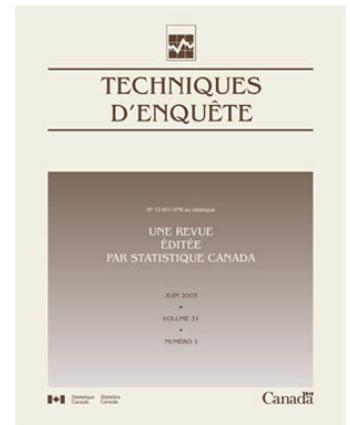


## Techniques d'enquête

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

par Jean D. Opsomer, Daifeng Han et Medha Uppala

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](http://www.statcan.gc.ca).

Vous pouvez également communiquer avec nous par :

**Courriel** à [infostats@statcan.gc.ca](mailto: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](http://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.*

---

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

Jean D. Opsomer, Daifeng Han et Medha Uppala<sup>1</sup>

## Résumé

Dans le présent exposé, nous complétons l'excellent aperçu fourni par les professeurs Lohr et Rao en abordant quelques sujets supplémentaires. Le premier sujet constitue un appel à une plus grande reconnaissance du rôle central de la modélisation dans l'estimation par enquête. Le second est un bref exposé sur l'utilisation de renseignements partiels sur la base de sondage dans le plan d'enquête. Finalement, nous attirons l'attention sur les augmentations récentes des méthodes synthétiques, en particulier la régression multiniveau et la poststratification (RMP) dans les applications d'estimation sur petits domaines.

**Mots-clés :** Inférence basée sur le plan de sondage; inférence basée sur un modèle; plan d'échantillonnage; régression multiniveau et poststratification (RMP).

## 1. Introduction

Nous félicitons les professeurs Lohr et Rao pour l'excellent aperçu opportun de l'état des statistiques d'enquête, démontrant qu'il s'agit toujours d'un domaine de recherche dynamique. Dans le présent exposé, nous mettons en évidence trois sujets qui complètent l'aperçu des auteurs : le rôle accru de la modélisation dans l'estimation d'enquête, l'application de renseignements partiels sur la base de sondage dans le plan d'enquête et la l'augmentation récente des méthodes synthétiques dans les applications d'estimation sur petits domaines.

## 2. Estimation fondée sur le plan de sondage généralisé : nous sommes maintenant tous des modélisateurs

Traditionnellement, l'estimation fondée sur le plan de sondage a deux propriétés principales : 1) la *représentativité* : les estimateurs sont statistiquement représentatifs de la population d'où provient l'échantillon, et 2) l'indépendance du modèle : l'inférence est possible sans recourir aux hypothèses du modèle. Ces deux propriétés sont garanties dans le cadre purement fondé sur la randomisation qui est à la base de l'inférence fondée sur le plan de sondage. Elles continuent aussi d'être vraies à l'intérieur du paradigme « assisté par un modèle », dans lequel les contributions du modèle sont pleinement saisies par la distribution de randomisation des estimateurs. En fait, les propriétés 1 et 2 sont souvent traitées comme interchangeables, la représentativité des estimateurs étant justifiée par le fait que la randomisation a été utilisée pour obtenir les données. C'est ce qu'exprime le terme « fondé sur le plan de sondage », que nous utilisons encore malgré

---

1. Jean D. Opsomer, Daifeng Han et Medha Uppala, Westat, Inc., 1600 Research Blvd., Rockville (Maryland), 14850, États-Unis. Courriel : [jeanopsomer@westat.com](mailto:jeanopsomer@westat.com).

le fait qu'en situation de non-réponse (qui peut maintenant atteindre 90 % pour certains modes d'enquête), les hypothèses du modèle et les approches de modélisation deviennent toutes deux une partie intrinsèque de la construction et de la distribution des estimateurs, c'est-à-dire que 2 ne tient plus.

Le non-respect de 2 n'invalide pas automatiquement 1, bien que le fait de prétendre que les estimateurs sont statistiquement représentatifs sans avoir recours à 2 exige une justification plus complète et plus rigoureuse. Comme le démontrent les auteurs dans leur article de synthèse, des méthodes statistiques sophistiquées ont été mises au point par la communauté de recherche des statistiques d'enquête, non seulement pour tenir compte de la non-réponse totale et partielle, mais aussi pour combiner les données de différentes enquêtes; elles permettent même d'intégrer les données d'enquête et les données autres que des données d'enquête. Par conséquent, la distinction entre « fondé sur le plan de sondage » et « fondé sur un modèle » n'est plus significative dans le sens de faire la distinction entre les méthodes et le cadre d'inférence utilisés pour créer des estimations d'enquête.

Cependant, il y a un certain nombre de caractéristiques clés de l'inférence des enquêtes qui la différencient tout de même des autres domaines de statistiques. En tant que statisticiens ayant une approche « fondée sur le plan de sondage généralisé », notre but inférentiel continue d'être de fournir des données qui permettent de décrire ou d'estimer des caractéristiques de  $\mathcal{U} = \{\mathbf{y}_i, i \in U\}$  en fonction des données-échantillons  $\mathcal{S} = \{\mathbf{y}_i, i \in S\}$ , où  $\mathbf{y}_i$  désigne un vecteur de variables pour l'unité de population  $i$ . Cela est contraire aux statisticiens ayant une approche « fondée sur un modèle », qui font généralement des inférences sur un modèle sous-jacent pour la variable aléatoire  $\mathbf{Y}$  observée dans l'échantillon, indiqué par  $\mathcal{F}_s(\mathbf{Y})$  (bien qu'ils pourraient faire des prétentions inférentielles à propos d'un modèle de population  $\mathcal{F}_U(\mathbf{Y})$  aussi). Même si les statisticiens ayant une approche fondée sur le plan de sondage généralisé doivent continuer de tenir compte du plan de sondage initial, l'inférence pour  $\mathcal{U}$  exige de modéliser le processus de sélection au complet, y compris toutes les formes de non-observation (y compris la non-réponse, l'absence d'emplacement, les erreurs de correspondance, les erreurs de base de sondage, etc.). Tant que le processus de sélection est correctement modélisé, la représentativité statistique des estimations en ce qui a trait à la population continue d'être valable, même si l'inférence n'est plus indépendante du modèle.

Les statisticiens d'enquête, tant les chercheurs que les praticiens, fonctionnent déjà dans ce cadre fondé sur le plan de sondage généralisé depuis plusieurs décennies. Cependant, l'équilibre entre les sources d'incertitude connues et inconnues (c'est-à-dire le plan d'échantillonnage et les autres mécanismes de sélection) a changé dans la mesure où les sources inconnues constituent maintenant souvent l'élément dominant de l'incertitude globale. Par conséquent, compte tenu de l'importance grandissante des modèles de sélection avec leurs hypothèses connexes dans l'estimation, il nous appartiendrait de reconnaître de manière plus explicite cette dépendance du modèle dans les renseignements fournis par les ensembles de données d'enquête. En particulier, les prétentions de « représentativité statistique » qui reposent sur des hypothèses de modèle pourraient être accompagnées d'un avis de non-responsabilité approprié, par exemple : « les poids d'enquête et les poids de rééchantillonnage reflètent le plan d'échantillonnage et les probabilités de sélections modélisées qui sont fondées sur les prédicteurs suivants : [...] En vertu de l'hypothèse selon

laquelle le modèle de sélection est correctement précisé, l'utilisation de ces poids dans l'estimation et la modélisation garantit que les résultats sont statistiquement représentatifs de la population ».

Cette transparence fera aussi comprendre aux analystes de données et aux chercheurs d'autres domaines de la statistique que les statisticiens d'enquête d'aujourd'hui n'appliquent plus seulement des méthodes fondées sur le plan de sondage; ils sont plutôt devenus aussi des modélisateurs sophistiqués. Notre expertise particulière se situe dans la modélisation des processus de sélection complexes qui donnent comme résultat les données recueillies, l'intégration de ces modèles dans la randomisation du plan d'échantillonnage et la création de méthodes d'estimation efficaces et représentatives. Cette expertise continue de représenter une contribution absolument importante à la science des statistiques. La vaste portée des avancées méthodologiques abordées par les professeurs Lohr et Rao démontre que nous sommes en bonne posture pour relever les défis actuels et futurs inhérents au fait de fournir une inférence statistiquement valide pour les populations visées.

### **3. Utilisation d'information auxiliaire incomplète dans l'échantillonnage**

Dans la section consacrée au plan d'enquête, les professeurs Lohr et Rao mentionnent que la répartition de l'échantillon continue d'être un sujet de recherche important. Nous sommes d'accord et souhaitons mettre brièvement en évidence un type particulier de répartition qui devient de plus en plus utile dans l'échantillonnage des ménages lorsqu'il n'y a aucun accès à une base de sondage de grande qualité. Dans cette situation, l'échantillonnage peut être très inefficace si l'objectif est de sélectionner des sous-populations d'intérêt (par exemple les ménages comportant des enfants) ou de suréchantillonner certaines sous-populations pour accroître la précision pour des estimations par domaine (par exemple les estimations pour des groupes de minorité raciale ou ethnique rares). Aux États-Unis, l'échantillonnage basé sur l'adresse est devenu largement utilisé au cours de la dernière décennie. La base de sondage est formée de toutes les adresses postales aux États-Unis, mais elle manque de renseignements sur les caractéristiques des ménages à ces adresses. Les fournisseurs de bases de sondage fondées sur l'adresse peuvent ajouter des caractéristiques des adresses comme le nombre d'adultes vivant à l'adresse, l'âge du responsable du ménage, l'origine hispanique et la présence d'enfants (Valliant, Hubbard, Lee et Chang, 2014; Roth, Caporaso et DeMatteis, 2022), mais la qualité de cette information varie et n'est disponible que pour un sous-ensemble des adresses.

Une approche générale à l'égard de l'utilisation de tels renseignements pour cibler une sous-population particulière est de former des strates d'échantillonnage de « grande densité » et de « faible densité » au moyen des variables de la base de sondage disponibles. Les unités d'échantillonnage qui comportent des renseignements manquants dans la base de sondage sont automatiquement classées dans la strate de faible densité. Il est alors possible d'allouer l'échantillon de sorte que la strate de « grande densité » soit suréchantillonnée, avec une réduction correspondante du taux d'échantillonnage dans la strate de « faible densité ». Il est clair que l'on peut s'attendre à ce que la sous-population visée soit présente dans les deux

strates, dans des proportions différentes. Deux conditions sont requises pour que cette approche améliore l'efficacité de l'estimation (Waksberg, Judkins et Massey, 1997). Premièrement, dans la strate de « grande densité », il doit y avoir une proportion élevée de cas qui appartiennent à la sous-population ou au domaine d'intérêt. Deuxièmement, parmi tous les cas qui appartiennent à la sous-population ou au domaine, ceux qui se trouvent dans la strate de « grande densité » doivent représenter une proportion suffisamment élevée de leur total de population. À moins que les renseignements du fournisseur ne soient de qualité suffisante, il est possible de déroger à une condition ou aux deux.

Tandis que les renseignements au niveau de la région (par exemple le groupe d'îlots de recensement ou le secteur de recensement) et au niveau de l'adresse peuvent être utilisés pour un tel ciblage, un avantage important d'utiliser les renseignements au niveau de l'adresse est que la sous-population n'a pas besoin de faire l'objet d'un groupement géographique (Chen et Kalton, 2015); Dutwin, Coyle, Lerner, Bilgen et English, 2024). Il est aussi possible d'utiliser la modélisation prédictive avec les caractéristiques au niveau de la région en combinaison avec les variables ajoutées par le fournisseur d'échantillonnage basé sur l'adresse aux fins de stratification et de suréchantillonnage. Une approche prometteuse dans cette direction est la méthode bayésienne reposant sur le nom de famille et le géocodage (Bayesian Surname and Geocoding (BSG)) pour permettre un suréchantillonnage des groupes de minorité raciale et de minorité ethnique. La méthode BSG et ses variants s'appuient sur des renseignements géographiques (par exemple le secteur de recensement) ainsi que sur le nom de famille pour prédire un ensemble de probabilités de faire partie de cinq groupes raciaux ou ethniques (c'est-à-dire Blancs, Noirs, Hispaniques, Asiatiques et autres) pour chaque enregistrement (Elliott, Morrison, Fremont, McCaffrey, Pantoja et Lurie, 2009; Khanna, Bertelsen, Rosenman, Olivella et Imai, 2022).

Comme indiqué, l'information auxiliaire au niveau de l'adresse provient de nombreuses sources dont la qualité et l'intégralité varient. La qualité des données de cette information auxiliaire a fait l'objet d'évaluations, mais peu d'entre elles permettent de savoir si l'information auxiliaire était assez précise pour obtenir un ciblage efficace (Battaglia, Dillman, Frankel, Harter, Buskirk, McPhee, DeMatteis et Yancey, 2016; Dutwin et coll., 2024). Il n'y a qu'une exception : Roth et coll. (2022), qui ont traité du compromis entre une taille d'échantillon nominal et l'effet de plan de sondage lors de l'utilisation d'information auxiliaire imparfaite pour cibler des sous-populations d'intérêt. Puisque le but du ciblage est d'améliorer la précision des estimations selon un coût fixe, il est essentiel d'examiner la taille de l'échantillon *efficace* (calculée comme la taille de l'échantillon nominal divisée par l'effet de plan de sondage) pour les estimations par domaine et les estimations globales.

#### **4. Régression multiniveau avec poststratification**

L'estimation sur petits domaines est un sujet de recherche actif et bien établi parmi les statisticiens d'enquête, alors qu'elle repose sur des approches de modélisation sophistiquées au niveau du domaine et au niveau de l'unité qui se prêtent à un grand nombre d'applications. Ces estimateurs sur petits domaines peuvent souvent s'exprimer sous la forme de combinaisons d'estimateurs synthétiques et directs et

possèdent l'importante propriété selon laquelle, à mesure que la taille de l'échantillon dans les petits domaines augmente, la contribution des estimateurs directs augmente, c'est-à-dire que les estimateurs sur petits domaines convergent vers les estimateurs directs. Dans le cadre d'un développement relativement nouveau en dehors de notre communauté de recherche, la régression multiniveau avec poststratification (RMP) est de plus en plus acceptée en tant que méthode de rechange d'estimation sur petits domaines, plus particulièrement en matière de santé publique et dans les domaines connexes. Voir par exemple Zhang, Onufrak, Holt et Croft (2013), Zhang, Holt, Lu, Wheaton, Ford, Greenlund et Croft (2014), Zhang, Holt, Yun, Lu, Greenlund et Croft (2015), Davila-Payan, DeGuzman, Johnson, Serban et Swann (2015), Zgodic, Eberth, Breneman, Wende, Kaczynski, Liese et McLain (2021) et Wang, Tevendale, Lu, Cox, Carlson, Li, Shulman, Morrow, Hastings et Barfield (2022).

Introduite à l'origine par Gelman et Little (1997) comme régression logistique hiérarchique avec poststratification, le but était d'améliorer les ajustements par poststratification avec estimateur de la méthode itérative du quotient en produisant des estimations pour un vaste ensemble de domaines plutôt que pour un ensemble limité de catégories. Bien que similaires aux estimateurs sur petits domaines traditionnels au niveau de l'unité, les estimateurs de RMP sont synthétiques et il n'est pas garanti qu'ils soient proches des estimateurs directs, même en ayant des tailles d'échantillon de grands domaines. Néanmoins, compte tenu de leur utilisation accrue dans les applications d'estimation sur petits domaines, nous croyons qu'il est utile que les statisticiens d'enquête soient au courant de cette avancée.

Nous décrivons brièvement la méthode de RMP dans un contexte d'estimation sur petits domaines simple au niveau de l'unité. Soit  $y_{ij}$  qui représente une variable cible au niveau de l'unité, avec l'unité  $j = 1, \dots, N_i$  à l'intérieur du petit domaine  $i = 1, \dots, m$ , et un vecteur de covariable catégorielle correspondant  $\mathbf{x}_{ij}$  observé dans l'échantillon. Le nombre d'unités d'échantillonnage dans une cellule définie par l'intersection entre tous les niveaux des covariables dans chaque petit domaine est connu à partir d'une source externe.

La première étape de la RMP consiste à établir un modèle de régression de  $y$  sur  $\mathbf{x}$  au moyen des données-échantillons. Comme il reflète la portion « multiniveau » de la RMP, ce modèle comprendra souvent des effets aléatoires, qui visent à améliorer le degré d'ajustement global et la stabilité du modèle et en particulier, n'ont pas à correspondre aux petits domaines. Une fois que les estimations pour tous les paramètres du modèle sont obtenues, par méthodes fréquentistes ou bayésiennes, le modèle est utilisé pour prédire la valeur attendue de  $y$  pour chaque combinaison des niveaux des covariables dans chaque petit domaine. La deuxième étape de la RMP, la poststratification, regroupe les prédictions de  $y$  en établissant la moyenne des cellules en proportion de leurs totaux connus dans chaque petit domaine.

Un avantage pratique de la RMP est qu'elle permet de prédire  $y$  pour les valeurs de  $\mathbf{x}$  qui n'ont pas été échantillonnées ou qui présentent de très petites tailles d'échantillon. Cependant, l'hypothèse implicite de la RMP d'un plan d'enquête ignorable est un obstacle de taille. L'incorporation d'échantillons en grappe, la disponibilité des chiffres de population et la détermination de solides prédicteurs au niveau du groupe sont d'autres obstacles pour les régressions multiniveaux. Une littérature croissante sur la RMP porte sur certains

de ces obstacles. De récents travaux réalisés par Gelman, Si et West (2024) incorporent des poids d'échantillonnage en tant que régresseurs et estiment la répartition de la population des poids pour éventuellement prédire  $y$  dans chaque cellule; ces prédictions fondées sur des enquêtes pondérées de  $y$  sont ensuite stratifiées *a posteriori* en calculant la moyenne sur les chiffres de population connus.

## Bibliographie

Battaglia, M.P., Dillman, D.A., Frankel, M.R., Harter, R., Buskirk, T.D., McPhee, C.B., DeMatteis, J.M. et Yancey, T. (2016). Sampling, data collection, and weighting procedures for address-based sample surveys. *Journal of Survey Statistics and Methodology*, 4, 476-500.

Chen, S., et Kalton, G. (2015). Geographic oversampling for race/ethnicity using data from the 2010 U.S. population census. *Journal of Survey Statistics and Methodology*, 3, 543-565.

Davila-Payan, C., DeGuzman, M., Johnson, K., Serban, N. et Swann, J. (2015). Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. *Preventing Chronic Disease*, 12, 140229.

Dutwin, D., Coyle, P., Lerner, J., Bilgen, I. et English, N. (2024). Leveraging predictive modelling from multiple sources of big data to improve sample efficiency and reduce survey nonresponse error. *Journal of Survey Statistics and Methodology*, 12, 435-457.

Elliott, M.N., Morrison, P.A., Fremont, A.M., McCaffrey, D.F., Pantoja, P.M. et Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9, 69-83.

Gelman, A., et Little, T.C. (1997). [Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3616-fra.pdf). *Techniques d'enquête*, 23(2), 135-145. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3616-fra.pdf>.

Gelman, A., Si, Y. et West, B.T. (2024). MRPW: Regression, poststratification and small-area estimation with sampling weights. Rapport technique, Columbia University. Disponible sur [http://www.stat.columbia.edu/~gelman/research/unpublished/weight\\_regression.pdf](http://www.stat.columbia.edu/~gelman/research/unpublished/weight_regression.pdf).

Khanna, K., Bertelsen, B., Rosenman, E., Olivella, S. et Imai, K. (2022). wru: Who are you? Bayesian prediction of racial category using surname and geolocation. R package version 1.0.0. Disponible sur <https://cran.r-project.org/web/packages/wru/index.html>.

- Roth, S., Caporaso, A. et DeMatteis, J. (2022). Variables appended to abs frame: Has their data quality improved? *PLoS ONE* 17, e0269110.
- Valliant, R., Hubbard, F., Lee, S. et Chang, C. (2014). Efficient use of commercial lists in U.S. household sampling. *Journal of Survey Statistics and Methodology*, 2, 182-209.
- Waksberg, J., Judkins, D. et Massey, J.T. (1997). [Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis](#). *Techniques d'enquête*, 23(1), 69-80. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3107-fra.pdf>.
- Wang, Y., Tevendale, H., Lu, H., Cox, S., Carlson, S.A., Li, R., Shulman, H., Morrow, B., Hastings, P.A. et Barfield, W.D. (2022). U.S. county-level estimation for maternal and infant health-related behavior indicators using pregnancy risk assessment monitoring system data, 2016-2018. *Population Health Metrics*, 20, 14.
- Zgodic, A., Eberth, J.M., Breneman, C.B., Wende, M.E., Kaczynski, A-T., Liese, A.D. et McLain, A.C. (2021). Estimates of childhood overweight and obesity at the region, state, and county levels: A multilevel small-area estimation approach. *American Journal of Epidemiology*, 190, 2618-2629.
- Zhang, X., Holt, J.B., Lu, H., Wheaton, A.G., Ford, E.S., Greenlund, K.J. et Croft, J.B. (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, 179, 1025-1033.
- Zhang, X., Holt, J.B., Yun, S., Lu, H., Greenlund, K.J. et Croft, J.B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, 182, 127-137.
- Zhang, X., Onufrak, S., Holt, J.B. et Croft, J.B. (2013). A multilevel approach to estimating small area childhood obesity prevalence at the census block-group level. *Preventing Chronic Disease*, 10, 120252.