids de départ pour le calage assisté par modèle pour la variable d'enquête y_1, de sorte qu'on obtient $w_{1i} = w_i g_{1i}$, si $\sum_{i\in\mathcal{S}} w_{1i}\hat{m}_1(\mathbf{x}_i) = \sum_{i=1}^{N} \hat{m}_1(\mathbf{x}_i)$. On prévoit donc que ces poids soient très efficaces pour obtenir une estimation du total de la population de y_1; par contre, si les ajustements g_{1i} sont importants, ils risquent d'avoir un effet sur les estimations d'autres variables, en particulier dans les domaines. Pour éviter cette possibilité, on peut notamment associer les ajustements g_{1i} aux valeurs de l'échantillon y_1 plutôt qu'aux poids, afin d'obtenir des valeurs *modifiées* $y_{1i}^* = g_{1i} y_{1i}$, et de conserver w_i comme système de poids à appliquer à toutes les variables. En fait,

$$\hat{Y}_{1,mc} = \sum_{i\in\mathcal{S}} w_{1i} y_{1i} = \sum_{i\in\mathcal{S}} w_i g_{1i} y_{1i} = \sum_{i\in\mathcal{S}} w_i y_{1i}^* = \hat{Y}_{1,c}^*.$$

Cette équation peut être reproduite pour chacune des P variables d'enquête clés. Cette approche, utilisée par Cantarella, Neri et Ranalli (2024), leur a permis de compenser la déclaration erronée de la richesse et de ses composantes dans l'enquête sur le comportement financier et de consommation des ménages de l'Eurosystème (Eurosystem Household Finance and Consumption Survey); elle est similaire, dans son esprit, au calage inversé proposé par Chambers et Ren (2004) pour rectifier les valeurs d'enquête aberrantes.

Très récemment, on a proposé d'avoir recours au calage (et particulièrement au calage assisté par un modèle) comme cadre général optionnel pour la pondération par l'inverse de la probabilité et pour obtenir une estimation doublement robuste de l'inférence causale au moyen de données d'observation (Zhang, Han et Wu, 2023; Huang, Wu et Zeng, 2024 et 2025).

4. Estimation sur petits domaines

La demande croissante d'estimations opportunes et détaillées, associée à la nécessité de réduire le fardeau des répondants et de comprimer les coûts, a stimulé la recherche sur les méthodes d'estimation sur petits domaines et leurs applications. À cet égard, je reconnaîtrais que l'estimation sur petits domaines a joué le rôle du cheval de Troie, car elle a autorisé l'utilisation des estimations basées sur un modèle dans l'inférence d'une population finie, notamment pour la production de statistiques officielles. Les orientations adoptées par les chercheurs dans ce domaine ont été nombreuses et j'en trouve quelques-unes particulièrement intrigantes, surtout du point de vue des perspectives qu'elles offrent pour la recherche future.

La première approche vise à examiner les méthodes d'estimation sur petits domaines pour découvrir les variables d'intérêt ne pouvant être mesurées directement. Il s'agit notamment de phénomènes comme le bien-être, la privation, l'invalidité, l'intégration sociale, les lacunes en matière d'éducation ou la fracture numérique : ceux-ci ne sont pas directement observables, mais on peut les mesurer en examinant de multiples éléments connexes. Ces éléments peuvent être combinés selon une règle déterministe pour obtenir une mesure finale de la variable d'intérêt. Les approches normalisées comprennent également la création d'un tableau de bord d'indicateurs, pour le calcul d'une moyenne pondérée des indicateurs au niveau local ou, plus généralement, d'un indicateur agrégé. Par ailleurs, les modèles à variables latentes – comme ceux utilisés dans l'analyse factorielle, la théorie de la réponse d'item et l'analyse de classe latente – peuvent servir à extraire des variables non observées sous-tendant l'ensemble des variables évidentes. Les valeurs des variables latentes peuvent ensuite être attribuées à toutes les unités échantillonnées à l'aide de prédictions *a posteriori* déterminées en fonction des données observées.

Si la taille des échantillons est suffisamment grande, ces variables latentes prédites pourraient être utilisées pour calculer des estimations directes à partir de la moyenne du domaine. Ainsi, l'estimation crée donc un problème double : elle nécessite d'abord une modélisation par variables latentes pour reconnaître les résultats latents provenant des éléments observés, mais nécessite ensuite des méthodes d'estimation sur petits domaines pour tenir compte des tailles d'échantillon limitées disponibles dans ces domaines. Dans les travaux de recherche antérieurs sur les résultats latents de l'estimation sur petits domaines, on a souvent abordé la question en deux étapes : on faisait d'abord une estimation concernant ces variables, puis on les traitait comme des variables réponses (comme si elles avaient été observées) dans un modèle d'estimation sur petits domaines. Cette approche risque cependant de produire des estimations de paramètres biaisées pour les relations entre la variable latente et les covariables, de sorte que les erreurs se propagent de la première étape à la seconde sans le moindre contrôle. Il devient donc très difficile d'obtenir une mesure fiable de l'incertitude, dans l'estimation sur petits domaines finale. Fabrizi, Montanari et Ranalli (2016) mentionnent une exception, dans laquelle le problème de la classification de la population et de l'obtention d'estimations sur petits domaines se fait en une seule étape dans un modèle hiérarchique bayésien où la probabilité d'appartenir à une classe latente unique varie en fonction des covariables. On présente alors un modèle de régression à niveaux unitaires de classes latentes pour extraire et estimer les décomptes de petites zones d'invalidité à partir de données d'enquête sur neuf éléments catégoriques ordonnés. Giovinazzi et Cocchi (2021) adoptent une approche similaire pour mesurer l'intégration sociale des élèves de deuxième génération dans le système scolaire italien. Moretti, Shlomo et Sakshaug (2021) utilisent des modèles d'analyse factorielle pour extraire des données sur le bien-être économique de quatre variables (privation matérielle sévère, revenu disponible équivalent, propriété du logement, densité de logements) et appliquer des modèles à niveaux unitaires en utilisant les scores factoriels comme variables de résultats. Ranalli et Bertarelli (2025) ont récemment analysé un ensemble de 33 indicateurs binaires recueillis dans le cadre de l'enquête sur les activités de la vie quotidienne menée en Italie, en se concentrant sur un échantillon de personnes âgées de 15 à 29 ans. Ils ont proposé l'utilisation de modèles multidimensionnels découlant de la théorie de la réponse d'item pour extraire les dimensions latentes des lacunes sur le plan des études. Pour

estimer ces dimensions latentes au niveau local, ils intègrent l'estimation sur petits domaines dans le modèle multidimensionnel de la théorie de la réponse d'item en permettant que les valeurs des facteurs latents varient en fonction des covariables et des effets aléatoires propres au domaine.

Les modèles à variables latentes peuvent également être appliqués à l'intégration des données lorsque la variable d'intérêt pourrait, en principe, être observée, mais que des enquêtes différentes ou des données administratives ne rendent disponibles que deux mesures indirectes (ou plus), sans qu'aucune puisse servir de point de référence définitif. C'est ce qui arrive, par exemple, dans le cas de la population active en Italie, où différentes mesures sont obtenues pour les mêmes unités à partir de l'enquête sur la population active et des données administratives fournies par la sécurité sociale et les autorités fiscales. Boeschoten, Filipponi et Varriale (2021) utilisent des modèles de Markov cachés, c'est-à-dire une extension longitudinale de l'analyse de modèles à classes latentes, pour classer les unités et exploiter pleinement la structure longitudinale des données. Filipponi, Guarnera et Varriale (2024) s'appuient sur cette approche en ayant recours à divers modèles de Markov cachés pour évaluer l'exactitude des sources administratives disponibles et générer des statistiques d'emploi à l'échelle régionale et nationale pour l'ensemble de la population.

Une autre orientation de recherche prometteuse consiste à utiliser des méthodes d'estimation sur petits domaines pour intégrer des données probabilistes et des données non probabilistes (généralement de grande taille), quand celles-ci représentent la principale source de données, comme on peut le lire dans la section 5 de l'article de Rao et Lohr (2025). Dans de tels cas, les modèles d'estimation sur petits domaines sont principalement utilisés pour corriger le biais de sélection. Schirripa Spagnolo, Bertarelli, Summa, Scannapieco, Pratesi, Marchetti et Salvati (2024) abordent par exemple le problème de l'estimation de la sensibilité des entreprises italiennes aux objectifs de développement durable des Nations Unies à l'échelle locale à l'aide de données moissonnées du Web (un échantillon non probabiliste), puisque cette variable n'est pas étudiée dans le cadre des enquêtes officielles. Pour atténuer le biais de sélection des estimations découlant de ces données non probabilistes, ils font appel à des variables auxiliaires d'une enquête probabiliste pour élaborer un estimateur doublement robuste qui combine i) la pondération par la propension afin d'accroître la représentativité de l'échantillon non probabiliste et ii) un modèle statistique conçu pour prédire les unités non incluses dans cet échantillon. Les deux modèles tiennent compte des effets aléatoires de domaine.

Bien que l'augmentation de la demande d'estimations sur petits domaines soit attribuable en partie aux exigences en matière de surveillance et d'établissement de politiques, les ouvrages publiés portant sur l'analyse causale à l'échelle locale sont limités. Ranjbar, Salvati et Pacini (2023) ont été les premiers à remédier à cette lacune en estimant les effets de traitement moyen propres aux domaines que l'on retrouvait dans des études par observation. Ils examinaient plus précisément les répercussions du traitement binaire en se basant sur le modèle à résultats potentiels d'abord créé par Rubin (1980). Pour ce faire, ils utilisent le modèle de régression à erreurs emboîtées au niveau de l'unité de Battese, Harter et Fuller (1988) et le modèle de régression M-quantile de Chambers et Tzavidis (2006), afin d'estimer les scores de propension et les résultats non observés pour la population. Pour l'estimation des effets de traitement moyen propres aux domaines, ils proposent une modification de l'estimateur de pondération par l'inverse de la probabilité en

fonction des scores de propension estimés et des résultats prédits. Leurs estimateurs conservent les propriétés d'estimateurs doublement robustes; en d'autres termes, ils demeurent convergents même si l'un des deux modèles (le modèle du score de propension ou celui de prédiction des résultats) affiche une erreur de spécification.

5. Conclusion

L'évolution constante des méthodes d'enquête témoigne de la complexité croissante des sources de données modernes et des outils d'analyse de données, en plus de la demande toujours grandissante d'estimations actuelles et détaillées. Les professeurs Rao et Lohr ont fourni une synthèse remarquable de ces tendances, et ont ainsi offert une évaluation exhaustive permettant de mieux comprendre les défis et les possibilités actuels en ce qui concerne les plans d'enquête et l'estimation. L'accent qu'ils mettent sur les méthodes innovantes, y compris l'apprentissage automatique et l'intégration des données, met en évidence l'adaptabilité de la théorie et de la méthodologie des enquêtes par sondage pour qui veut profiter des technologies émergentes et de la diversité des sources de données.

Cette discussion s'est concentrée sur trois domaines spécifiques mis en évidence par Rao et Lohr (2025), à savoir l'échantillonnage équilibré, le calage et l'estimation sur petits domaines. Les principales caractéristiques de ces sujets ont été soulignées pour relever les défis émergents, tels que l'intégration de techniques de régression modernes et de nouvelles sources de données, ainsi que leur application dans la vérification et l'intégration de données et l'inférence causale. Ces approches tirent avantage de l'utilisation efficace de l'information auxiliaire et des modèles statistiques. Le pouvoir prédictif des variables auxiliaires est essentiel pour rehausser l'efficacité ou rectifier le biais d'échantillonnage. Des approches statistiques plus sophistiquées, comme celles qui reposent sur l'apprentissage automatique ou les méthodes doublement robustes ou multirobustes, peuvent être utiles pour extraire tous (et seulement) les renseignements disponibles. La prudence s'impose toutefois, car certaines études – en particulier celles qui reposent sur des panels volontaires – ont révélé que le recours à des méthodes statistiques sophistiquées n'améliore pas nécessairement la qualité des estimations : les nouvelles sources de données peuvent être à tel point imparfaites qu'elles rendent impossible la correction par la modélisation statistique (Dutwin et Buskirk, 2017; Kennedy, Mercer et Lau, 2024). Par conséquent, il est essentiel d'effectuer des diagnostics solides et d'évaluer en profondeur la qualité des variables auxiliaires pour assurer la validité statistique de la stratégie d'échantillonnage globale.

Remerciements

Ces travaux ont été réalisés grâce au soutien de l'Union européenne dans le cadre de son plan de relance NextGenerationEU et, plus précisément, de la composante 1 de la mission 4 CUP J33C 22002910001 – Projet INCLUDE-ALL : inclusion sociale et estimation des indicateurs liés au numérique à l'échelle locale.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Boeschoten, L., Filipponi, D. et Varriale, R. (2021). Combining multiple imputation and hidden Markov modeling to obtain consistent estimates of employment status. *Journal of Survey Statistics and Methodology*, 9(3), 549-573.

Breidt, F.J., et Chauvet, G. (2012). Penalized balanced sampling. *Biometrika*, 99(4), 945-958.

Breidt, F.J., Opsomer, J.D. et Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *Journal of the American Statistical Association*, 111(514), 822-833.

Cantarella, M., Neri, A. et Ranalli, M.G. (2024). Estimating the distribution of household wealth: methods for adjusting survey data estimates using national accounts and rich list data. *Review of Income and Wealth*, 70(3), 551-580.

Chambers, R.L., et Ren, R. (2004). Outlier robust imputation of survey data. *Proceedings of the Survey Research Methods Section,* American Statisical Association, 3336-3344.

Chambers, R., et Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2), 255-268.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Dutwin, D., et Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81(S1), 213-239.

Fabrizi, E., Montanari, G.E. et Ranalli, M.G. (2016). A hierarchical latent class model for predicting disability small area counts from survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179(1), 103-131.

Filipponi, D., Guarnera, U. et Varriale, R. (2024). Multi-source inference via mixture of hidden Markov models: Application to regional labour statistics in italy. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae111.

Giovinazzi, F., et Cocchi, D. (2021). Social integration of second generation students in the italian school system. *Social Indicators Research*, 1-21.

Huang, J., Wu, C. et Zeng, L. (2024). Pseudo-empirical likelihood methods for causal inference. *arXiv preprint arXiv:2401.06919*.

Huang, J., Wu, C. et Zeng, L. (2025). Sample empirical likelihood methods for causal inference. *Canadian Journal of Statistics,* à paraître.

Kennedy, C., Mercer, A. et Lau, A. (2024). Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi. *Techniques d'enquête*, 50(1), 5-26. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00013-fra.pdf.

Kott, P. (2007). Randomisation-assisted model-based survey sampling. *Session IPM 55*. ISI 56th WSC, Lisboa.

Montanari, G.E., et Ranalli, M.G. (2009). La calibration à des modèles multiples. Atelier sur le calage et l'estimation dans les enquêtes : Recueil. Statistique Canada. Disponible à l'adresse https://www150.statcan.gc.ca/n1/fr/catalogue/11-536-X200900110807.

Moretti, A., Shlomo, N. et Sakshaug, J.W. (2021). Small area estimation of latent economic well-being. *Sociological Methods & Research*, 50(4), 1660-1693.

Ranalli, M.G., et Bertarelli, G. (2025). Small area estimation of educational poverty using item response theory models. *Methodological and Applied Statistics and Demography I, SIS 2024, Short Papers, Plenary and Specialized Sessions*, Italian Statistical Society Series on Advances in Statistics, Suisse: Springer Nature.

Ranjbar, S., Salvati, N. et Pacini, B. (2023). Estimating heterogeneous causal effects in observational studies using small area predictors. *Computational Statistics & Data Analysis*, 184, 107742.

Rao, J.N.K., et Lohr, S.L. (2025). Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage. *Techniques d'enquête*, 51(1), 87-122. Article accessible à l'adresse http://www.statcan.gc.ca/ n1/fr/pub/12-001-x/2025001/article/00014-fra.pdf.

Rubin, D.B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371), 591-593.

Ruppert, D., Wand, M.P. et Carroll, R. (2003). *Semiparametric Regression*. New York: Cambridge University Press, Cambridge.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schirripa Spagnolo, F., Bertarelli, G., Summa, D., Scannapieco, M., Pratesi, M., Marchetti, S. et Salvati, N. (2024). Inference for big data assisted by small area methods: An application on sustainable development goals sensitivity of enterprises in italy. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae115.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

Zhang, S., Han, P. et Wu, C. (2023). Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *Revue Internationale de Statistique*, 91(2), 165-192.