

Programme de recherche et développement en méthodologie : réalisations, 2025-2026

Date de diffusion : le 8 octobre 2026



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, 2026

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.

**Programme de recherche
et
développement en méthodologie**

Réalisations, 2025-2026

Le présent rapport fait la synthèse des réalisations en 2025-2026 du Programme de recherche et développement en méthodologie (PRDM) de la Direction des méthodes statistiques modernes et de la science des données de Statistique Canada. Ce programme couvre un éventail d'activités de recherche et développement dans les méthodes statistiques et de science des données susceptibles d'être appliquées à grande échelle aux programmes statistiques de l'organisme; ce sont des activités qui, autrement, ne seraient vraisemblablement pas menées dans le cadre des services réguliers de méthodologie offerts à ces programmes. Ajoutons que, dans le but de promouvoir l'utilisation des résultats des travaux de recherche et de développement, le PRDM comporte des activités de soutien aux clients pour la mise en application de travaux de développement antérieurs fructueux. Des renseignements supplémentaires sur chacun des projets décrits peuvent être obtenus des personnes-ressources mentionnées. Pour en savoir davantage sur le PRDM dans son ensemble, veuillez communiquer avec :

Jean-François Beaumont

(Courriel : jean-francois.beaumont@statcan.gc.ca)

Programme de recherche et développement en méthodologie

Réalisations, 2025-2026

Table of Contents

1	Intégration de données.....	3
1.1	Intégration de données d'échantillons probabilistes et non probabilistes.....	3
1.2	Couplage d'enregistrements	4
1.3	Estimation sur petits domaines	6
2	Méthodes et applications de la science des données	11
3	Problèmes d'estimation dans les enquêtes	19
4	Confidentialité et accès aux données	27
5	Soutien (centres de ressources).....	29
5.1	Centre de recherche et d'analyse en séries chronologiques	29
5.2	Centre de ressources sur l'innovation et les outils de statistiques économiques	32
5.3	Centre de ressources en couplage d'enregistrements.....	35
5.4	Centre de ressources en analyse de données	36
5.5	Centre de ressources pour la confidentialité et l'accès aux données	37
5.6	Centre de ressources en conception de questionnaires	38
5.7	Secrétariat de l'éthique des données.....	38
5.8	Secrétariat de la qualité.....	40
6	Autres activités	41
6.1	Revue <i>Techniques d'enquête</i>	41
6.2	Transfert de connaissances – Formation en statistique.....	42
6.3	Symposium international sur les questions de méthodologie de Statistique Canada.....	43
7	Documents de recherche parrainés par le Programme de recherche et développement en méthodologie.....	43

1 Intégration de données

1.1 Intégration de données d'échantillons probabilistes et non probabilistes

PROJET : Inférence par prédiction pour des totaux ou des moyennes de population finie en l'absence, ou en présence limitée, de données d'enquête probabiliste

Depuis plusieurs années, les organismes nationaux de statistique (ONS) de plusieurs pays cherchent à réduire les coûts de collecte des données des enquêtes probabilistes ainsi que le fardeau des répondants, en tirant parti de sources de données de rechange et de techniques de prédiction modernes. Cela a conduit certains ONS à étudier le recours à des échantillons non probabilistes pour la production de statistiques officielles. À Statistique Canada, on évalue actuellement la possibilité de réduire la fréquence des enquêtes probabilistes et de remplacer les données d'enquête manquantes par des prédictions fondées sur un modèle. Par exemple, une enquête probabiliste annuelle pourrait être menée tous les deux ans, et les données d'enquête manquantes au cours des années sans enquête pourraient être remplacées par des prédictions, à condition que des données auxiliaires ainsi que des données d'entraînement soient disponibles au cours des années sans enquête (p. ex. des données historiques ou des données actuelles issues d'échantillons non probabilistes).

Dans ce contexte, nos recherches ont porté sur les deux questions suivantes :

- 1) Selon quelles méthodes et hypothèses peut-on obtenir des prédicteurs approximativement sans biais de totaux ou de moyennes de population finie lorsque les variables d'intérêt ne sont pas recueillies dans le cadre d'une enquête probabiliste?
- 2) Comment évaluer la qualité de ces prédicteurs? Plus précisément, comment estimer la variance de prédiction?

Progrès :

Dans le cadre de travaux antérieurs, nous avons élaboré une théorie, y compris des estimateurs de la variance de prédiction, pour le cas où les données d'enquête manquantes sont remplacées par des prédictions fondées sur un modèle de travail linéaire ou par des prédictions fondées sur la méthode du plus proche voisin. Cette année, nous avons mené à bien une étude par simulation comparant ces méthodes de prédiction ainsi que des prédictions fondées sur un modèle de forêt aléatoire, et avons rédigé un article (Beaumont, Bocci, Bosa, Dasylva et Maranda, 2026) qui a été présenté en juillet 2025 lors de l'Italian Conference on Survey Methodology de 2025.

Dans notre article, nous examinons également brièvement le cas où des données auxiliaires sont disponibles pour l'ensemble de la population, ainsi que des données d'enquête issues d'un petit échantillon probabiliste. Le fait d'avoir des données d'enquête nous permet de faire des inférences valides sans nous fier à des hypothèses de modèle. Une nouvelle approche, appelée inférence par prédiction (prediction-powered inference), a récemment été proposée pour gérer ce cas lorsque les observations sont indépendantes et identiquement distribuées. Elle consiste à éliminer le biais des prédictions d'apprentissage automatique en tirant parti des données d'enquête observées. L'inférence par prédiction s'avère essentiellement équivalente à l'approche assistée d'un modèle, laquelle est bien connue dans la littérature.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., Bocci, C., Bosa, K., Dasylda, A. et Maranda, G. (2026). Prediction inference for finite population totals or means. Soumis pour publication dans une revue à comité de lecture.

1.2 Couplage d'enregistrements

PROJET : Modélisation conjointe des erreurs de couplage et des enregistrements en double dans le cadre du recensement de la population canadienne

Dans le recensement de la population canadienne, on parle d'enregistrements en double lorsque certaines personnes sont dénombrées à plusieurs reprises. Bien que ces enregistrements en double soient une source de biais, ils sont difficiles à déceler sans identificateur unique. Jusqu'à maintenant, l'Étude sur le surdénombrement du recensement (ESuR) a permis de se pencher sur ce défi, car elle consiste à coupler des enregistrements du recensement et à effectuer des vérifications administratives (c.-à-d. manuelles) coûteuses afin de repérer d'éventuels cas d'enregistrements en double à partir de petits groupes d'enregistrements couplés (Statistique Canada, 2024). L'ESuR pourrait être améliorée (meilleure qualité) et menée plus rapidement (c.-à-d. à moindre coût) s'il était possible d'évaluer les enregistrements en double et les erreurs de couplage à l'aide d'un modèle statistique adéquat et sans recourir à des vérifications manuelles. Dans le cadre de travaux antérieurs, Dasylda et Goussanou (2022) ont proposé un tel modèle, qui était limité par l'hypothèse selon laquelle chaque personne ne génère au plus que deux enregistrements du recensement, c'est-à-dire que seuls les dénombrements simples ou doubles étaient permis. Toutefois, le nombre de dénombrements d'ordre supérieur n'est pas négligeable dans le recensement.

Progrès :

La méthodologie a été généralisée et améliorée afin de permettre de choisir librement le nombre maximal de dénombrements par personne et de fournir des estimations ponctuelles et par intervalle pour les enregistrements en double et les taux d'erreur de couplage (Dasylda et Goussanou, 2026). Elle donne des résultats encourageants lors des simulations et des essais réalisés à partir de données de recensements antérieurs.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Abel Dasylda (abel.dasylda@statcan.gc.ca).

Bibliographie

Dasylda, A. et Goussanou, A. (2022). [A new model for the automated identification of duplicate records](#). Actes de la réunion annuelle de la Société statistique du Canada (SSC), Section des méthodes d'enquête. Accessible à l'adresse : https://ssc.ca/sites/default/files/imce/dasylda_ssc2022.pdf.

Dasylda, A. et Goussanou, A. (2026). A joint model of record duplication and linkage error for the Canadian Census of Population. Rapport interne, Statistique Canada, Ottawa.

Statistique Canada (2024). [Rapport technique sur la couverture, Recensement de la population, 2021](#). N° 98-303-X au catalogue, numéro 2021001. Accessible à l'adresse : <https://www12.statcan.gc.ca/census-recensement/2021/ref/98-303/98-303-x2021001-fra.pdf>.

PROJET : Méthode de couplage d'enregistrements différenciellement confidentielle

Statistique Canada pourrait produire de nombreuses statistiques utiles s'il pouvait coupler ses microdonnées à celles d'autres organismes sans exiger de ces derniers qu'ils divulguent l'intégralité de leurs microdonnées dans le cadre d'une entente de partage de données, qui peut s'avérer difficile à négocier. Il faut des solutions novatrices pour coupler les microdonnées et effectuer des calculs à partir de celles-ci lorsqu'elles sont protégées par une méthode de perturbation ou par des formes particulières de chiffrement, comme le chiffrement homomorphe. Bien qu'il existe de nombreuses solutions de ce type, aucune d'entre elles n'a, à ce jour, permis de satisfaire pleinement aux exigences rigoureuses de la confidentialité différentielle, qui est devenue, dans la pratique, la norme de référence en matière de contrôle de la divulgation statistique (Dwork et Roth, 2014). En effet, chacune de ces solutions comporte au moins une étape où la confidentialité repose entièrement sur le chiffrement, qui est reconnu comme étant insuffisant pour garantir la confidentialité différentielle (Escobar, Canard, Laguillaumie et Phan, 2024). Le fait de remédier à cette limite actuelle pourrait faciliter l'adoption de ces technologies par les organismes nationaux de statistique, qui se sont montrés réticents à ce que leurs microdonnées quittent leurs locaux sans entente de partage de données, que le processus fasse appel ou non au chiffrement.

Progrès :

Une nouvelle procédure respectant entièrement la confidentialité différentielle a été mise au point (Dasylda, 2025) afin d'estimer un total d'une population finie en couplant deux sources de données, lesquelles appartiennent respectivement à deux organismes de statistique, et en calculant un total sur les couplages ainsi établis. Dans le cadre de cette procédure, qui s'inspire des travaux d'Inan, Kantarcioglu, Ghinita et Bertino (2010), chaque organisme effectue séparément le hachage de ses enregistrements en un nombre fini de groupes. Ensuite, le premier organisme calcule un total pour chaque groupe en fonction des enregistrements qui y sont sélectionnés, puis il transmet une version perturbée du résultat à son homologue après y avoir ajouté du bruit conformément aux paramètres de perte de confidentialité négociés. Les totaux perturbés des groupes sont fournis à l'organisme partenaire sous la forme d'un ensemble de données composé de microagrégats (un pour chaque groupe), qui s'apparente aux microdonnées à grande diffusion en ce qui concerne la confidentialité. À partir de ces renseignements, le deuxième organisme estime le total souhaité pour la population finie, puis publie une version perturbée de cette estimation après y avoir ajouté davantage de bruit, également conformément aux paramètres de perte de confidentialité convenus. Par rapport aux travaux antérieurs (p. ex. Dasylda, De Cubellis, De Fausti et Franssen, 2025; Dasylda, Santos, Franssen, De Cubellis, De Fausti, Pappagallo, Berrios et Fitzsimons, 2025), la procédure proposée pourrait être moins utile et procurer un degré de souplesse moindre en ce qui concerne les statistiques pouvant être tirées des microdonnées. Toutefois, une étude par simulation a montré qu'il est possible d'atteindre un niveau d'utilité élevé avec une perte modérée de confidentialité (Dasylda et Goussanou, 2026). La nouvelle procédure présente aussi de nombreux avantages importants par rapport aux solutions existantes, notamment son faible coût (aucun matériel spécialisé requis), sa capacité d'adaptation à des ensembles de données comportant des dizaines de millions d'enregistrements ainsi que son niveau renforcé de protection de la confidentialité, puisque tous les résultats statistiques et toutes les microdonnées échangées sont différenciellement confidentiels. De plus, cette procédure constitue un exemple concret de calcul multipartite, puisqu'elle ne nécessite aucune relation de confiance entre les parties et qu'aucun tiers de confiance n'intervient.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Abel Dasylda (abel.dasylda@statcan.gc.ca).

Bibliographie

Dasylda, A. (2025). A differentially private record linkage for the production of official statistics. Communication présentée au Comité consultatif sur les méthodes statistiques, octobre 2025, Statistique Canada.

Dasylda, A., De Cubellis, M., De Fausti, F. et Franssen, L. (2025). Linking trade data from different national statistical offices through a private set intersection. *Journal of Official Statistics*, numéro spécial, 1-29. Accessible à l'adresse : <https://journals.sagepub.com/doi/10.1177/0282423X251329407>.

Dasylda, A. et Goussanou, A. (2026). A joint model of record duplication and linkage error for the Canadian Census of Population. Rapport interne, Statistique Canada, Ottawa.

Dasylda, A., Santos, B., Franssen, L., De Cubellis, M., De Fausti, F., Pappagallo, A., Berrios, N. et Fitzsimons, J. (2025). Private linkage of international trade microdata in a cloud-based secure enclave. Rapport interne.

Dwork, C. et Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9, 211-407.

Escobar, F., Canard, S., Laguillaumie, F. et Phan, D.-H. (2024). [Computational differential privacy for encrypted databases supporting linear queries](https://petsymposium.org/popets/2024/popets-2024-0131.pdf). *Proceedings on Privacy Enhancing Technologies*, 2024(4), 583-604. Accessible à l'adresse : <https://petsymposium.org/popets/2024/popets-2024-0131.pdf>.

Inan, A., Kantarcioglu, M., Ghinita, G. et Bertino, E. (2010). Private record matching using differential privacy. *Proceedings of the 13th International Conference on Extending Database Technology*, 123-134. Accessible à l'adresse : <https://doi.org/10.1145/1739041.1739059>.

1.3 Estimation sur petits domaines

PROJET : Utilisation de forêts aléatoires pour l'estimation sur petits domaines

Lorsque la taille d'échantillon de domaines est petite, les estimateurs directs convergents par rapport au plan des paramètres de population sont susceptibles d'être instables. Pour améliorer la précision des estimateurs directs, le modèle de Fay-Herriot au niveau du domaine est souvent utilisé. Il comprend deux composantes : un modèle d'échantillonnage et un modèle de liaison. Ce dernier précise la relation entre les paramètres d'intérêt de la population et les variables auxiliaires disponibles au niveau du domaine. Dans sa forme initiale, le modèle de Fay-Herriot suppose un modèle de liaison linéaire avec variance d'erreur constante. Il exige également d'estimer la variance d'échantillonnage lissée des estimateurs directs, c'est-à-dire l'espérance par rapport au modèle de la variance d'échantillonnage des estimateurs directs.

Les hypothèses sous-jacentes aux modèles de Fay-Herriot et de lissage ne sont pas toujours satisfaites en pratique, et il peut être difficile et long de corriger adéquatement les modèles. Dans ce contexte, il peut être souhaitable d'avoir accès à des méthodes non paramétriques, surtout lorsque le nombre de domaines est élevé, parce qu'elles dépendent moins fortement de la validité des hypothèses de modèle et peuvent accélérer la production d'estimations sur petits domaines. Nous nous intéressons particulièrement aux forêts aléatoires pour trois raisons : i) elles peuvent être facilement appliquées dans le cas d'une combinaison de variables auxiliaires catégoriques et continues; ii) elles ne nécessitent pas de spécification d'interactions; et iii) elles produisent des prédictions qui demeurent toujours dans la fourchette des valeurs observées.

En premier lieu, nous avons élaboré et évalué, au moyen d'études empiriques, une version non paramétrique du meilleur prédicteur empirique lorsque des forêts aléatoires sont utilisées pour remplacer des modèles entièrement paramétriques. Par la suite, nous avons utilisé des prédictions « out-of-bag » comme variable auxiliaire dans un modèle linéaire de Fay-Herriot. Nos résultats montrent que les forêts aléatoires offrent de la robustesse par rapport à une spécification incorrecte du modèle, accroissent l'efficacité des estimations sur petits domaines et simplifient l'effort de modélisation (mais sans l'éliminer).

En 2024-2025, ce projet a été présenté à l'International Conference on Establishment Statistics à Glasgow en juin 2024 et à la conférence INCASS à Ottawa en juillet 2024. Nous avons également rédigé un article (Bosa, Beaumont, Bocci et Sombo, 2025) qui a été présenté au Comité consultatif des méthodes statistiques de Statistique Canada en juin 2025.

Progrès :

Nous avons précédemment développé une approche bootstrap pour estimer l'erreur quadratique moyenne lorsque les prédictions provenant des forêts aléatoires sont utilisées comme variable auxiliaire. Au cours de la dernière année, des fonctions R ont été développées pour mettre en œuvre la méthodologie proposée, incluant la méthode bootstrap. Nous avons également reprogrammé la méthodologie de base du modèle de Fay-Herriot, ce qui a permis de réduire le temps d'exécution en comparaison avec des fonctions R et SAS existantes. Une version beta des fonctions est disponible pour les employés de Statistique Canada.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca) ou

Keven Bosa (keven.bosa@statcan.gc.ca).

Bibliographie

Bosa, K., Beaumont, J.-F., Bocci, C. et Sombo, S. (2025). The use of a random forest algorithm in small area estimation. Article présenté au Comité consultatif sur les méthodes statistiques, juin 2025, Statistique Canada.

PROJET : Rétro-ingénierie d'un processus de réconciliation hypothétique pour obtenir la matrice de covariance des estimations directes dans le modèle de Fay-Herriot

Le modèle d'estimation sur petits domaines (EPD) de base de Fay-Herriot (FH) au niveau du domaine suppose l'indépendance des estimations directes du domaine. Dans le contexte de l'Enquête sur la population active (EPA), on a observé que pour l'EPD de l'emploi total pour une sous-division des provinces, les estimations directes sont très négativement corrélées. Cette situation est devenue évidente, car un processus de réconciliation est appliqué aux estimations sur petits domaines et les corrélations ont une incidence considérable sur les indicateurs de qualité de ces estimations. Ce projet porte sur l'extension du modèle de base de FH au cas de forte dépendance. On cherche à obtenir une meilleure estimation de la matrice de variance d'échantillonnage des estimations directes, qui est censée être connue tant dans le modèle de base que dans le modèle étendu. Pour y parvenir, on procède à la rétro-ingénierie d'un processus de réconciliation hypothétique, afin d'obtenir les termes de covariance compte tenu des termes de variance (lissée) et d'une matrice de variance des estimations directes

agrégées (les estimations provinciales totales dans le cas de l'EPA). De plus, on examine à la fois les EPD réconciliées et non réconciliées.

Progrès :

En 2025-2026, les travaux menés en 2024-2025 ont été présentés lors de la conférence de 2025 de la Société statistique du Canada et publiés dans le recueil des actes de la conférence (Verret et Walker, 2025). Les travaux portaient sur l'élaboration et la mise en œuvre de la matrice de variance complète dans le modèle de FH étendu, y compris l'extension de nombreuses caractéristiques du système généralisé de Statistique Canada, G-Est (sélection descendante de variables, diagnostics, détection des valeurs aberrantes). De plus, cette méthodologie a été utilisée à compter de juillet 2025 pour publier la version révisée des estimations sur petits domaines de l'EPA, découlant des données du Recensement de 2021. Les programmes sont passés de SAS à R plus tard au cours de l'année. Afin de valider davantage la méthode, une vaste étude par simulations de Monte Carlo (MC) a été lancée; les données du questionnaire détaillé du Recensement de 2016 ont servi de population cible, et l'EPA a été simulée selon un plan de sondage simplifié (échantillonnage aléatoire simple) et sans erreur de couverture ni erreur due à la non-réponse. Les matrices de variance théoriques des estimateurs de Horvitz-Thompson et par la régression généralisée ont été produites, puis comparées d'abord aux matrices de MC correspondantes. Elles ont ensuite été comparées à la matrice obtenue par la méthode de rétro-ingénierie et à une matrice de compromis décrite dans Isaki, Tsay et Fuller (2000). Les simulations ont mis en évidence des domaines d'amélioration potentiels concernant les termes de covariance; on pourrait notamment mieux tenir compte de la structure de dépendance de l'échantillonnage et du calage des poids par rapport au total de la population des domaines les plus vastes. D'autres simulations sont prévues pour répondre à ces besoins et évaluer l'incidence des différentes matrices de variance sur l'EPD à l'aide du modèle de FH étendu.

Pour obtenir plus de renseignements, veuillez communiquer avec :

François Verret (francois.verret@statcan.gc.ca) ou

Braedan Walker (braedan.walker@statcan.gc.ca).

Bibliographie

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2000). [Estimation des facteurs de correction au recensement. Techniques d'enquête](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000001/article/5176-fra.pdf), 26(1), 37-49. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000001/article/5176-fra.pdf>.

Verret, F. et Walker, B. (2025). Application of the Fay-Herriot model with correlated design errors: The Labour Force Survey experience. Actes de la réunion annuelle 2025 de la Société statistique du Canada (SSC), Saskatoon, Canada.

PROJET : Estimation sur petits domaines avec étalonnage : évaluation et étude par simulation

Dans le cadre de ce projet, nous évaluons et étudions la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE) ainsi que des estimateurs étalonnés du MPLSBE pour l'estimation sur petits domaines. Nous examinons les MPLSBE autoétalonnés de Wang, Fuller et Qu (2008) et de You, Rao et Hidirolou (2013) selon le modèle de Fay-Herriot, quand la modélisation est exacte et inexacte. Nous appliquons les estimateurs du MPLSBE aux données sur le taux de chômage de l'Enquête sur la population active et comparons les estimations basées sur un modèle aux valeurs issues du recensement. Nous menons également des études par simulation afin d'évaluer et de comparer le MPLSBE et les MPLSBE

étalonnés quand la modélisation est exacte et inexacte. Ce projet a pour objectif principal d'évaluer le MPLSBE et le MPLSBE étalonné quand on utilise une modélisation exacte et inexacte.

Progrès :

Nous avons mené une analyse des données et une étude par simulation. Nos résultats montrent que les MPLSBE étalonnés de Wang, Fuller et Qu (2008) et de You, Rao et Hidiroglou (2013) affichent un rendement très semblable. Toutefois, la procédure d'étalonnage n'améliore pas le MPLSBE lorsque la modélisation est inexacte ou qu'il y a erreur de spécification du modèle. Un rapport de recherche (You, 2026) a été rédigé et soumis à une revue à comité de lecture.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Yong You (yong.you@statcan.gc.ca).

Bibliographie

Wang, J., Fuller, W.A. et Qu, Y. (2008). [Estimation pour petits domaines sous une contrainte](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10619-fra.pdf). *Techniques d'enquête*, 34(1), 33-40. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10619-fra.pdf>.

You, Y. (2026). Comparison and evaluation of small area EBLUP estimators under benchmarking with application and simulation study. Soumis à une revue à comité de lecture en vue d'une publication éventuelle.

You, Y., Rao, J.N.K. et Hidiroglou, M.A. (2013). [De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11830-fra.pdf). *Techniques d'enquête*, 39(1), 243-255. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11830-fra.pdf>.

PROJET : Étude empirique, évaluation des effets d'emprunt et construction de distributions *a priori* informatives en estimation sur petits domaines

La modélisation hiérarchique bayésienne (HB) est très populaire en estimation sur petits domaines, et la spécification des distributions *a priori* est très importante dans cette approche. Dans le cadre de ce projet, nous étudions le rendement d'estimateurs HB sur petits domaines, reposant sur les modèles bayésiens de You et Chapman (2006) et de You (2021), et considérons des distributions *a priori* non informatives et informatives pour les paramètres de régression et les composantes de variance. Nous appliquons ces estimateurs aux données de l'Enquête sur la population active (EPA) du Canada. Nous étudions l'incidence de distributions *a priori* informatives et non informatives appropriées et inappropriées pour l'estimation HB sur petits domaines, à l'aide d'une application aux données de l'EPA et d'une étude par simulation.

Progrès :

Nous avons publié un article (You et Bosa, 2025) reposant sur ce projet. Dans la lignée des travaux de You et Bosa (2025), nous proposons une approche itérative de spécification des distributions *a priori* pour l'utilisation des données du mois précédent. Nous évaluons et comparons ensuite les estimations HB à l'aide de distributions *a priori* non informatives et informatives dans le cadre d'une étude empirique. Nos résultats montrent que les distributions *a priori* informatives peuvent donner d'excellents résultats et que

l'approche itérative proposée pour la spécification des distributions *a priori* peut conduire à de meilleurs résultats et à des résultats stables, caractérisés par un biais et un coefficient de variation plus faibles. L'ébauche d'un article de recherche (You et Bosa, 2026) a été rédigée.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Yong You (yong.you@statcan.gc.ca).

Bibliographie

You, Y. (2021). [Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00007-fra.pdf). *Techniques d'enquête*, 47(2), 389-399. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00007-fra.pdf>.

You, Y. et Bosa, K. (2025). [Rendement des estimateurs hiérarchiques bayésiens sur petits domaines reposant sur des distributions *a priori* non informatives et informatives et application à l'Enquête sur la population active du Canada](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025002/article/00010-fra.pdf). *Techniques d'enquête*, 51(2), 461-488. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025002/article/00010-fra.pdf>.

You, Y. et Bosa, K. (2026). An empirical study on the use of informative priors for the hierarchical Bayes small area estimation with application. Article de recherche en cours de rédaction, destiné à être soumis pour une publication éventuelle.

You, Y. et Chapman, B. (2006). [Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf). *Techniques d'enquête*, 32(1), 107-114. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf>.

PROJET : Estimation sur petits domaines pour les indices des prix à la production

Ce projet avait pour objectif d'évaluer des méthodes d'estimation sur petits domaines (EPD) afin d'obtenir des estimations plus détaillées des indices des prix à la production pour des industries et des groupes de produits, ainsi que d'améliorer la qualité des estimations actuelles dans un contexte de taux de réponse faibles et de contraintes budgétaires. Nous avons examiné le modèle transversal de Fay-Herriot (FH), le modèle de Rao-Yu (RY) et un modèle multiniveau comportant un terme de série chronologique.

Progrès :

Au cours de l'année 2025-2026, nous avons adapté la théorie établie de l'EPD afin d'estimer un indice de prix de Lowe ou de Young. Nous avons démontré que la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE), dans le cadre du modèle de FH, produit un indice de prix transitif selon certaines conditions, mais que le MPLSBE du modèle de RY et les estimations hiérarchiques bayésiennes pour le modèle multiniveau n'ont pas les propriétés souhaitées pour les indices de prix. Nous avons modélisé l'Indice des prix des services du commerce de détail, qui est mensuel, à l'aide d'un grand nombre de séries chronologiques de covariables, notamment des indicateurs généraux du marché, des caractéristiques de l'industrie et d'autres indices de prix. L'analyse exploratoire et la sélection du modèle ont malheureusement révélé l'absence de bons prédicteurs. Des EPD mensuelles ont été établies, mais elles présentaient une erreur quadratique moyenne élevée et ne semblaient pas d'une qualité suffisante pour être mises en œuvre. Cela souligne l'importance d'une collecte rigoureuse des données sur les prix, car la modélisation ne peut pas facilement remplacer ces données.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Justin Francis (justin.francis@statcan.gc.ca).

2 Méthodes et applications de la science des données

PROJET : Utilisation de l'affinage efficace des paramètres sur de petits modèles de langage pour traiter la classification statistique de codes et la vérification des faits

Ce projet vise à explorer le potentiel des modèles de langage transformeurs plus petits et à poids ouverts pour des tâches spécialisées de classification statistique de codes. En utilisant des modèles libres, nous avons cherché à déterminer si l'application de l'affinage efficace des paramètres (parameter-efficient fine-tuning) à des ensembles de données triés pouvait permettre à un modèle compact d'apprendre des schémas complexes de raisonnement et de prise de décision. Nous voulions démontrer que des modèles plus petits et efficaces sur le plan des ressources peuvent offrir un rendement comparable à celui de modèles exclusifs plus vastes, tout en facilitant un déploiement rentable.

Progrès :

Il y a eu deux cas pratiques distincts, soit l'affinage de modèles de plongement et de reclassement préentraînés comme première étape de la classification statistique des codes du Système de classification des produits de l'Amérique du Nord, et l'affinage de grands modèles de langage libres plus petits en s'appuyant sur le raisonnement d'un modèle enseignant GPT-5, afin d'effectuer le changement de base de la Classification nationale des professions pour passer de la version de 2016 à la version de 2021 pour l'Enquête sur la population active.

Le premier cas pratique n'a pas été concluant, les modèles affinés ayant affiché un rendement inférieur par rapport aux modèles de base. Parmi les difficultés, il y avait les données bruitées ainsi que le manque de ressources adéquates de processeur graphique pour effectuer l'affinage sur des données suffisantes. De plus, l'affinage de modèles de plongement est moins bien documenté et dépend fortement de la mise en forme et de la qualité des données, ce qui rend la tâche en soi plus difficile.

En ce qui concerne le deuxième cas pratique, nous avons réussi à mettre en œuvre l'affinage efficace des paramètres et l'adaptation par modèle auxiliaire quantifiée (quantized low-rank adaptation) avec un processeur graphique A100. Nous avons créé des données d'entraînement tirant parti des réponses libres, des définitions de codes et des réponses du modèle enseignant, ce qui a permis au modèle étudiant d'imiter le raisonnement et l'exactitude de classification du modèle enseignant. L'affinage a été effectué sur 10 000 et 30 000 exemples; dans ce dernier cas, le modèle étudiant libre a surpassé le modèle enseignant exclusif en ce qui concerne la mesure F1 pondérée moyenne (+1,4 %; 79,93 % par rapport à 78,56 %). Parmi les explications possibles, qui méritent d'être approfondies, il y a le fait qu'en entraînant le modèle étudiant exclusivement à partir des extraits de raisonnement positifs du modèle enseignant, le modèle étudiant a pu assimiler les schémas d'évaluation plus efficaces, ce qui a finalement conduit à une mesure légèrement supérieure. Mis à part cette modeste amélioration de la mesure, le fait de disposer d'un modèle libre ayant un bon rendement offre un meilleur contrôle sur le déploiement de fonctionnalités, comme les probabilités logarithmiques brutes, qui n'étaient pas offertes avec le modèle GPT-5 mini, et l'accès à celles-ci. Cet accès permet d'avoir des fonctionnalités supplémentaires, comme la quantification de l'incertitude en fonction de ces probabilités logarithmiques.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Alexandre Istrate (alexandre.istrate@statcan.gc.ca).

PROJET : Amélioration de la classification textuelle à l'aide de la prédiction conforme

Les systèmes de classification textuelle à grande échelle nécessitent souvent de classer les réponses textuelles dans l'une des nombreuses catégories possibles, tout en fournissant des mesures fiables de l'incertitude des prédictions. Ce projet vise à mettre au point des méthodes de classification tenant compte de l'incertitude, qui reposent sur la prédiction conforme, et ce, en passant d'expériences initiales à petite échelle (25 classes) à des systèmes comportant quelques milliers de classes. Cette approche permet une quantification fiable de l'incertitude dans le cadre de l'attribution de codes, fondée sur un grand modèle de langage, à des réponses d'enquête sous forme de texte. En s'appuyant sur des ensembles de données de référence à trois et à sept chiffres du Système de classification des produits de l'Amérique du Nord (SCPAN), ces travaux visent à rendre compte explicitement de l'incertitude des prédictions afin de favoriser une plus grande automatisation, de réduire les vérifications manuelles et d'améliorer la confiance à l'égard des systèmes de classification textuelle à grande échelle.

Progrès :

Dans le cadre de ce projet de recherche, un cadre de prédiction conforme fondé sur les logits a été mis au point et évalué pour des problèmes de classification en présence d'un grand nombre de classes. Un pipeline expérimental de bout en bout a été mis en place pour extraire les résultats des modèles, générer des ensembles de prédiction conforme et mener des expériences à grande échelle. Les premiers essais ont été menés à l'aide d'un ensemble de données de mathématiques simples (un ensemble de données synthétiques composé de 125 questions portant sur des problèmes arithmétiques élémentaires et regroupées en 25 classes définies par les réponses) afin de valider la méthodologie dans un environnement contrôlé et d'étudier l'incidence des contraintes imposées aux résultats des modèles sur la fiabilité des prédictions. L'approche a ensuite été évaluée à l'aide d'ensembles de données issus du SCPAN qui comprenaient à la fois des tâches de classification par grand groupe de produits (à trois chiffres) et des tâches de classification détaillée par code de produits à sept chiffres.

Les résultats ont montré que les modèles sous-jacents donnaient généralement de bons résultats lorsqu'il s'agissait de classer des classes candidates dans des sous-ensembles plus petits, mais qu'ils éprouvaient des difficultés lorsque les prédictions étaient agrégées dans des espaces de classes très vastes. Lorsqu'il a été combiné à la prédiction conforme, le cadre a donné des résultats prometteurs pour des tâches de classification plus générales; toutefois, dans des contextes très détaillés concernant des milliers de classes, il a produit des ensembles de prédiction trop vastes pour être utiles sur le plan opérationnel. Ces résultats indiquent que l'efficacité de la prédiction conforme dans des contextes où il y a un grand nombre de classes dépend fortement du rendement et du calage du modèle de langage sous-jacent. D'autres travaux de recherche sont prévus afin d'évaluer des modèles ayant un meilleur rendement et de déterminer si les modèles logit peuvent constituer une base fiable pour la prédiction conforme dans le cadre d'applications de classification à grande échelle.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mihir Gajjar (mihir.gajjar@statcan.gc.ca).

PROJET : Étalonnage de l'attribution de codes fondée sur des grands modèles de langage pour les données d'enquête à grande échelle

Ce projet vise à évaluer l'utilisation de grands modèles de langage (GML) pour automatiser l'attribution de codes à des réponses d'enquête selon des systèmes de classification normalisés (comme ceux se

rapportant aux codes de profession et d'industrie), par rapport à des modèles d'apprentissage automatique (AA) existants, comme fastText et DistilBERT. On cherche principalement à soutenir l'objectif de Statistique Canada visant à moderniser et à élargir les processus d'attribution de codes pour de grands volumes de données non structurées.

Un cadre d'étalonnage normalisé a été créé et comprend notamment un ensemble de données de référence, un pipeline d'évaluation unifié et des mesures de rendement uniformes (p. ex. un taux d'automatisation, des mesures F1 et des taux d'erreur). Dans le cadre du projet, on a également mis au point un pipeline d'attribution de codes basé sur des GML; il effectue le prétraitement des données, la recherche et la vérification des codes potentiels, ainsi que l'attribution de scores de confiance calés, afin de permettre des comparaisons équitables avec les modèles traditionnels entraînés.

Progrès :

Le cadre a été appliqué à des données tirées de l'Enquête sur la population active et de l'Enquête sur les postes vacants et les salaires, et mis à l'essai à l'aide de ces données. Les résultats montrent que les modèles d'AA classiques (fastText et DistilBERT) surpassent les GML tant au chapitre de l'exactitude que des taux d'automatisation, en particulier lorsque les seuils de confiance sont plus élevés. Ces résultats mettent en évidence les limites actuelles des GML préalablement entraînés pour effectuer des tâches d'attribution de codes qui reposent sur des sources d'information externes et une expertise en la matière, et pour déterminer les défis liés à la génération de scores de confiance calés pour les prédictions fondées sur des GML.

L'étape 2 de la recherche visera à renforcer les approches fondées sur l'AA et à peaufiner les stratégies d'étalonnage, tout en évaluant les répercussions sur les résultats des enquêtes en collaboration avec des spécialistes du domaine.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Soufiane Fadel (soufiane.fadel@statcan.gc.ca).

PROJET : Estimation fondée sur la prédiction pour les statistiques officielles

Les organismes nationaux de statistique doivent de plus en plus composer avec une baisse des taux de réponse, une réduction de la taille des échantillons et des contraintes budgétaires, tout en ayant désormais accès à de riches sources de données auxiliaires, comme des dossiers administratifs, des images et des textes. Parallèlement, les progrès réalisés dans les domaines de l'apprentissage automatique (AA), de l'intelligence artificielle (IA) et des grands modèles de langage (GML) offrent de puissantes capacités prédictives susceptibles d'améliorer l'estimation statistique. Cependant, les estimateurs d'enquête classiques assistés par un modèle nécessitent généralement que les modèles soient ajustés à l'aide des données tirées de l'échantillon d'enquête et ne peuvent donc pas tirer pleinement parti des grands modèles à boîte noire préentraînés avec des ensembles de données externes.

Ces travaux de recherche permettent de se pencher à nouveau sur l'utilisation de modèles préentraînés pour l'estimation fondée sur le plan et présentent un nouveau cadre visant à intégrer l'IA au domaine des statistiques officielles tout en préservant les propriétés fondées sur le plan de sondage. L'article propose l'estimateur assisté par des prédictions (prediction-powered estimator), l'estimateur par la différence assisté par des prédictions (prediction-powered difference estimator) et une adaptation de l'estimateur par la régression généralisée pour laquelle les prédictions des modèles servent d'information auxiliaire.

Les méthodes proposées permettent d'utiliser des modèles d'AA et des GML préentraînés sans qu'il soit nécessaire de les réentraîner à l'aide de données d'enquête, donc des systèmes d'IA modernes et de nouvelles sources de données comptent parmi les outils à la disposition des méthodologistes chargés des enquêtes.

Progrès :

Ces travaux de recherche ont abouti à une publication évaluée par des pairs (Denis et Haddou, 2026) montrant la façon dont des modèles d'AA et d'IA préentraînés peuvent être intégrés à l'estimation fondée sur le plan tout en garantissant une estimation sans biais et une estimation valide de la variance. L'étude a évalué les estimateurs proposés à l'aide de données d'enquête réelles provenant de Statistique Canada, auxquelles s'ajoutaient des expériences par simulation et des applications novatrices fondées sur des GML comprenant des données non structurées comme du texte et des images. Pour plusieurs mesures de rendement, dont le biais, l'erreur quadratique moyenne, la couverture de l'intervalle de confiance et la largeur de l'intervalle, les approches proposées ont systématiquement surpassé les estimateurs classiques servant de points de référence.

L'étude s'inscrit dans le domaine émergent de l'IA inférentielle en proposant des méthodes fondées sur des principes afin d'intégrer des modèles d'IA pionniers au contexte des statistiques officielles. Les résultats indiquent que les modèles préentraînés peuvent améliorer l'efficacité de l'estimation, réduire potentiellement la taille des échantillons d'enquête requis et alléger le fardeau imposé aux répondants tout en préservant la rigueur statistique. Ces travaux jettent également les bases d'applications futures de l'inférence optimisée par l'IA au sein des organismes nationaux de statistique.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mohammed Haddou (mohammed.haddou@statcan.gc.ca).

Bibliographie

Denis, N. et Haddou, M. (2026). Prediction-powered estimation: Unbiased model-assisted estimation. *Journal of Official Statistics*. Accessible à l'adresse : <https://doi.org/10.1177/0282423X261451320>.

PROJET : Guide de la Commission économique des Nations Unies pour l'Europe sur la quantification de l'incertitude pour l'inférence statistique améliorée par l'intelligence artificielle et l'apprentissage automatique

Comme les organismes nationaux de statistique (ONS) recourent de plus en plus à l'apprentissage automatique (AA) et à l'intelligence artificielle (IA) pour améliorer l'efficacité et l'étendue de la production statistique, garantir la validité et la fiabilité statistiques des résultats reposant sur l'IA est devenu un défi crucial. Bien que les méthodes d'AA puissent améliorer considérablement l'exactitude des prévisions et l'efficacité opérationnelle, elles ne fournissent souvent pas de garanties en ce qui concerne la quantification de l'incertitude et l'inférence, lesquelles sont habituellement exigées pour les statistiques officielles. Cela soulève une question méthodologique importante, à savoir comment les ONS peuvent-ils tirer parti de l'IA moderne tout en préservant la rigueur et la confiance du public qui sont à la base des statistiques officielles?

Pour relever ce défi, Statistique Canada a dirigé, en collaboration avec plusieurs ONS, l'élaboration du Handbook on Uncertainty Quantification for AI/ML-Enhanced Statistical Inference (guide sur la quantification de l'incertitude pour l'inférence statistique améliorée par l'IA et l'AA) de la Commission économique des Nations Unies pour l'Europe (CEE-ONU), dans le cadre de l'initiative ModernStats du Groupe de haut niveau sur la modernisation des statistiques officielles. Ce guide présente des cadres reposant sur des principes afin d'intégrer les méthodes modernes d'IA au domaine des statistiques officielles tout en préservant la validité inférentielle. En plus de passer en revue les approches classiques de quantification de l'incertitude, on y fournit des conseils détaillés sur deux cadres complémentaires : la prédiction conforme, qui permet d'avoir des garanties liées à l'incertitude des prédictions pour des échantillons de taille finie, et ce, sans supposer de distribution particulière, et l'inférence par prédiction, qui permet d'utiliser des modèles préentraînés tout en préservant la validité fondée sur le plan.

Progrès :

Le guide a été rédigé et publié par la CEE-ONU (CEE-ONU, 2026). Statistique Canada a dirigé l'équipe de travail internationale chargée d'élaborer le guide et de coordonner les contributions de plusieurs ONS. Ce guide comprend des conseils méthodologiques, des considérations relatives à la mise en œuvre, des cadres décisionnels ainsi que des études de cas opérationnelles provenant de plusieurs organismes nationaux de statistique afin d'illustrer les applications pratiques des méthodes de quantification de l'incertitude dans des environnements de production. Les études de cas portent sur l'attribution de codes automatisée, la classification économique et l'estimation de la proportion à l'aide de techniques modernes d'IA.

Ce guide s'inscrit dans le domaine émergent de l'IA inférentielle en démontrant que les méthodes modernes d'IA peuvent être intégrées aux cadres méthodologiques d'enquête existants sans compromettre la rigueur ou la fiabilité statistiques. En élargissant l'estimation classique assistée par un modèle afin d'intégrer des modèles d'AA et des modèles fondateurs complexes, ce guide offre aux spécialistes des outils pratiques pour le déploiement de systèmes d'IA, tout en préservant la validité des intervalles de confiance, l'inférence fondée sur le plan et la confiance du public à l'égard des statistiques officielles.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mohammed Haddou (mohammed.haddou@statcan.gc.ca).

Bibliographie

Commission économique pour l'Europe des Nations Unies (CEE-ONU) (2026). UNECE handbook on Uncertainty quantification in ML-enhanced statistical inference - A practitioner handbook with applications in official statistics. Commission économique pour l'Europe des Nations Unies (CEE-ONU), HLG-MOS ModernStats Initiative. Accessible à l'adresse :

<https://unece.org/statistics/documents/2026/06/reports/uncertainty-quantification-ml-enhanced-statistical-inference>.

PROJET : Validité et fiabilité des grands modèles de langage utilisés comme évaluateurs

L'adoption rapide des grands modèles de langage (GML) a conduit à leur utilisation croissante non seulement comme systèmes génératifs, mais aussi comme évaluateurs des résultats de l'apprentissage automatique, communément appelés en anglais « LLMs-as-judges » (GML utilisés comme évaluateurs).

Ces systèmes sont désormais grandement utilisés dans les tâches d'évaluation, de définition de points de référence et de comparaison de modèles liées à la génération automatique de texte (GAT), en raison de leur capacité à s'ajuster à l'échelle et de leur faible coût. Cependant, malgré leur popularité croissante, des questions fondamentales subsistent quant à leur validité et à leur fiabilité statistiques en tant qu'instruments de mesure. En particulier, on ne sait pas clairement si les GML utilisés comme évaluateurs fournissent des évaluations uniformes, impartiales et conformes au jugement humain, d'une manière qui se prête à l'inférence scientifique et à la prise de décision.

Dans ces travaux de recherche, on applique la théorie de la mesure issue des sciences sociales afin d'examiner de manière systématique les hypothèses qui sous-tendent l'utilisation des GML comme évaluateurs. On examine d'un œil critique quatre hypothèses fondamentales : i) les résultats des GML utilisés comme évaluateurs peuvent servir d'approximations fiables du jugement humain; ii) les GML utilisés comme évaluateurs constituent des évaluateurs valides des résultats des modèles; iii) ils s'ajustent efficacement à l'échelle sans nuire à la qualité des mesures; iv) ils offrent des avantages considérables sur le plan des coûts par rapport aux méthodes d'évaluation traditionnelles. On analyse ensuite trois applications représentatives, soit la génération de résumés, l'annotation de données et la mise en conformité en matière de sécurité, afin d'illustrer la façon dont ces hypothèses peuvent s'avérer erronées en pratique et la façon dont cela peut influencer sur les résultats des évaluations en aval.

Progrès :

Ces travaux de recherche ont été publiés dans les actes de la conférence Neural Information Processing Systems de 2026 (Chehbouni, Haddou, Cheung et Farnadi, 2026) après avoir été acceptés en vue de leur présentation à la conférence. On propose une analyse critique structurée du paradigme émergent d'évaluation fondée sur les GML et on le situe dans un cadre officiel de la théorie de la mesure. Les résultats soulignent que, bien que les GML utilisés comme évaluateurs soient attrayants en raison de leur capacité de s'ajuster à l'échelle et de leur efficacité, leur adoption généralisée pourrait être prématurée en l'absence d'une validation rigoureuse de leur fiabilité et de leur validité conceptuelle.

Cet article s'inscrit dans le cadre élargi de l'évaluation responsable de l'IA en déterminant les principales limites des pratiques actuelles d'évaluation fondée sur les GML et en soulignant l'importance de la validité des mesures dans les pipelines d'évaluation de l'apprentissage automatique. Il soutient l'élaboration de normes d'évaluation plus rigoureuses pour les systèmes de GAT et souligne la nécessité d'une rigueur méthodologique lors du déploiement des GML dans des rôles allant au-delà de la génération, en particulier dans les contextes où les résultats des évaluations orientent les priorités en matière de recherche, les points de référence et les décisions pertinentes sur le plan des politiques.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mohammed Haddou (mohammed.haddou@statcan.gc.ca).

Bibliographie

Chehbouni, K., Haddou, M., Cheung, J.C.K. et Farnadi, G. (2026). Neither valid nor reliable? Investigating the use of LLMs as judges. Proceedings of NeurIPS 2025. Accessible à l'adresse : https://papers.nips.cc/paper_files/paper/2025/hash/829e8f32d76a6248815bf5b01633811d-Abstract-Position_Paper_Track.html.

PROJET : Cadre d'intelligence artificielle responsable pour les statistiques officielles

Les progrès rapides de l'intelligence artificielle (IA) ont conduit à une hausse de son adoption au sein des organismes nationaux de statistique (ONS), ce qui a ouvert la voie à de nouvelles capacités en matière de traitement, d'analyse et de diffusion des données. Toutefois, parallèlement à ces possibilités, l'IA comporte des risques importants liés aux biais, à la protection des renseignements personnels, à la transparence, à la reddition de comptes et à la confiance du public. Dans le cas des ONS, qui sont soumis à des mandats stricts visant à produire des statistiques officielles fiables dans l'intérêt public, ces risques exigent une approche en matière de gouvernance de l'IA qui est structurée et fondée sur des principes. Parmi les principaux défis, il y a donc le fait de veiller à ce que les systèmes d'IA soient déployés de manière éthique, transparente et conforme aux principes fondamentaux du domaine des statistiques officielles.

Ces travaux de recherche ont contribué à l'élaboration d'un cadre de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) sur l'IA responsable dans le domaine des statistiques officielles, lequel découle de l'initiative du Groupe de haut niveau sur la modernisation des statistiques officielles (GHN-MSO). Ce cadre résume les lignes directrices internationales d'organismes de statistique, d'organisations et d'institutions, comme Statistique Canada, l'Organisation de coopération et de développement économiques et la Commission européenne, et établit un ensemble de principes directeurs visant à assurer la conception, le déploiement et la surveillance responsables des systèmes d'IA au sein des ONS. Parmi ces principes, il y a le respect des normes éthiques et l'intérêt public, la reddition de comptes, la transparence et l'explicabilité, l'équité et l'inclusivité, la protection des renseignements personnels et la sécurité, ainsi que la validité et la robustesse. Ce cadre fournit à la fois les fondements conceptuels et les directives opérationnelles nécessaires pour intégrer l'IA aux systèmes de production statistique tout en préservant la confiance institutionnelle.

Progrès :

Ces travaux ont été publiés sous forme de rapport de la CEE-ONU dans le cadre de l'initiative ModernStats du GHN-MSO (CEE-ONU, 2025). Le cadre a été élaboré en collaboration avec des partenaires internationaux et s'appuie sur les approches de gouvernance existantes, lesquelles ont été adaptées précisément aux besoins et aux contraintes du domaine des statistiques officielles. Il fournit aux ONS des conseils pratiques sur la manière de mettre en œuvre les principes d'IA responsable tout au long du cycle de vie de la production statistique, de l'élaboration et de la validation des modèles à leur déploiement et à leur surveillance.

Le cadre renforce les fondements de la gouvernance en vue de l'adoption de l'IA dans le domaine des statistiques officielles, en permettant de trouver un équilibre entre l'innovation, la reddition de comptes et la confiance du public. Il souligne que les systèmes d'IA doivent non seulement être techniquement robustes, mais aussi aller de pair, sur les plans social et institutionnel, avec le mandat des ONS. Ces travaux favorisent l'établissement de normes internationales communes en matière d'IA responsable dans le domaine des statistiques officielles et soutiennent l'intégration sécuritaire et efficace des technologies d'IA dans les environnements de production.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mohammed Haddou (mohammed.haddou@statcan.gc.ca).

Bibliographie

Commission économique pour l'Europe des Nations Unies (CEE-ONU) (2025). [Framework on Responsible AI for Official Statistics](https://unece.org/sites/default/files/2025-10/Responsible%20AI%20for%20Official%20Statistics%20Framework_ADSaMM%20Group_HLG-MOS_0.pdf). Accessible à l'adresse : https://unece.org/sites/default/files/2025-10/Responsible%20AI%20for%20Official%20Statistics%20Framework_ADSaMM%20Group_HLG-MOS_0.pdf.

PROJET : Équité et biais en apprentissage automatique pour les statistiques officielles

À mesure que les systèmes d'intelligence artificielle (IA) et d'apprentissage automatique (AA) sont de plus en plus intégrés aux processus de production statistique et aux systèmes d'aide à la décision, les préoccupations relatives à l'équité et aux biais sont devenues des enjeux centraux pour leur déploiement responsable. Les biais dans les systèmes d'IA peuvent se manifester sous diverses formes et à différentes étapes du cycle de vie de l'apprentissage automatique, et mener à des résultats systématiquement inégaux entre les différents groupes de population. Dans le contexte des statistiques officielles, où les résultats orientent les politiques publiques et l'affectation des ressources, il est essentiel de comprendre et d'atténuer les biais afin de garantir l'équité, la transparence et la confiance du public.

Ces travaux ont servi à rédiger un document d'accompagnement de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) relatif au cadre sur l'IA responsable dans le domaine des statistiques officielles, et mettaient particulièrement l'accent sur le concept d'équité et les différentes formes de biais dans les systèmes d'AA. Dans cet article, on explique comment la notion de « biais » est définie dans différentes disciplines, en distinguant les notions statistiques de biais des interprétations juridiques et éthiques liées à la discrimination et au préjudice. En outre, on élabore une taxonomie structurée des biais présents dans les systèmes d'AA, notamment les biais historiques, les biais de représentation, les biais liés au traitement algorithmique, les biais liés au déploiement et les biais liés aux boucles de rétroaction. On souligne que ces formes de biais sont souvent interdépendantes et peuvent conjointement influencer les résultats dans des systèmes sociotechniques complexes.

Progrès :

Ces travaux ont été menés à bien et publiés sous forme de document d'accompagnement de la CEE-ONU dans le cadre de l'initiative ModernStats du Groupe de haut niveau sur la modernisation des statistiques officielles (Haddou, Seimandi et Del Monaco, 2025). L'article s'ajoute au cadre élargi de l'IA responsable en proposant une analyse conceptuelle et pratique ciblée de l'équité et des biais dans les applications d'AA destinées au domaine des statistiques officielles. Il résume les données issues de la statistique, de l'informatique, du droit et de l'éthique afin de favoriser une compréhension commune des biais dans toutes les disciplines et d'orienter l'élaboration de stratégies d'atténuation dans les milieux opérationnels.

Cette contribution renforce les fondements conceptuels de l'IA responsable dans le domaine des statistiques officielles en fournissant une classification claire et structurée des types de biais et leurs répercussions sur les systèmes statistiques. L'article aide les organismes nationaux de statistique à repérer les sources potentielles d'iniquité dans les pipelines d'IA et d'AA et souligne l'importance de tenir compte du contexte sociotechnique lors du déploiement de modèles d'AA dans des environnements à enjeux élevés. Les travaux contribuent à l'élaboration de lignes directrices internationales harmonisées sur l'équité en matière d'IA pour le domaine des statistiques officielles.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Mohammed Haddou (mohammed.haddou@statcan.gc.ca).

Bibliographie

Haddou, M., Seimandi, T. et Del Monaco, A. (2025). [Fairness in Machine Learning: Companion paper to the Framework on Responsible AI for Official Statistics](https://unece.org/sites/default/files/2025-10/Companion%20Paper%20on%20Fairness%20in%20Machine%20Learning_Responsible%20AI%20Framework_ADSaMM%20group.pdf). Accessible à l'adresse : https://unece.org/sites/default/files/2025-10/Companion%20Paper%20on%20Fairness%20in%20Machine%20Learning_Responsible%20AI%20Framework_ADSaMM%20group.pdf.

3 Problèmes d'estimation dans les enquêtes

PROJET : Degrés de liberté liés à l'estimation du questionnaire détaillé du recensement

Statistique Canada publie désormais des intervalles de confiance afin d'exprimer la qualité des estimations du recensement. La longueur d'un intervalle de confiance peut témoigner de la précision d'une estimation, à condition que la couverture annoncée soit respectée. Un paramètre jouant un rôle important dans le calcul des intervalles de confiance est le nombre de degrés de liberté. En pratique, cette valeur est généralement déterminée à l'aide d'une règle empirique. Dans le cas de petits domaines, cette règle tend toutefois à surestimer le nombre réel de degrés de liberté, ce qui entraîne une sous-couverture des intervalles de confiance.

Dans ce projet, l'approximation de Satterthwaite est utilisée afin d'obtenir une estimation plus précise des degrés de liberté dans le contexte d'estimations calculées à partir des données du questionnaire détaillé du recensement de la population. La méthode d'estimation de la variance repose sur une adaptation de la méthode des demi-échantillons équilibrés, telle que décrite par Devin et Verret (2016). Une étude de simulation évalue le gain au niveau de la couverture des intervalles de confiance pour l'estimation de totaux de variables continues et dichotomiques. Les résultats suggèrent qu'en utilisant un nombre de degrés de liberté plus précis, la couverture est améliorée et atteint plus souvent le seuil nominal dans le cas, souvent problématique, des petits domaines.

L'objectif du projet pour l'année 2025-2026 était de procéder aux corrections demandées à l'article de Toupin et Martin (2026) en vue de sa publication dans la revue *Techniques d'enquête*.

Progrès :

L'article de Toupin et Martin (2026) a été accepté pour publication dans la revue *Techniques d'enquête*. Celui-ci paraîtra dans le numéro de juin 2026.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Marie-Hélène Toupin (marie-helene.toupin@statcan.gc.ca).

Bibliographie

Devin, N. et Verret, F. (2016). The development of a variance estimation methodology for large-scale dissemination of quality indicators for the 2016 Canadian census long form sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria Virginie.

Toupin, M.-H. et Martin, V. (2026). [Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application au recensement canadien](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2026001/article/00005-fra.pdf). *Techniques d'enquête*, 52(1), 209-233. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2026001/article/00005-fra.pdf>.

PROJET : Estimation de la variance par groupes aléatoires dépendants à l'aide de la méthode de Fay

Les méthodes fondées sur les répliques sont grandement utilisées en échantillonnage pour estimer la variance des estimations des paramètres de la population. Ce processus consiste à prélever de manière répétitive des sous-échantillons à partir de l'échantillon initial et à mesurer la variation des estimations obtenues. Toutefois, les méthodes des répliques peuvent s'avérer instables lors de l'estimation des variances pour des ratios portant sur des caractéristiques rares, car ces caractéristiques peuvent disparaître de certains sous-échantillons. Pour résoudre ce problème, Robert Fay a proposé une approche de lissage des poids de rééchantillonnage dans la méthode des répliques répétées équilibrées, garantissant ainsi que toutes les unités échantillonnées contribuent à chaque estimation de sous-échantillon. Dans le cadre de ce projet, nous avons étendu l'approche de lissage de Fay (Dippo, Fay et Morganstein, 1984), appelée « méthode de Fay », à la méthode de réplication par groupes aléatoires dépendants, qui sert à estimer les variances pour le fichier de microdonnées à grande diffusion (FMGD) des particuliers du recensement canadien.

Progrès :

Nous avons modifié le système de pondération de la méthode de réplication par groupes aléatoires dépendants en nous inspirant des idées proposées par Robert Fay pour la méthode des répliques répétées équilibrées. Afin de compenser la sous-estimation de la variance qu'apportent les poids de rééchantillonnage lissés, nous avons obtenu un facteur de correction et l'avons intégré à la formule de variance des répliques. Nous avons ensuite mené une série d'études par simulation afin d'évaluer le biais et la variabilité de l'estimateur de variance des groupes aléatoires dépendants ainsi obtenu pour un éventail de statistiques, en utilisant différents niveaux de lissage des poids de rééchantillonnage. Les simulations ont montré qu'un lissage modéré suffisait à stabiliser l'estimation de la variance des ratios rares, tout en maintenant un bon rendement pour les autres types d'estimateurs. Nous avons également intégré un facteur de correction pour population finie aux poids de rééchantillonnage lissés, ce qui a permis d'obtenir des estimations de variance non biaisées en présence de fractions d'échantillonnage non négligeables. Les renseignements techniques et les résultats des simulations ont été réunis dans un article qui a été présenté à la revue *Techniques d'enquête* (Imbrogno et Verret, 2025). Les améliorations méthodologiques ont été intégrées au FMGD des particuliers du recensement afin d'améliorer la stabilité des estimations de variance à l'avenir.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Alex Imbrogno (alexander.imbrogno@statcan.gc.ca) ou

François Verret (francois.verret@statcan.gc.ca).

Bibliographie

Dippo, C.S., Fay, R.E. et Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginie.

Imbrogno, A. et Verret, F. (2025). Dependent random group variance estimation using Fay's method. Article soumis pour publication dans la revue *Techniques d'enquête*.

PROJET : Estimation bootstrap de la variance pour des estimateurs par calage

Les estimations d'enquête sont souvent calées pour qu'elles correspondent aux totaux connus de la population. L'estimateur par la régression généralisée (estimateur RG) est un estimateur de calage populaire qui suppose une relation linéaire entre la variable d'enquête et les variables auxiliaires. La théorie qui sous-tend cet estimateur est très bien établie et il existe des estimateurs de la variance par linéarisation. Deville et Särndal (1992) ont montré que, pour une famille d'estimateurs par calage, tous les membres de la famille sont asymptotiquement équivalents à l'estimateur RG. Cela suggère l'utilisation de l'estimateur de la variance de l'estimateur RG, ou une adaptation de ce dernier, pour d'autres estimateurs par calage. C'est ce type d'estimateur de la variance, similaire au cas RG, qui est généralement mis en œuvre en pratique.

Certaines méthodes de calage comportent des contraintes sur l'étendue des poids, comme le calage ridge. La linéarisation en présence de telles contraintes n'est pas toujours simple, et l'estimateur habituel de la variance de l'estimateur RG peut ne pas convenir. L'objectif de la présente recherche était de trouver un autre estimateur de la variance dans cette situation.

Nous avons débuté ce projet en 2023-24 et avons élaboré un estimateur de la variance bootstrap qui tient correctement compte des contraintes sur les poids et utilise des totaux de contrôle adéquats. Une étude par simulations comparant quatre approches bootstrap pour estimer la variance d'un estimateur par calage ridge a été complétée. Dans cette étude, nous avons considéré deux populations différentes et quatre tailles d'échantillon. Les résultats de cette étude ont démontré l'importance d'utiliser les contraintes et les totaux de contrôle adéquats afin de bien estimer la variance et l'erreur quadratique moyenne.

Progrès :

Nous avons considéré un nouvel estimateur bootstrap où les contraintes sur les poids nécessitent un ajustement supplémentaire lorsque le calage est fait sur les totaux de populations à chaque réplique bootstrap. Les résultats montrent que l'approche développée en 2023-24 où les poids bootstraps sont calés sur les estimations Horvitz-Thompson est légèrement plus stable que la nouvelle approche. Nous avons également rédigé un article expliquant la théorie et contenant les principaux résultats de l'étude par simulations. L'objectif est de soumettre cet article dans une revue statistique en 2026-27.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Keven Bosa (keven.bosa@statcan.gc.ca).

Bibliographie

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

PROJET : Pondération automatique pour la non-réponse

La méthode des scores est une méthode populaire utilisée pour corriger les poids d'enquête afin de neutraliser les effets de la non-réponse dans les enquêtes sur échantillon. Parmi les éléments clés de la méthode des scores, il y a la spécification d'un modèle de non-réponse qui modélise la propension de réponse de chaque unité en fonction de l'information auxiliaire disponible. Dans certains cas, notamment lorsque l'on dispose de peu de renseignements sur la relation entre le mécanisme de réponse et les covariables disponibles, spécifier avec efficacité ce modèle de réponse peut s'avérer difficile et prendre beaucoup de temps. C'est pourquoi nous avons étudié l'utilisation de différents algorithmes d'apprentissage automatique afin de remplacer cette étape de modélisation dans la méthode des scores.

Ces travaux s'appuient sur ceux de Larbi, Tsang, Haziza et Dagdoug (2025), qui ont étudié l'utilisation de méthodes d'apprentissage automatique pour corriger la pondération. Ce projet vise à analyser l'effet d'une stratégie de pondération, fondée sur la méthode des scores et l'apprentissage automatique, sur le biais relatif et l'efficacité des estimations d'enquête. De plus, d'autres aspects ont été intégrés à l'analyse, notamment l'effet du recours à la validation croisée pour sélectionner le « meilleur » modèle, l'agrégation de modèles à l'aide de la régression LASSO et de réseaux neuronaux, ainsi que diverses méthodes de formation de groupes de réponses homogènes (GRH).

Progrès :

Ce projet est maintenant terminé. Nous avons mené une étude par simulation afin d'étudier notre question de recherche. Nous avons créé un ensemble de données synthétiques comprenant une variable d'enquête, des variables auxiliaires et des indicateurs de réponse générés selon un mécanisme de réponse défini. Nous avons également mis au point un flux de travaux automatisé qui sélectionne des variables auxiliaires pour un modèle de réponse et corrige des modèles de réponse à l'aide 1) d'une régression logistique, 2) de méthodes d'apprentissage automatique choisies parmi un ensemble de méthodes reposant sur des mesures de validation croisée et 3) d'une méthode agrégée qui englobe toutes les méthodes d'apprentissage automatique envisagées. Plusieurs méthodes de regroupement des unités en GRH ont été mises à l'essai. Le biais relatif et l'efficacité de Monte Carlo ont été calculés pour chaque estimation pour 1 000 répliques afin de comparer le rendement des différentes méthodes de modélisation.

Les résultats de ce projet ont démontré qu'un modèle agrégé (empilé) peut constituer une solution de rechange valable au modèle de régression logistique et pourrait potentiellement offrir une certaine robustesse par rapport à l'erreur de spécification du modèle. Certains modèles d'apprentissage automatique ont également démontré leur potentiel pour la modélisation de la réponse, mais la généralisation des résultats de cette étude présente certaines limites, car peu de mécanismes de non-réponse ont été mis à l'essai. Par ailleurs, ce projet a mis en lumière les répercussions de la méthode utilisée pour créer les GRH sur la correction des biais et l'efficacité.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Gavi Arganosa (gavriel.arganosa@statcan.gc.ca), **Ilya Hekimi** (ilya.hekimi@statcan.gc.ca) ou **Oleksii Volkov** (oleksii.volkov@statcan.gc.ca).

Bibliographie

Larbi, K., Tsang, J., Haziza, D. et Dagdoug, M. (2025). [Utilisation de méthodes d'apprentissage automatique pour le traitement de la non-réponse totale dans les enquêtes](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025001/article/00003-fra.pdf). *Techniques d'enquête*, 51(1), 303-333. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025001/article/00003-fra.pdf>.

PROJET : Plans d'échantillonnages pour enquêtes sociales intégrant la propension de réponse

Plusieurs étapes du processus d'enquête se fondent sur l'hypothèse que les propensions de réponse réelles d'une enquête, ρ_i , sont exactement ou approximativement bien estimées par un modèle de propension $\hat{\rho}_i$ pour toutes les unités i de l'échantillon. C'est le point de départ de l'ajustement de la non-réponse et de plusieurs stratégies de collecte adaptative.

On observe que les facteurs influençant la réponse sont relativement stables entre les enquêtes sociales. Ce constat suggère que des indicateurs de propensions de réponse pourraient aussi être intégrés dès le plan d'échantillonnage, à partir de modèles élaborés lors de cycles antérieurs et de l'information auxiliaire disponible sur la base de sondage. Cette pratique étant peu répandue, ce projet en a exploré les intérêts et limites.

Progrès :

Plusieurs explorations théoriques ont été menées, comme la dérivation d'une allocation optimale pour un plan d'échantillonnage stratifié aléatoire simple sans remise avec les strates confondues avec les groupes de réponse homogènes. Cette allocation a également été étendue pour des plans à forte entropie.

Des analyses reposant sur la fonction de perte d'efficacité introduite par Kish (1976), similaire à l'effet de plan, ont été effectuées pour quantifier la perte selon l'homogénéité de la propension de réponse dans la population ou due à une approximation inexacte du modèle de propension. On déduit que, par exemple, l'optimalité est atteinte lorsque la propension de réponse estimée est proportionnelle à la propension de réponse réelle. L'allocation optimale ajustée par les propensions de réponse est plus efficace lorsque les strates sont intra-homogènes et inter-hétérogènes par rapport aux propensions de réponse.

Des simulations ont été développées pour mesurer les gains d'efficacité selon différents scénarios de modèle et de strates homogènes par rapport à la propension de réponse. Des gains en efficacité entre 5% et 10% sont attendus, selon la distribution des propensions et la qualité de leur modélisation.

Ce projet continuera. Il visera à intégrer dans l'allocation un indicateur réaliste et pratique de coût et à valider les gains en efficacité par la simulation d'un contexte réaliste de production d'enquête sociale.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Gautier Gissler (gautier.gissler@statcan.gc.ca).

Bibliographie

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, 139, Part 1, 80-95.

PROJET : Méthodes d'élagage des poids pour les données d'enquête

Habituellement, le processus de pondération des données d'enquête comporte trois grandes étapes : 1) attribuer un poids de base; 2) corriger les poids pour la non-réponse totale; 3) modifier les poids pour assurer l'uniformité entre les estimations de l'enquête et les totaux de population connus, opération souvent appelée calage. Dans certains cas, une dernière modification est apportée aux poids : on les élague afin d'améliorer l'efficacité des estimations d'enquête en limitant la variabilité des poids et en garantissant ainsi une plus grande stabilité des estimations.

Les procédures d'élagage exigent la spécification d'une constante de réglage, où les poids supérieurs à la constante sont réduits à la valeur de la constante et le poids excédentaire est redistribué entre les unités restantes. Souvent, les procédures utilisées en pratique sont ponctuelles; il manque en effet à l'heure actuelle une méthode formelle et théorique pour traiter l'élagage de poids. Ce projet a les objectifs suivants : 1) élaborer et évaluer des procédures d'élagage des poids pour traiter de multiples variables d'enquête; 2) effectuer une vaste étude par simulation pour comparer le rendement de plusieurs procédures d'élagage des poids (voir Potter, 1988, 1990; Potter et Zheng, 2015; Haziza et Beaumont, 2017) pour ce qui est du biais et de l'efficacité et pour évaluer le rendement des estimateurs de l'erreur quadratique moyenne pour ce qui est du biais; puis 3) appliquer et démontrer la méthodologie proposée sur un ensemble de données existant.

Progrès :

Une étude par simulation a été conçue et réalisée dans R. On a utilisé le système généralisé de calage de Statistique Canada, Yoho, et une contrainte de poids maximal au cours du processus de calage. Cette étude compare 10 méthodes d'élagage des poids en fonction du biais et de la variance lorsqu'elles sont appliquées à des variables d'enquête catégoriques et continues présentant des degrés de corrélation variables avec les poids. Les résultats montrent que, en l'absence de corrélation, il n'y a pas grand-chose à perdre en élaguant les poids, et qu'un élagage plus dynamique a tendance à offrir les réductions de variance les plus importantes sans compromettre le biais. Lorsque l'on introduit une corrélation, les méthodes qui élaguent le plus grand nombre de poids ne donnent plus les résultats escomptés; ce sont au contraire celles qui en élaguent un petit nombre qui ont le meilleur rendement. Cette différence est d'autant plus marquée à mesure que les poids sont plus dispersés. La méthode de l'erreur quadratique moyenne (EQM) minimale prend en compte certaines variables d'enquête importantes et vise à sélectionner un seuil d'élagage qui permet d'obtenir l'EQM combinée la plus faible pour celles-ci. Cette méthode révèle une adaptabilité et offre un bon compromis en ce qui concerne l'EQM pour les variables d'enquête retenues. Le rendement des méthodes était semblable, que l'on utilise des variables continues ou catégoriques présentant la même corrélation avec les poids.

Ces méthodes ont été appliquées à un ensemble de données fourni par l'équipe de l'Enquête sociale générale et l'analyse est en cours. Les prochaines étapes consisteraient notamment à achever l'exemple basé sur des données réelles, à comparer d'autres méthodes, comme le lissage des poids, et à mesurer l'effet sur l'estimation de la variance à l'aide de poids bootstrap.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Jody Krahn (jody.krahn@statcan.gc.ca) ou

David Haziza (dhaziza@uottawa.ca).

Bibliographie

Haziza, D. et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 453-458.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Potter, R. et Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

PROJET : Estimation de la variance à l'aide de méthodes d'imputation avancées

L'apprentissage automatique est utile pour l'imputation de données d'enquête manquantes, car il permet de bien rendre compte des relations complexes. Pour estimer la variance due à l'imputation (VDI), Statistique Canada utilise le Système d'estimation de la variance due à la non-réponse et à l'imputation (SEVANI) (Beaumont, Bissonnette et Bocci, 2010; Beaumont et Bissonnette, 2011). La variance étant un indicateur clé de la qualité des données, ce système doit être à la hauteur des nouvelles méthodes. Afin de combler cette lacune, ce projet vise à explorer différentes options permettant d'estimer la VDI lorsque l'on utilise des modèles à arborescence.

Progrès :

Nous avons examiné deux approches pour estimer la VDI lorsque l'imputation est effectuée à l'aide de méthodes d'apprentissage automatique. La première étend le SEVANI aux arbres de régression et aux forêts aléatoires. Nous avons ajusté ces modèles avec les progiciels R libres rpart et ranger, puis nous avons décomposé la structure des modèles afin de déterminer dans quelle mesure chaque répondant avait contribué à chaque valeur imputée. La deuxième repose sur la méthode du bootstrap par pseudo-population (BPP) (Chen, Haziza, Léger et Mashreghi, 2019), qui vise à répéter l'ensemble du processus, de l'échantillonnage à l'estimation.

Dans le cas du SEVANI, nous avons eu recours à l'ajustement croisé afin de limiter le surajustement dans les estimations d'erreur des répondants. Nous avons ensuite utilisé le modèle du plus proche voisin pour estimer les erreurs des valeurs imputées, comme dans Beaumont et coll. (2010). Dans le cadre de notre simulation, cela a permis de réduire le biais des estimations de variance lorsque les modèles d'imputation utilisés étaient des arbres de régression et des forêts aléatoires. Dans le cas de la méthode du BPP, plutôt que de décomposer les arbres de régression et le modèle Cubist, nous avons répété l'ensemble du processus pour toutes les répliques bootstrap. Cette approche a donné de bons résultats pour les deux méthodes d'imputation dans la simulation. Dans l'ensemble, le SEVANI reste solide lorsqu'il est possible de facilement trouver les contributions. La méthode du BPP s'avère pratique lorsque la décomposition est complexe, comme dans le cas du modèle Cubist, mais le calcul prend plus de temps en raison du processus de bootstrap.

L'étude par simulation met en évidence deux prochaines étapes. Tout d'abord, nous pouvons examiner la façon dont le matériel d'intelligence artificielle (IA), comme les processeurs graphiques, peut faciliter l'estimation de la variance. Le traitement parallèle et réparti peut accélérer le processus et rendre les deux méthodes plus pratiques à l'échelle d'un organisme. Par conséquent, les avantages que procurent les outils d'IA pourraient avoir des retombées positives et renforcer les méthodes statistiques classiques. Nous pourrions également étudier les moyens de réduire l'incidence des modèles d'imputation

inadéquats sur l'inférence. Parmi les options possibles, il y a le recours à une méthode d'inférence indépendante du modèle.

Pour obtenir plus de renseignements, veuillez communiquer avec :

John Tsang (john.tsang@statcan.gc.ca) ou

Darren Gray (darren.gray@statcan.gc.ca).

Bibliographie

Beaumont, J.-F. et Bissonnette, J. (2011). [Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11605-fra.pdf). *Techniques d'enquête*, 37(2), 183-192. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11605-fra.pdf>.

Beaumont, J.-F., Bissonnette, J. et Bocci, C. (2010). SEVANI, Guide de méthodologie. Statistique Canada.

Chen, S., Haziza, D., Léger, C. et Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika*, 106(2), 369-384.

PROJET : Estimation à partir d'échantillons non probabilistes à l'aide de plusieurs enquêtes de référence

Ce projet avait pour objectif de proposer une méthode de calcul des pseudo-poids pour des enquêtes non probabilistes à partir d'au moins deux échantillons probabilistes de référence.

Progrès :

Le mécanisme de participation aux échantillons non probabilistes est inconnu, et les paramètres cibles estimés à partir de ces échantillons peuvent être touchés par un biais de sélection. Les approches actuelles mises au point pour faciliter l'inférence statistique à partir d'échantillons non probabilistes ne permettent pas de prendre en compte plus d'un échantillon de référence. Dans cet article, nous proposons un cadre général permettant de tenir compte du biais de participation dans les échantillons non probabilistes à l'aide de plusieurs enquêtes de référence. Nous nous concentrons principalement sur les estimateurs par calage. Ces estimateurs offrent une grande souplesse dans les situations où l'accès à des microdonnées d'enquête est limité et sont faciles à mettre en œuvre en pratique. Nous décrivons deux méthodes d'estimation de la variance qui tiennent compte de toutes les sources de variabilité des estimateurs proposés. Nous évaluons le rendement des différentes méthodes dans le cadre d'une vaste étude par simulation, qui a démontré que l'estimateur par calage selon la méthode du ratissage croisé offrait des résultats satisfaisants dans des situations où les probabilités de participation étaient très dispersées dans des échantillons non probabilistes, et qu'il produisait des estimations de variance nettement plus faibles pour les résultats continus. Enfin, nous illustrons l'application de ces méthodes à l'aide de données tirées d'une étude réelle menée auprès d'adultes actifs au Canada.

Au cours de l'année 2025-2026, nous avons publié l'article de Landsman, Wang, Carrillo-Garcia, Mitani, Smith, Graubard, Bui et Carnide (2026) dans la revue *Statistics in Medicine*.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Ivan Carrillo-Garcia (ivan.carrillo-garcia@statcan.gc.ca).

Bibliographie

Landsman, V., Wang, L., Carrillo-Garcia, I., Mitani, A.A., Smith, P.M., Graubard, B.I., Bui, T. et Carnide, N. (2026). Correction for participation bias in nonprobability samples using multiple reference surveys. *Statistics in Medicine*, 45, 3-5, e70403. Accessible à l'adresse : <https://doi.org/10.1002/sim.70403>.

4 Confidentialité et accès aux données

Les travaux de recherche sur la confidentialité à Statistique Canada continuent à mettre l'accent sur l'élaboration de nouvelles méthodes et idées qui offrent d'autres formes d'accès tout en continuant à garantir que les renseignements privés des particuliers et des entreprises ne sont divulgués d'aucune façon. Des progrès ont été réalisés dans le cadre des projets décrits ci-dessous. L'équipe responsable du Centre de confidentialité et d'accès aux données à Statistique Canada a également continué d'offrir des services de consultation aux partenaires internes et externes pour contribuer au renforcement des capacités en matière de détermination et de traitement des risques de divulgation (voir la [section 5.5](#)).

PROJET : Données synthétiques

Les travaux se sont poursuivis à Statistique Canada en vue d'élaborer des cadres de gouvernance pour les données synthétiques. Les données synthétiques peuvent permettre d'aborder les problèmes de confidentialité liés aux données personnelles tout en conservant la plus grande valeur analytique possible. Les données synthétiques peuvent être particulièrement utiles lors de la recherche d'occasions de collaboration avec des intervenants externes qui n'ont peut-être pas accès aux microdonnées confidentielles.

Progrès :

Un groupe de travail interne a été mis sur pied afin de mieux comprendre les possibilités d'utilisation des données synthétiques et de moderniser les documents d'orientation pour favoriser leur utilisation et leur élaboration. Il s'agit entre autres de mettre à jour les principaux textes de gouvernance ainsi que de créer un manuel plus pragmatique (Gauvin, 2026).

Pour obtenir plus de renseignements, veuillez communiquer avec :

Héloïse Gauvin (heloise.gauvin@statcan.gc.ca) ou

Steven Thomas (steven.thomas@statcan.gc.ca).

Bibliographie

Gauvin, H. (2026). Handbook for the creation of public synthetic data files (version préliminaire). Document interne, Statistique Canada.

PROJET : Stratégies d'optimisation pour la suppression de cellules complémentaires

La méthode de suppression de cellules complémentaires (SCC) qu'utilise Statistique Canada dans le cadre d'un grand nombre de ses enquêtes économiques est une méthode typique permettant de supprimer les cellules sensibles lors de la diffusion de données quantitatives tabulaires. Cette méthode est bien élaborée

et appuyée par le logiciel SAS G-Confid de Statistique Canada. Les méthodes reposent sur l'optimisation par programmation linéaire, qui permet d'obtenir des solutions de suppression optimales garantissant que les modèles de suppression sont valides et réduisent le plus possible la quantité de renseignements supprimés.

Progrès :

Statistique Canada a poursuivi le développement d'une solution basée sur Python pour G-Confid. Le lancement du logiciel à l'interne est prévu pour juin 2026. Le logiciel mis à jour comprend une solution optimale d'arrondissement additif, un outil de calcul des mesures de sensibilité, la création d'un modèle optimal de SCC et un programme de vérification pour s'assurer que les modèles sont valides. On a évalué les solutions d'optimisation libres offertes par le progiciel Python PuLP pour savoir si elles pouvaient permettre de trouver des solutions optimales pour le modèle de SCC. Après un examen attentif des options, la solution Python a été mise en œuvre avec succès pour traiter les cas les plus complexes à Statistique Canada. Le groupe a également étudié d'autres options visant à garantir qu'un modèle de suppression valide puisse être trouvé selon les contraintes propres aux problèmes à grande échelle. Les solutions combinent la mise à l'échelle des microdonnées, la modification des contraintes, et le recours aux options de solveurs existantes et à l'infrastructure informatique.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Steven Thomas (steven.thomas@statcan.gc.ca) ou

Haoluan Chen (Haoluan.chen@statcan.gc.ca).

PROJET : Évaluation des facteurs fondés sur l'utilité et le risque de divulgation au regard de la mise en œuvre de la confidentialité différentielle dans le contexte de la diffusion de données tabulaires

Les stratégies de contrôle de la divulgation qui permettent d'assurer la confidentialité différentielle (CD) font toujours l'objet d'études à Statistique Canada, tant sur le plan de leurs répercussions pratiques sur l'utilité des renseignements diffusés que sur celui des garanties en matière de risques. Nous continuons de concentrer nos efforts sur la diffusion des données du recensement canadien.

Progrès :

Un article (Miller, 2026) a été rédigé; il comprend une analyse des différentes versions de la CD, leurs avantages et inconvénients, ainsi qu'une analyse formelle des propriétés fondamentales de la CD. Une interprétation statistiquement intuitive de la CD a également été proposée. Selon les principales conclusions de cet article, la CD offre des garanties solides et rigoureuses en matière de protection des renseignements personnels, mais possiblement au prix d'une perte importante d'utilité par rapport à ce que Statistique Canada tolère habituellement. La mise en œuvre d'un cadre de CD s'accompagne de nombreux défis pratiques, son adoption n'est donc pas une mince affaire. De plus, nous avons conclu que l'adoption à grande échelle de la CD n'est pas, à l'heure actuelle, possible sur le plan opérationnel dans le contexte de la diffusion des données du recensement. Toutefois, la CD pourrait actuellement être facilement appliquée aux données administratives lorsque les statistiques à diffuser sont des dénombrements et que le nombre de façons distinctes dont chaque participant fournit ses renseignements est faible. Il serait également envisageable de produire des ensembles de données synthétiques et de les diffuser dans des outils de totalisation à distance, comme le système d'accès à distance en temps réel. Nous n'excluons pas la possibilité de découvrir à l'avenir des algorithmes

conformes à la CD offrant des compromis plus attrayants entre la confidentialité et l'utilité. Si cela devait se produire, nous pourrions réexaminer l'adoption d'algorithmes conformes à la CD.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Joshua Miller (Joshua.miller@statcan.gc.ca) ou

Steven Thomas (steven.thomas@statcan.gc.ca).

Bibliographie

Miller, J. (2026). *A Privacy-Utility Based Study of Differential Privacy*. Document de travail du service Modern Statistical Methods and Data Science Branch (SSMD-2026-001E).

5 Soutien (centres de ressources)

5.1 Centre de recherche et d'analyse en séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques maintient une expertise de haut niveau dans le domaine des séries chronologiques (désaisonnalisation, estimation de la tendance-cycle, rapprochement des données, étalonnage, couplage, calendarisation, modélisation, prévisions et prévisions immédiates, et autres). Le Centre offre des services spécialisés visant à garantir que les méthodes d'analyse des séries chronologiques utilisées à Statistique Canada sont appropriées et vont de pair avec les avancées récentes. Il développe également des logiciels d'analyse et de traitement des séries chronologiques et offre des services, des consultations et de la formation en matière de séries chronologiques à l'échelle de l'organisme ainsi qu'à des clients externes.

Les projets peuvent être répartis en quatre sous-sujets :

- Consultation et formation en séries chronologiques;
- Méthodes relatives aux séries chronologiques;
- Soutien et amélioration du système et des outils de traitement de séries chronologiques;
- Soutien méthodologique au programme de l'Indice des prix à la consommation et au programme de l'Indice des prix à la production.

Progrès :

Consultation et formation en séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques est responsable d'élaborer des formations sur les méthodes de séries chronologiques, y compris la désaisonnalisation, l'établissement de données repères, le rapprochement des données et la modélisation de séries chronologiques, et de les offrir à des participants de Statistique Canada ainsi qu'à des participants d'autres organismes. Cette année, le Centre a offert des cours sur les composantes des séries chronologiques, la désaisonnalisation, l'établissement de données repères et le rapprochement des données aux participants internes et externes par l'entremise de l'Institut de formation de Statistique Canada (Statistique Canada, 2026). Les exemples utilisés dans le cadre du cours sur le rapprochement des données ont été convertis de SAS à R. Le Centre a également participé à des activités de sensibilisation et à des formations ponctuelles destinées à

d'autres groupes au sein de Statistique Canada sur des sujets liés aux séries chronologiques (cours spéciaux offerts à des groupes d'analystes, démonstrations du tableau de bord de la désaisonnalisation).

Par ailleurs, le Centre fournit des conseils et des consultations sur les projets de séries chronologiques en général pour les programmes de l'ensemble de Statistique Canada. Cette année, le Centre a offert ses services de consultation à divers partenaires internes, pour divers sujets : statistique de l'état civil (désaisonnalisation des séries hebdomadaires), analyse économique (établissement de données repères), comptes économiques nationaux (désaisonnalisation), comptes des industries (établissement de données repères), commerce international des services (désaisonnalisation), statistique de la population active (tableau personnalisé de données désaisonnalisées) et statistique du secteur public (prévisions immédiates et prévisions). Des représentants du Centre assistent également périodiquement à un forum hebdomadaire d'analystes, afin de maintenir une présence dans la communauté des analystes. Le Centre mène régulièrement des consultations sur la rétropolation pour préserver ou rétablir la comparabilité au fil du temps. De plus, le Centre a participé à des consultations avec des organismes publics fédéraux et provinciaux, des organismes nationaux de statistique, et d'autres organismes externes. Le Centre a notamment mené une consultation sur la calendarisation des données relatives au commerce de détail.

Méthodes relatives aux séries chronologiques

La pandémie de COVID-19 a eu une incidence considérable sur de nombreuses séries chronologiques économiques au cours de la période de 2020 à 2022. Les données étant maintenant disponibles jusqu'en 2025, l'incidence de la pandémie sur les tendances saisonnières, la tendance-cycle et la volatilité a été évaluée. À l'aide de séries chronologiques représentatives provenant de divers programmes statistiques (main-d'œuvre, commerce, fabrication, permis de construction et investissement en construction de bâtiments, tourisme, et Indice des prix à la consommation), le Centre a fait une étude afin de déterminer si ces séries sont revenues à des tendances prépandémie ou sont passées à un nouveau régime postpandémie. Les résultats de cette étude ont été présentés lors de l'atelier de 2025 de l'Organisation de coopération et de développement économiques et de l'Institut national de la statistique et des études économiques, qui portait sur l'analyse des séries chronologiques pour le domaine des statistiques officielles (Cheng, An et Guest, 2025).

En raison des modifications apportées aux règles d'admissibilité à l'assurance-emploi au début de 2017, deux ensembles d'options de désaisonnalisation sont actuellement utilisés pour les séries chronologiques du programme des Statistiques de l'assurance-emploi (SAE). Une étude menée par le Centre a démontré que cela n'était plus nécessaire et que la désaisonnalisation pouvait être effectuée à l'aide d'un seul ensemble d'options tout en conservant une qualité similaire. Cette mesure sera mise en œuvre lors de la prochaine révision historique des données de l'enquête SAE et permettra d'améliorer considérablement l'efficacité de l'examen annuel des options de désaisonnalisation pour ce programme.

Le Centre a continué d'évaluer l'apprentissage automatique afin de détecter la saisonnalité dans les séries chronologiques, un problème complexe dans le domaine de la désaisonnalisation. Les algorithmes mis à l'essai ont donné de bons résultats à partir des données simulées et même à partir des données qui n'avaient pas été utilisées lors de l'entraînement. Malheureusement, le rendement des algorithmes s'est considérablement dégradé lorsque des données réelles ont été utilisées. Le problème réside, d'une part, dans la quantité limitée de données étiquetées accessibles pour l'entraînement et, d'autre part, dans la difficulté de simuler des séries présentant des caractéristiques semblables à celles de séries réelles afin d'augmenter la quantité de données étiquetées. L'évaluation se poursuit.

Le Centre a achevé un nouvel ensemble de lignes directrices sur les méthodes de séries chronologiques à Statistique Canada, qui seront publiées en 2026-2027. Outre la mise à jour des lignes directrices antérieures, les sections consacrées aux techniques d'établissement de données repères et de rapprochement des données ont été étoffées, et une section sur la comparabilité temporelle a été ajoutée.

Soutien et amélioration du système et des outils de traitement de séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques élabore et tient à jour un certain nombre d'outils importants utilisés pour traiter et analyser les données de séries chronologiques pour les programmes de Statistique Canada produisant des données désaisonnalisées, en particulier le système généralisé G-Series, aux fins d'établissement de données repères et de rapprochement des données (ratissage et équilibrage) (Statistique Canada, 2016; Ferland, 2025), le Système de traitement des séries chronologiques (Ferland, 2022) et le tableau de bord de la désaisonnalisation (Verret, 2021).

Les travaux restants, portant sur une nouvelle version libre de G-Series dans R, ont été menés à bien cette année. La version 3 de G-Series a été officiellement lancée en juin 2025 sous la forme d'un progiciel de R appelé [gseries](#) (Ferland, 2025).

Le Système de traitement des séries chronologiques (STSC) est une application SAS personnalisable permettant d'appliquer des techniques de séries chronologiques, dont des techniques de désaisonnalisation, d'établissement de données repères et de rapprochement des données, largement utilisées dans la production d'estimations désaisonnalisées pour des sous-programmes annuels de Statistique Canada (essentiels au mandat dans de nombreux cas). Le système est dans un état mature et stable. Toutefois, il doit être constamment mis à jour, afin d'élargir ses fonctionnalités et de répondre aux nouveaux besoins des programmes de l'organisme. Tout comme G-Series, le STSC fait actuellement l'objet d'une refonte en R. Les travaux ont débuté cette année et devraient s'achever en décembre 2027. Cela offrira la souplesse nécessaire pour intégrer les outils et les nouvelles techniques qu'offrent les logiciels libres.

Cette année, une fonctionnalité a été ajoutée au tableau de bord de la désaisonnalisation afin de permettre aux analystes de filtrer facilement les séries chronologiques en fonction du fait que le mois de référence actuel présente une aberrance ou non selon le processus de désaisonnalisation. Le Centre a également offert de la formation à des analystes spécialisés.

Le Centre a continué à améliorer un système de désaisonnalisation de beaucoup de séries qui pourrait permettre de traiter un grand nombre de séries chronologiques sans intervention manuelle. La désaisonnalisation en découlant ne répondrait pas aux critères de qualité requis pour une publication officielle, mais pourrait offrir une qualité acceptable pour des produits expérimentaux.

Soutien méthodologique au programme de l'Indice des prix à la consommation et au programme de l'Indice des prix à la production

Le Centre de recherche et d'analyse en séries chronologiques compte également une sous-section chargée de fournir un soutien méthodologique au programme de l'Indice des prix à la consommation (IPC) et au programme de l'Indice des prix à la production.

Le Centre a poursuivi ses travaux visant à élaborer un ensemble d'indicateurs de qualité pour l'IPC. Des mesures de couverture ont été établies pour les quatre dimensions de l'échantillonnage : la couverture géographique, la couverture du marché liée aux points de vente, la couverture des produits et la couverture temporelle. Dans le cas des dimensions associées aux produits et au temps, on a également estimé le biais d'échantillonnage de l'IPC pour les principaux points de vente fournissant des données de lecteurs optiques.

Le Centre a étudié des approches de modélisation susceptibles d'offrir un cadre méthodologique solide permettant d'intégrer des données de rechange aux données d'enquête de certaines estimations de l'Indice des prix à la production (produits industriels, matières premières). Un cadre bayésien a été proposé et présenté à un comité d'examen scientifique interne. Cette approche a également fait l'objet de discussions internes avec des experts des statistiques bayésiennes et sera présentée à un comité d'experts internationaux (le Comité consultatif des méthodes statistiques de Statistique Canada) en juin 2026 (Francis, Dewan et Carrillo-Garcia, 2026).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Etienne Rassart (etienne.rassart@statcan.gc.ca).

Bibliographie

Cheng, B., An, L. et Guest, J. (2025). Assessing the impact of the COVID-19 pandemic on Statistics Canada's economic time series. Document interne, Statistique Canada, présenté à l'atelier 2025 de l'OCDE-INSEE sur l'analyse des séries chronologiques pour les statistiques officielles.

Ferland, M. (2022). Time series processing system – v3.08. Document interne, Statistique Canada.

Ferland, M. (2025). [gseries: Improve the Coherence of Your Time Series Data](https://StatCan.github.io/gensol-gseries/fr/). Librairie R, version 3.0.2. Accessible à l'adresse : <https://StatCan.github.io/gensol-gseries/fr/>. (la librairie [gseries](https://CRAN.R-project.org/package=gseries) est également disponible sur CRAN à : <https://CRAN.R-project.org/package=gseries>).

Francis, J., Dewan, A. et Carrillo-Garcia, I. (2026). *Bayesian Methods to Incorporate Auxiliary Price Data into Price Indices*. Présentation au Comité consultatif sur les méthodes statistiques, juin 2026, Statistique Canada.

Statistique Canada (2016). *G-Series 2.00.001 User Guides*. Document interne, Statistique Canada.

Statistique Canada (2026). [Formation](https://www.statcan.gc.ca/fr/evenements-formation/formation/cours-statistiques). Accessible à l'adresse : <https://www.statcan.gc.ca/fr/evenements-formation/formation/cours-statistiques>.

Verret, F. (2021). Le tableau de bord de la désaisonnalisation de Statistique Canada. *Recueil : Symposium 2021, Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société*, Statistique Canada, Ottawa, Canada.

5.2 Centre de ressources sur l'innovation et les outils de statistiques économiques

La Sous-section des systèmes généralisés fait partie de la Section de l'infrastructure du Programme intégré de la statistique des entreprises (PISE) et est chargée du soutien et de l'élaboration de trois systèmes

généralisés existants, c'est-à-dire **G-Sam** (échantillonnage), **Banff** (vérification des données statistiques) et **G-Est** (estimation), ainsi que de deux nouveaux systèmes, **Jasper** (le remplaçant libre de G-Sam) et **Yoho** (le remplaçant libre de G-Est).

Progrès :

Les projets de la Sous-section peuvent généralement être classés dans les catégories du soutien (pour les utilisateurs du système), de la recherche et du développement.

Elle a traité un volume habituel de demandes de soutien. Il s'agit notamment des demandes de soutien pour les systèmes actuels basés sur SAS, ainsi que pour les nouveaux systèmes libres basés sur Python (Banff) et R (Jasper et Yoho). Un nombre important de demandes de soutien portait sur la prestation de directives concernant les changements entre les différentes versions des systèmes et de conseils sur la migration de SAS vers des logiciels libres. Même si certaines demandes ont été résolues en quelques échanges de courriels, d'autres ont nécessité une participation plus approfondie et ont comporté des recommandations sur la mise en œuvre appropriée du système pour atteindre les objectifs statistiques. Vous trouverez ci-dessous un résumé de certaines demandes de soutien.

Équipe de G-Sam et de Jasper :

- A fourni de l'aide à diverses équipes d'enquête lors de la conversion de leurs programmes d'échantillonnage SAS utilisant G-Sam en programmes d'échantillonnage R utilisant Jasper. Parmi ces équipes figuraient celle du Relevé trimestriel des états financiers et celle du Compte d'urgence pour les entreprises canadiennes.

Équipe de Banff :

- A fourni des conseils et une assistance à l'équipe de l'Enquête sur les infrastructures publiques essentielles du Canada dans le cadre de la conversion de son processus de vérification et d'imputation SAS en un processus axé sur les métadonnées reposant sur le processeur Banff, basé sur Python. Dans le cadre du processus de migration, l'équipe de Banff a également travaillé sur deux modules faits sur mesure, *ApplyRules* et *CubistR*.
- A collaboré directement avec l'équipe du Fichier des familles T1 Plus afin de mettre au point un processus pour le processeur Banff permettant de traiter des règles de vérification déterministes, c'est-à-dire des règles si-alors. Cela comprenait une stratégie pour traiter les fichiers dont la taille dépassait celle de la mémoire dans le processeur Banff.

Équipe de G-Est et de Yoho :

- A fourni un soutien pour une demande de partenaires des Territoires du Nord-Ouest utilisant G-Est.
- A fourni une aide à l'équipe de l'Enquête sociale générale pour une migration de G-Est vers Yoho liée au calage.

Au cours de l'exercice, la majeure partie de la recherche et du développement a été consacrée à l'initiative de Transition vers les sources ouvertes de Statistique Canada. L'équipe a joué un rôle actif pour tenir informés les utilisateurs, au sein et hors de la direction, de l'état d'avancement des systèmes. Cela comprenait notamment des présentations au Comité directeur des systèmes généralisés, ainsi que lors de séminaires de la direction et des réunions du Conseil de planification du secteur. Un rapport d'étape pour l'ensemble des systèmes généralisés a été remis aux utilisateurs externes dans Gray (2026).

Le progiciel R **Jasper** a été lancé à l'interne, par l'intermédiaire de GitLab, le 31 octobre 2025. Il propose des fonctions libres de répartition et de sélection destinées à remplacer celles du système généralisé G-Sam, basé sur SAS. Ces fonctions vont au-delà des macros équivalentes de G-Sam et permettent de résoudre une plus grande variété de problèmes. Parmi les améliorations, on retrouve la résolution par optimisation d'un problème de répartition élargi ainsi que l'ajout d'algorithmes de sélection. Le progiciel Jasper a également permis de réduire considérablement le code base par rapport à celui de G-Sam, c'est-à-dire qu'il est environ 10 fois plus petit, et il a seulement deux dépendances externes, ce qui réduit les coûts futurs d'entretien et d'amélioration, tout en augmentant énormément le rendement. La méthodologie Jasper a été présentée au Comité d'examen scientifique de la direction (Stinner, 2026), et l'examen par les pairs des aspects mathématiques est en cours. Un guide de migration a été créé sur GitLab afin d'aider les utilisateurs actuels de G-Sam à migrer vers Jasper. Une publication externe sur GitHub et le Comprehensive R Archive Network est prévue l'année prochaine.

Les méthodes de stratification de G-Sam ont été converties en R sous la forme d'un ensemble distinct de fonctions, hors de Jasper, et ajoutées dans un dépôt GitLab. Une version bêta de ce dépôt a été mise à la disposition des utilisateurs internes, et la date de lancement prévue est en mai 2026. Les recherches sur de nouvelles méthodes de stratification (Chase, Gray et Stinner, 2026) se sont révélées prometteuses et se poursuivront au cours de la nouvelle année.

Le développement de la version Python de **Banff**, initialement lancée en janvier 2025, se poursuit. Dans le cadre du cycle de développement, diverses versions de production et versions bêta des procédures de Banff et des systèmes du processeur Banff ont été publiées à l'interne sur GitLab et à l'externe sur GitHub et PyPi. Des progrès importants ont été réalisés concernant les nouveaux modules du processeur Banff, afin de proposer de nouvelles méthodes d'imputation modernes. Le module *ApplyRules* applique des règles de vérification déterministes de type si-alors mises en œuvre à l'aide de la syntaxe Pandas, ce qui offre aux utilisateurs une approche simple fondée sur des métadonnées pour définir, exécuter et suivre des règles de vérification, parallèlement à des procédures de vérification et d'imputation plus complexes. Un deuxième module, *CubistR*, permettant de procéder à une imputation à l'aide de l'algorithme Cubist, a été élaboré en collaboration avec des membres de l'équipe d'imputation du PISE, dans le cadre d'un projet de recherche visant à explorer l'utilisation de l'imputation Cubist sur des données d'enquête (Mireuta, Bowditch, Gray, Kuromi, Ngabo Nsengiyaremye et Sivathayalan, 2026).

Enfin, une version bêta de **Yoho**, le remplaçant de G-Est en langage R, a été lancée le 23 décembre 2025. La version bêta comprend des fonctions de calage, d'estimation ponctuelle et d'estimation de la variance, ainsi que des répliques bootstrap. Les essais internes menés par l'équipe de développement et les utilisateurs de la version bêta montrent une uniformité avec les résultats de G-Est, ainsi qu'une amélioration considérable du rendement. Le code base de Yoho est remarquablement allégé par rapport à celui de G-Est : il ne représente qu'environ 1 % de sa taille, mesurée en kilo-octets. Yoho ne nécessite que deux dépendances externes, ce qui réduit les coûts d'entretien prévus. Une version officielle de Yoho, comprenant les fonctionnalités mentionnées ci-dessus, est prévue en juin 2026.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Darren Gray (darren.gray@statcan.gc.ca) ou

Pierre-Olivier Julien (pierre-olivier.julien@statcan.gc.ca).

Bibliographie

Chase, P., Gray, D. et Stinner, M. (2026). Stratification via hierarchical clustering. Présentation interne, Statistique Canada.

Gray, D. (2026). Statistics Canada's generalized systems are migrating to R and Python. *The Survey Statistician*, 93, janvier 2026.

Mireuta, M., Bowditch, F., Gray, D., Kuromi, G., Ngabo Nsengiyaremye, B. et Sivathayalan, A. (2026). Modern Imputation methods applied via Open-Source and BANFF. Rapport interne, Statistique Canada.

Stinner, M. (2026). *The Jasper Package – An Open-Source-Style Sampling System*. Présentation au comité d'examen scientifique de la Direction des méthodes statistiques modernes et de la science des données, Réunion 59, Statistique Canada.

5.3 Centre de ressources en couplage d'enregistrements

Les objectifs du Centre de ressources en couplage d'enregistrements (CRCE) sont d'offrir des services de consultation aux utilisateurs internes et externes de méthodes de couplage d'enregistrements, ce qui comprend la formulation de recommandations sur les logiciels et les méthodes à utiliser, et le travail de collaboration sur les applications de couplage d'enregistrements. Il facilite également la diffusion de renseignements sur les méthodes, les logiciels et les politiques de couplage d'enregistrements, ainsi que l'analyse de données couplées pour les personnes intéressées au sein ou à l'extérieur de Statistique Canada.

Progrès :

Au cours de l'année, l'essentiel des travaux méthodologiques a porté sur la maintenance et l'accompagnement des utilisateurs de la version 3.5 de G-Coup sur les serveurs SAS en environnement infonuagique. Un volume typique de demandes de soutien pour G-Coup a été traité par l'équipe du projet. La plupart d'entre elles ont été résolues par des suggestions sur la manière d'appliquer le système en pratique, mais plusieurs ont nécessité un rôle plus important.

L'équipe a également entrepris des travaux visant à ordonner et à mettre à jour les documents sur la stratégie de couplage et les données synthétiques dans le cadre de l'initiative d'accélération méthodologique. Les produits qui en découlent doivent s'ajouter aux ressources mises à la disposition des utilisateurs du futur logiciel Web G-Coup basé sur Python.

Par ailleurs, le CRCE a travaillé sur divers autres couplages probabilistes dans l'Environnement de couplage de données sociales. Ces couplages l'ont aidé à analyser le rendement des logiciels et des solutions à fournir. Des travaux ont également été menés concernant la réalisation et la planification du remaniement des couplages du Registre canadien du cancer, ainsi que la migration vers Python de la version régulière et de celle d'échantillonnage du modèle de groupe de population du Programme d'élaboration de données sociales longitudinales (PEDSL), en prévision de la cessation des activités du PEDSL, afin qu'il puisse être utilisé dans de futurs programmes, notamment la base de données sociodémographiques.

Les membres de l'équipe ont offert des cours officiels au Centre de formation de Statistique Canada, ont commencé à planifier un atelier destiné au personnel chargé des statistiques officielles de la cohorte colombienne du recensement et ont mis à jour les documents de référence et les documents sur les stratégies destinés aux utilisateurs des techniques de couplage probabiliste d'enregistrements.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Eric Hortop (eric.hortop@statcan.gc.ca).

5.4 Centre de ressources en analyse de données

Le Centre de ressources en analyse de données (CRAD) a pour objectif premier de fournir des conseils sur le bon usage des outils et des méthodes d'analyse de données et de promouvoir l'adoption de pratiques exemplaires dans ce domaine. Les services du CRAD, axés principalement sur les enquêtes, les recensements et les données administratives, visent les employés de Statistique Canada et d'autres ministères et organismes, ainsi que les analystes et les chercheurs du milieu universitaire et des centres de données de recherche.

Progrès :

Consultation

Des services de consultation ont été fournis à la demande de clients internes et externes. Du 1^{er} avril 2025 au 31 mars 2026, le CRAD a répondu à environ 40 demandes. Les questions présentaient divers degrés de complexité et portaient notamment sur l'analyse normalisée selon l'âge, le test de chevauchement d'intervalles de confiance, la régression logistique pour des événements rares à partir de données d'enquête, le regroupement de données provenant de plusieurs sources, les diagnostics de modèles, la correction de Fay dans l'estimation de la variance, les intervalles de confiance pour les petites proportions, la comparaison de sous-populations dépendantes, les comparaisons multiples, les marges prédictives et la variance des médianes. Le CRAD a également aidé des clients à mettre en œuvre des méthodes statistiques dans les logiciels R, SAS, SUDAAN et STATA. De plus, le CRAD a passé en revue des articles analytiques destinés à être publiés à l'interne et à l'externe.

Formation

Le CRAD a fait une présentation sur l'analyse de données d'enquête complexes, en français et en anglais, lors de l'Atelier d'interprétation de données de Statistique Canada.

Le CRAD a aussi présenté, en anglais, les séances sur la régression linéaire et logistique au moyen de données d'enquête complexes, dans le cadre du cours sur la modélisation statistique à Statistique Canada.

Le CRAD a préparé et animé les séminaires bilingues suivants au niveau de la direction :

- Éviter les pièges que peuvent présenter les intervalles de confiance : conseils et principes directeurs
- Résolution de problèmes méthodologiques assistée par l'IA : une étude de cas
- Répondre aux besoins de désagrégation des données en réalisant une stratégie de combinaison d'estimations supérieure à la moyenne

Collaboration

Le CRAD a collaboré avec le Secrétariat du Conseil du Trésor à l'élaboration de stratégies de mesure pour le projet de mesure du rendement en santé mentale en milieu de travail. Pour ce projet, les données recueillies dans le cadre des cycles de 2019, de 2020, de 2022 et de 2024 du Sondage auprès des fonctionnaires fédéraux ont été utilisées pour mesurer des variables latentes, comme les facteurs de risque psychologique et les comportements, et pour calculer les scores factoriels pour différents niveaux d'agrégation. Les scores factoriels établis pour le projet ont servi à créer le Tableau de bord de la santé mentale en milieu de travail dans la fonction publique fédérale : [Tableau de bord de la santé mentale — Canada.ca \(tbs-sct.gc.ca\)](https://tableau.bord.santemurale.ca/tbs-sct.gc.ca). Les modèles de mesure ont été élaborés au moyen d'une analyse factorielle et d'un modèle d'équation structurelle, comme l'ont expliqué Blais, Mach, Michaud et Simard (2020) et Blais, Michaud, Simard, Mach et Houle (2021). Cette année, le modèle a été amélioré grâce à l'ajout de certains facteurs. Tous les scores depuis 2019 ont fait l'objet d'une nouvelle estimation à l'aide du nouveau modèle.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Ryan Chepita (ryan.chepita@statcan.gc.ca) ou

Isabelle Michaud (isabelle.michaud@statcan.gc.ca).

Bibliographie

Blais, A.-R., Mach, L., Michaud, I. et Simard, J.-F. (2020). Analysis of the Public Service Employee Survey Items as Measures of the Psychosocial Risk Factors. Présentation au Workplace Mental Health Performance Measurement Steering Committee, 7 octobre 2020.

Blais, A.-R., Michaud, I., Simard, J.-S., Mach, L. et Houle, S. (2021). [Mesurer les facteurs psychosociaux en milieu de travail au sein du gouvernement fédéral](https://www150.statcan.gc.ca/n1/fr/pub/82-003-x/2021012/article/00001-fra.pdf). *Rapports sur la santé*, 32, 1-13. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/82-003-x/2021012/article/00001-fra.pdf>.

5.5 Centre de ressources pour la confidentialité et l'accès aux données

Le groupe de la méthodologie responsable de la confidentialité et des méthodes d'accès a continué d'offrir des services de consultation et de soutien aux partenaires internes et externes en ce qui concerne les diverses solutions d'accès et stratégies de prévention de la divulgation.

Données anonymisées

Le Centre de ressources pour la confidentialité et l'accès aux données a continué d'offrir gouvernance, orientation et conseils stratégiques en matière de production d'ensembles de données ouvertes. Cette année, plus de 12 nouveaux fichiers de microdonnées à grande diffusion ont été examinés et offerts sur le site Web de Statistique Canada. Statistique Canada a aussi continué d'offrir son expertise à plusieurs groupes à l'étranger et au Canada.

Consultation

À l'échelle nationale, nous continuons de collaborer avec Santé Canada et l'Agence de la santé publique du Canada pour élaborer des stratégies visant à diffuser davantage d'information. Nous avons également

collaboré avec l'Université de Sherbrooke et des intervenants internes afin de mettre au point des options d'accès modernes pour les analyses fédérées. Sur la scène internationale, nous consultons le Bureau australien de la statistique en vue d'améliorer notre méthodologie de suppression des cellules complémentaires, soutenue par notre outil libre G-Confid.

Rationalisation du processus de contrôle des données de sortie

Le Centre collabore avec ses collègues des centres de données de recherche pour élaborer des stratégies de contrôle normalisées visant à simplifier et à rationaliser le processus de contrôle des données de sortie. Ces normes ouvriront la voie à un contrôle plus poussé, et des processus automatisés sont envisagés pour l'avenir.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Steven Thomas (steven.thomas@statcan.gc.ca).

5.6 Centre de ressources en conception de questionnaires

Le Centre de ressources en conception de questionnaires (CRCQ) est le centre d'expertise de Statistique Canada en matière de conception et d'évaluation de questionnaires. Le CRCQ offre des services de consultation et de soutien et mène des recherches et des projets relatifs à l'élaboration, à la mise à l'essai et à l'évaluation de questionnaires d'enquête.

Il joue un rôle très important dans la gestion de la qualité et répond aux exigences de l'ensemble des programmes de Statistique Canada en consultant les clients, les répondants et les utilisateurs de données et en procédant à l'essai préliminaire des questionnaires d'enquête.

Même si une grande partie du travail du CRCQ est effectuée selon le principe du recouvrement des coûts, il est fréquemment sollicité, de manière ponctuelle, pour effectuer des évaluations d'expert et offrir des services de consultation relativement à un large éventail d'enquêtes. Le groupe propose aussi des cours sur la conception de questionnaires.

Progrès :

Le CRCQ a effectué de nombreux examens de questionnaires d'enquête. La plupart de ces examens portaient sur les questionnaires de Statistique Canada, mais plusieurs ont été réalisés pour des enquêtes menées par d'autres organisations gouvernementales, notamment le Service canadien d'appui aux tribunaux administratifs, la Banque du Canada, la Commission canadienne des droits de la personne, l'Agence du revenu du Canada, Ressources naturelles Canada et Services partagés Canada. Le groupe a également participé à diverses initiatives de consultation de l'organisme.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Jeremy Solomon (jeremy.solomon@statcan.gc.ca).

5.7 Secrétariat de l'éthique des données

Le Secrétariat de l'éthique des données a été mis sur pied pour appuyer le Comité d'éthique des données de l'organisme et déterminer les enjeux et les questions de nature éthique ayant trait aux initiatives de collecte de données à Statistique Canada. Il facilite également les activités en formulant des recommandations équilibrées, à l'intention du dirigeant principal de l'éthique des données, de la qualité

et de l'intégrité scientifique, sur les questions liées à l'éthique des données. Ces recommandations permettent de déterminer si un cas doit être soumis au Comité d'éthique des données (interne) et au Conseil consultatif sur l'éthique et la modernisation de l'accès aux microdonnées (externe).

PROJET : Éthique et intelligence artificielle

Ce projet vise à établir une distinction entre les problèmes difficiles et faciles dans le domaine de l'éthique de l'intelligence artificielle (IA) et de l'apprentissage automatique (AA). Il reflète une distinction bien connue dans la littérature sur la philosophie de l'esprit entre les problèmes difficiles et faciles de la conscience. Cette distinction est utilisée pour souligner l'importance des directives éthiques existantes et montrer la façon dont nous pouvons améliorer les chances de trouver des solutions concrètes dans le domaine de l'éthique de l'IA et de l'AA, ce qui est particulièrement utile pour les organismes régis par des instruments de gouvernance établis portant sur des questions comme l'intégrité scientifique ou l'éthique des données. Ces organismes ont déjà des attentes d'ordre éthique envers leurs membres. L'essence de ces attentes demeure pertinente même s'ils automatisent entièrement ou partiellement les tâches effectuées par des personnes à l'aide d'algorithmes d'IA ou d'AA. Nous pouvons remplacer la participation d'une personne dans un processus en utilisant des algorithmes, mais nous ne pouvons pas pour autant ignorer les directives éthiques existantes relatives à ce processus.

Progrès :

Le projet s'est conclu par la publication d'un article de Rochefort-Maranda (2026a) dans une revue à comité de lecture.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Guillaume Maranda (guillaume.maranda@statcan.gc.ca) ou

Ryan Chepita (ryan.chepita@statcan.gc.ca).

Bibliographie

Rochefort-Maranda, G. (2026a). The hard and easy problems of AI ethics: how to leverage existing guidance. *AI Ethics*, 6, 278. Accessible à l'adresse : <https://doi.org/10.1007/s43681-025-00907-1>.

PROJET : Défis d'éthique des organismes nationaux de statistique

Ce projet visait à comprendre la façon dont les défis d'éthique des organismes nationaux de statistique (ONS) peuvent se manifester, à l'aide de la littérature sur la psychologie morale, et plus particulièrement de celle sur le situationnisme. Les variables situationnelles peuvent influencer les actions ayant une portée morale. Il est possible de tirer des leçons importantes de cette situation pour améliorer les orientations éthiques. Cette étude est intéressante dans la mesure où l'émergence de défis d'éthique pour les ONS ne semble pas découler d'un changement au chapitre de la moralité des employés ni d'un manque de sensibilisation envers les responsabilités ou les principes pertinents.

Progrès :

Le projet s'est conclu par la publication d'un article de Rochefort-Maranda (2026b) dans une revue à comité de lecture.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Guillaume Maranda (guillaume.maranda@statcan.gc.ca) ou

Ryan Chepita (ryan.chepita@statcan.gc.ca).

Bibliographie

Rocheffort-Maranda, G. (2026b). Data ethics: Lessons from situationism for a National Statistical Organization. *Statistical Journal of the IAOS* (IOS Press/IAOS). Accessible à l'adresse : <https://doi.org/10.1177/18747655261439346>.

5.8 Secrétariat de la qualité

Le Secrétariat de la qualité veille à la production de données pertinentes de grande qualité provenant de sources éthiques en soutenant l'élaboration et la mise en œuvre de politiques et de procédures qui favorisent de saines pratiques en matière de gestion de la qualité. On a défini une nouvelle mission pour le Secrétariat qui sera examinée par le Comité de gestion stratégique de Statistique Canada.

PROJET : Indicateurs de qualité pour les données administratives

Alors que les indicateurs de qualité (comme l'exactitude) pour les données issues d'enquêtes sur échantillon sont bien établis, il n'existe actuellement aucun moyen uniforme et établi de fournir des indicateurs de qualité pour les données administratives. Bien que des recherches aient été menées à l'échelle internationale, les indicateurs de qualité improvisés qu'utilise Statistique Canada pour les programmes ingérant des sources de données de rechange ont été élaborés en grande partie au niveau des programmes. Souvent, ces indicateurs prennent la forme de métadonnées, compte tenu des difficultés que pose la quantification de la qualité d'un tel ensemble de données ou d'estimations. Ce projet a pour objectif de déterminer des indicateurs de qualité potentiels qui sont appropriés, robustes, et faciles à appliquer et à comprendre.

Progrès :

En nous appuyant sur des études internationales, notamment sur les travaux menés par un sous-groupe du Groupe d'experts des Nations Unies des cadres nationaux d'assurance de la qualité, nous avons défini une série de bons indicateurs de qualité potentiels. L'importance relative de chaque indicateur peut être déterminée par l'utilisateur des données, ce qui correspond au concept bien établi selon lequel la qualité tient de l'adéquation des données à leur utilisation. Il a été proposé que les indicateurs pondérés puissent ensuite être combinés pour former une mesure globale. Cette proposition générale a été présentée au groupe d'accélération méthodologique lors de la séance de mars 2026. À la lumière des commentaires formulés par le groupe, les prochaines étapes consisteront à choisir un programme pour mettre à l'essai les mesures proposées, à mettre en œuvre le processus et à consigner les difficultés éprouvées, ainsi qu'à déterminer les éléments à conserver et ceux à modifier en vue d'applications futures.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Ryan Chepita (ryan.chepita@statcan.gc.ca) ou

Yuliya Romanyuk (yuliya.romanyuk@statcan.gc.ca).

PROJET : Indicateurs de qualité pour le Modèle générique du processus de production statistique

Dans le cadre des travaux en cours du Groupe d'experts des Nations Unies des cadres nationaux d'assurance de la qualité, un sous-groupe a été formé afin de déterminer indirectement des indicateurs de qualité correspondant à chaque sous-étape du Modèle générique du processus de production statistique (GSBPM). Le GSBPM décrit huit étapes chronologiques qui composent le processus de production statistique (détermination des besoins, conception, construction, collecte, traitement,

analyse, diffusion et évaluation), et chaque étape comporte de trois à huit sous-processus. La nécessité de moderniser les indicateurs de qualité s'explique par l'élargissement des possibilités pour certains sous-processus (p. ex. la manière dont les données sont ingérées) et la complexité croissante des types de données dont disposent les organismes nationaux de statistique et qui correspondent à leurs besoins.

Progrès :

Tout au long de 2025-2026, le Secrétariat de la qualité a joué un rôle actif en participant aux réunions mensuelles du sous-groupe. Il a collaboré avec un autre pays membre afin de déterminer indirectement des indicateurs de qualité supplémentaires et des pratiques exemplaires pour certaines étapes du GSBPM. Il a ensuite participé à l'examen d'autres étapes. À la fin de l'exercice financier, les coresponsables du groupe mettaient la dernière touche à l'ébauche d'un document dans le but de le rendre accessible à des fins de consultation en mai 2026.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Ryan Chepita (ryan.chepita@statcan.gc.ca).

6 Autres activités

6.1 Revue *Techniques d'enquête*

Techniques d'enquête est une revue statistique en ligne gratuite à comité de lecture publiée deux fois par année par Statistique Canada depuis 1975. Cette revue vise à publier des articles novateurs de recherche théorique ou appliquée, et parfois des articles de synthèse, qui offrent de nouvelles perspectives sur les méthodes statistiques présentant un intérêt pour les organismes nationaux de statistique et d'autres organismes statistiques. Les articles sont publiés gratuitement dans les deux langues officielles et sont accessibles à l'adresse suivante : www.statcan.gc.ca/techniques-enquete. Le [comité de rédaction](#) est composé de chefs de file de renommée mondiale du domaine des méthodes d'enquête issus des secteurs public, universitaire et privé.

Progrès :

Les numéros de juin 2025 et de décembre 2025 (numéros 51-1 et 51-2) ont été publiés. Le numéro de [juin 2025](#) est un numéro spécial célébrant le 50^e anniversaire de *Techniques d'enquête*. Il comprend deux articles avec commentaires. Le premier, rédigé par Carl-Erik Särndal, s'intitule « Progrès de la science et de la pratique des enquêtes : hier, aujourd'hui et demain », alors que le deuxième, rédigé par J.N.K. Rao et Sharon Lohr, s'intitule « Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage ». Ces deux articles sont suivis de commentaires d'éminents statisticiens d'enquête et d'une réponse. Dans le cadre de ce numéro spécial, nous avons lancé une nouvelle section de la revue consacrée aux entretiens. Ces articles sont fondés sur des conversations avec des statisticiens ou des méthodologistes d'enquête ayant eu un impact important sur le domaine des enquêtes par sondage. Le numéro spécial de juin 2025 comprend deux entretiens, un avec Ivan P. Fellegi, ancien statisticien en chef du Canada, et l'autre avec Geoffrey Hole, ancien directeur à Statistique Canada. Il comprend également sept articles sollicités rédigés par des spécialistes renommés de la statistique et de la méthodologie d'enquête.

Treize articles ont été publiés dans le numéro de [décembre 2025](#), dont une communication brève et l'article sollicité Waksberg de 2025 de Mike Hidioglou, s'intitulant « L'échantillonnage pour les enquêtes-entreprises à Statistique Canada ».

En 2025, 77 articles ont été soumis à la revue. Le nombre moyen de jours entre la soumission et l'évaluation initiale était de 46 jours. Tous les articles soumis, sauf deux, ont été évalués dans un délai de 150 jours, et 82 % d'entre eux ont été évalués dans un délai de 90 jours. Parmi ces 77 articles, 49 ont été rejetés, 20 ont été acceptés et 8 n'avaient pas fait l'objet d'une décision définitive (y compris les articles qui n'avaient pas été révisés par les auteurs avant la date limite) en date du 23 mai 2026. D'avril 2025 à mars 2026, les pages de *Techniques d'enquête* ont été consultées 51 204 fois.

Nous prévoyons actuellement publier deux numéros spéciaux en 2026 ou en 2027. Le premier comprendra certains articles présentés à la septième International Conference on Establishment Statistics, à Glasgow, en juin 2024. Paul Smith et Mojca Bavdaz sont les rédacteurs invités de ce numéro spécial, dont la parution est prévue en décembre 2026. Le deuxième numéro spécial, dont les rédactrices invitées sont Maria Rosaria Ferrante et Natalie Shlomo, est prévu pour décembre 2027. Il portera sur le thème suivant : « Façonner l'avenir de la statistique d'enquête à l'ère des données ».

Pour obtenir plus de renseignements, veuillez communiquer avec :

Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

6.2 Transfert de connaissances – Formation en statistique

Le Groupe de travail sur le développement de talent statistique a pour mandat principal la formation en statistique au sein de la Direction et de l'organisme. Plusieurs cours ont été offerts cette année, dont ceux reliés aux séries chronologiques, à la conception de questionnaires, à l'échantillonnage, au couplage d'enregistrements, à l'imputation, à la pondération, à la modélisation et à l'analyse statistique avec des données d'enquête, ainsi qu'un atelier d'introduction aux simulations.

Pour ce qui est des nouvelles activités, le groupe a continué à concevoir et à accorder la priorité à des activités d'apprentissage qui peuvent être élaborées en temps opportun et axées sur l'apprentissage actif. Cette année, le travail pour remanier le deuxième cours d'analyse statistique avec des données d'enquête ainsi que le cours sur l'estimation de la variance a continué.

Concernant la prochaine année, nous continuerons d'offrir les cours du cursus en fonction de la demande et de la disponibilité des enseignants. De plus, on vise offrir le cours sur l'estimation de la variance dans les deux langues officielles.

Le Groupe de travail sur le développement de talent offre divers types de possibilités de formation afin que les employés puissent jouir d'une certaine souplesse dans leur perfectionnement professionnel. En plus des activités mentionnées précédemment, il existe de nombreuses possibilités d'autoformation et d'autoapprentissage ainsi que des communautés de pratique.

Pour obtenir plus de renseignements, veuillez communiquer avec :

Keven Bosa (keven.bosa@statcan.gc.ca).

6.3 Symposium international sur les questions de méthodologie de Statistique Canada

Le *Recueil* du Symposium international de 2024 sur les questions de méthodologie de Statistique Canada, dont le thème était « Façonner l'avenir des statistiques officielles » a été publié en septembre 2025. Le *Recueil* est disponible sur notre site web au : <https://www150.statcan.gc.ca/n1/fr/catalogue/11-522-X>.

Statistique Canada organisera le prochain Symposium international sur les questions de méthodologie dans une année ultérieure à déterminer.

À titre de mesure intérimaire, on a proposé un atelier interne aux employés de Statistique Canada. Les sujets suggérés comprennent le leadership, l'intelligence artificielle responsable, la définition de la statistique officielle, l'intégration des données et la gouvernance.

Pour obtenir plus de renseignements, veuillez communiquer avec:

Peter Wright (peter.wright@statcan.gc.ca).

7 Documents de recherche parrainés par le Programme de recherche et développement en méthodologie

Beaumont, J.-F., Bocci, C., Bosa, K., Dasylva, A. et Maranda, G. (2026). Prediction inference for finite population totals or means. Soumis pour publication dans une revue à comité de lecture.

Bosa, K., Beaumont, J.-F., Bocci, C. et Sombo, S. (2025). The use of a random forest algorithm in small area estimation. Article présenté au Comité consultatif sur les méthodes statistiques, juin 2025, Statistique Canada.

Chase, P., Gray, D. et Stinner, M. (2026). Stratification via hierarchical clustering. Présentation interne, Statistique Canada.

Chehbouni, K., Haddou, M., Cheung, J.C.K. et Farnadi, G. (2026). Neither valid nor reliable? Investigating the use of LLMs as judges. Proceedings of NeurIPS 2025. Accessible à l'adresse : https://papers.nips.cc/paper_files/paper/2025/hash/829e8f32d76a6248815bf5b01633811d-Abstract-Position_Paper_Track.html.

Cheng, B., An, L. et Guest, J. (2025). Assessing the impact of the COVID-19 pandemic on Statistics Canada's economic time series. Document interne, Statistique Canada, présenté à l'atelier 2025 de l'OCDE-INSEE sur l'analyse des séries chronologiques pour les statistiques officielles.

Commission économique pour l'Europe des Nations Unies (CEE-ONU) (2025). [Framework on Responsible AI for Official Statistics](#). Accessible à l'adresse : https://unece.org/sites/default/files/2025-10/Responsible%20AI%20for%20Official%20Statistics%20Framework_ADSaMM%20Group_HLG-MOS_0.pdf.

Commission économique pour l'Europe des Nations Unies (CEE-ONU) (2026). UNECE handbook on Uncertainty quantification in ML-enhanced statistical inference - A practitioner handbook with applications in official statistics. Commission économique pour l'Europe des Nations Unies (CEE-ONU), HLG-MOS ModernStats Initiative. Accessible à l'adresse :

<https://unece.org/statistics/documents/2026/06/reports/uncertainty-quantification-ml-enhanced-statistical-inference>.

Dasylyva, A. (2025). A differentially private record linkage for the production of official statistics. Communication présentée au Comité consultatif sur les méthodes statistiques, octobre 2025, Statistique Canada.

Dasylyva, A. et Goussanou, A. (2026). A joint model of record duplication and linkage error for the Canadian Census of Population. Rapport interne, Statistique Canada, Ottawa.

Denis, N. et Haddou, M. (2026). Prediction-powered estimation: Unbiased model-assisted estimation. *Journal of Official Statistics*. Accessible à l'adresse : <https://doi.org/10.1177/0282423X261451320>.

Ferland, M. (2025). [gseries: Improve the Coherence of Your Time Series Data](https://statcan.github.io/gensol-gseries/fr/). Librairie R, version 3.0.2. Accessible à l'adresse : <https://statcan.github.io/gensol-gseries/fr/>. (la librairie [gseries](https://cran.r-project.org/package=gseries) est également disponible sur CRAN à : <https://cran.r-project.org/package=gseries>).

Francis, J., Dewan, A. et Carrillo-Garcia, I. (2026). *Bayesian Methods to Incorporate Auxiliary Price Data into Price Indices*. Présentation au Comité consultatif sur les méthodes statistiques, juin 2026, Statistique Canada.

Gauvin, H. (2026). Handbook for the creation of public synthetic data files (version préliminaire). Document interne, Statistique Canada.

Gray, D. (2026). Statistics Canada's generalized systems are migrating to R and Python. *The Survey Statistician*, 93, janvier 2026.

Haddou, M., Seimandi, T. et Del Monaco, A. (2025). [Fairness in Machine Learning: Companion paper to the Framework on Responsible AI for Official Statistics](https://unece.org/sites/default/files/2025-10/Companion%20Paper%20on%20Fairness%20in%20Machine%20Learning_Responsible%20AI%20Framework_ADSaMM%20group.pdf). Accessible à l'adresse : https://unece.org/sites/default/files/2025-10/Companion%20Paper%20on%20Fairness%20in%20Machine%20Learning_Responsible%20AI%20Framework_ADSaMM%20group.pdf.

Imbrogno, A. et Verret, F. (2025). Dependent random group variance estimation using Fay's method. Article soumis pour publication dans la revue *Techniques d'enquête*.

Landsman, V., Wang, L., Carrillo-Garcia, I., Mitani, A.A., Smith, P.M., Graubard, B.I., Bui, T. et Carnide, N. (2026). Correction for participation bias in nonprobability samples using multiple reference surveys. *Statistics in Medicine*, 45, 3-5, e70403. Accessible à l'adresse : <https://doi.org/10.1002/sim.70403>.

Miller, J. (2026). *A Privacy-Utility Based Study of Differential Privacy*. Document de travail du service Modern Statistical Methods and Data Science Branch (SSMD-2026-001E).

Mireuta, M., Bowditch, F., Gray, D., Kuromi, G., Ngabo Nsengiyaremye, B. et Sivathayalan, A. (2026). Modern Imputation methods applied via Open-Source and BANFF. Rapport interne, Statistique Canada.

Rochefort-Maranda, G. (2026a). The hard and easy problems of AI ethics: how to leverage existing guidance. *AI Ethics*, 6, 278. Accessible à l'adresse : <https://doi.org/10.1007/s43681-025-00907-1>.

Rochefort-Maranda, G. (2026b). Data ethics: Lessons from situationism for a National Statistical Organization. *Statistical Journal of the IAOS* (IOS Press/IAOS). Accessible à l'adresse : <https://doi.org/10.1177/18747655261439346>.

Statistique Canada (2026). [Formation](#). Accessible à l'adresse : <https://www.statcan.gc.ca/fr/evenements-formation/formation/cours-statistiques>.

Stinner, M. (2026). *The Jasper Package – An Open-Source-Style Sampling System*. Présentation au comité d'examen scientifique de la Direction des méthodes statistiques modernes et de la science des données, Réunion 59, Statistique Canada.

Toupin, M.-H. et Martin, V. (2026). [Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application au recensement canadien](#). *Techniques d'enquête*, 52(1), 209-233. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2026001/article/00005-fra.pdf>.

Verret, F. et Walker, B. (2025). Application of the Fay-Herriot model with correlated design errors: The Labour Force Survey experience. Actes de la réunion annuelle 2025 de la Société statistique du Canada (SSC), Saskatoon, Canada.

You, Y. (2026). Comparison and evaluation of small area EBLUP estimators under benchmarking with application and simulation study. Soumis à une revue à comité de lecture en vue d'une publication éventuelle.

You, Y. et Bosa, K. (2026). An empirical study on the use of informative priors for the hierarchical Bayes small area estimation with application. Article de recherche en cours de rédaction, destiné à être soumis pour une publication éventuelle.