Nº 12-206-X au catalogue
ISSN 1705-0812

Programme de recherche et développement en méthodologie : réalisations, 2024-2025

Date de diffusion : le 10 octobre 2025

Statistique Canada
Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Programme de recherche et développement

en

méthodologie

Réalisations, 2024-2025

Le présent rapport fait la synthèse des réalisations en 2024-2025 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, 2024-2025

Table des matières

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 probabilistes

Depuis plusieurs années, les organismes nationaux de statistique de plusieurs pays cherchent à réduire les coûts de collecte des données et le fardeau des répondants, tout en augmentant l'utilisation de sources de données de rechange et de techniques de prédiction modernes. À Statistique Canada, une idée actuellement explorée pour relever ces défis est de réduire la fréquence des enquêtes probabilistes et de remplacer les données d'enquête manquantes par des prédictions. Par exemple, une enquête probabiliste pourrait être menée tous les deux ans, au lieu de chaque année, 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. L'objectif de ce projet est d'étudier les deux questions suivantes : 1) Comment, et quelles hypothèses sont requises pour obtenir des prédicteurs approximativement sans biais de totaux ou de moyennes de population finie en l'absence, ou en présence limitée, de données d'enquête probabilistes?; 2) Comment estimer la qualité (variance de prédiction) de ces prédicteurs? Ce projet fait suite à un projet précédent de DaSylva, Beaumont, Bosa et Maranda (2023).

Progrès :

Nous avons examiné deux cas. Dans le premier cas, des données auxiliaires pour une période sans enquête sont disponibles dans un échantillon probabiliste, mais aucune donnée d'enquête n'est observée dans cet échantillon probabiliste. Nous avons montré que des prédicteurs approximativement sans biais de totaux ou de moyennes de population finie peuvent être obtenus à l'aide d'un modèle de travail linéaire, même dans le cas d'écarts par rapport à l'hypothèse de linéarité. Cette propriété surprenante exige que chaque observation dans l'échantillon d'entraînement soit correctement pondérée afin de s'assurer qu'une certaine contrainte de calage soit satisfaite. Nous avons également élaboré un estimateur simple de la variance de prédiction. Nous avons montré que ce prédicteur linéaire pondéré avait des propriétés semblables à celles d'un prédicteur par la méthode du plus proche voisin, ce qui explique sa robustesse. Quelques résultats de simulation ont été obtenus pour illustrer les propriétés des prédicteurs. Nous avons également commencé à élaborer une stratégie bootstrap pour estimer la variance de prédiction lorsque des prédictions d'apprentissage automatique sont utilisées comme solution de rechange aux prédictions par modèle de travail linéaire ou par la méthode du plus proche voisin.

Dans le deuxième cas, des données auxiliaires pour l'ensemble de la population sont disponibles, ainsi que des données d'enquête pour un petit échantillon probabiliste. La disponibilité de 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. Dans le cas de l'échantillonnage aléatoire simple, l'inférence par prédiction s'avère essentiellement équivalente à l'approche assistée d'un modèle, laquelle est bien connue dans la littérature.

Nos résultats théoriques et empiriques seront présentés à l'Italian Conference on Survey Methodology de 2025 en juillet 2025.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Dasylva, A., Beaumont, J.-F., Bosa, K. et Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. *Recueil de la Section des méthodes d'enquête*, Société Statistique du Canada, mai 2023.

1.2 Couplage d'enregistrements

PROJET : Couplage privé de microdonnées sur le commerce international avec erreurs de couplage

De nouveaux renseignements pourraient découler du couplage de microdonnées sur le commerce international entre les organismes statistiques nationaux. Des préoccupations relatives à la protection de la vie privée font actuellement obstacle, mais elles pourraient être résolues en tirant parti de technologies d'amélioration de la protection de la vie privée, comme les calculs sécurisés multipartites, l'intersection d'ensembles privés ou les enclaves sécurisées. Dans des travaux antérieurs, une solution a été proposée fondée sur une modification d'un protocole d'intersection d'ensembles privés de Dasylva, de Cubellis, de Fausti et Franssen (2025). Cette solution couple deux ensembles de données internationaux sans identificateur unique, évalue les taux d'erreurs de couplage avec un modèle et calcule des statistiques sommaires tout en tenant compte des erreurs de couplage. Toutefois, il comporte de nombreuses lacunes, comme une souplesse limitée en matière de couplage, des contraintes sur l'analyse statistique possible et le manque de contrôle de divulgation. Ces travaux visent à aborder ces problèmes en couplant les ensembles de données au sein d'une enclave infonuagique sécurisée.

Progrès :

Une méthodologie a été élaborée pour coupler des ensembles de données de façon probabiliste et confidentielle, au sein d'une enclave infonuagique sécurisée, tout en tenant compte de l'exactitude du couplage. À l'aide des données couplées, une régression linéaire ou logistique est effectuée, et certaines données synthétiques sont générées pour permettre une certaine exploration des données. De plus, toutes ces opérations sont menées en appliquant des techniques de protection de la confidentialité différentielle, afin de garantir que tous les produits statistiques soient protégés contre la divulgation statistique. La méthodologie a été testée avec succès sur une enclave infonuagique sécurisée. La méthodologie et les essais ont été décrits dans un rapport interne (Dasylva, Santos, Franssen, de Cubellis, de Fausti, Pappagallo, Berrios et Fitzsimons, 2025), qui a été accepté aux fins de publication dans le Statistical Journal of the International Association for Official Statistics.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Abel Dasylva (abel.dasylva@statcan.gc.ca).

Bibliographie

Dasylva, 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*, 41(2), 569-597, OnlineFirst, numéro spécial, https://journals.sagepub.com/doi/10.1177/0282423X251329407.

Dasylva, 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. À paraître dans *Statistical Journal of the International Association for Official Statistics*.

PROJET : Intervalles de confiance pour les taux d'erreurs de couplage d'enregistrements dans le cadre d'un échantillonnage systématique stratifié

La méthodologie de l'Environnement de couplage de données sociales (ECDS) estime les taux d'erreurs pour ses couplages à l'aide de la procédure suivante : l'ensemble de paires définies et rejetées est divisé en strates en fonction des centiles de la répartition du poids total de la paire et, au sein de ces strates, des échantillons (habituellement de taille 100) sont tirés selon un échantillonnage systématique (dans le cadre duquel les paires sont ordonnées en fonction du poids total de la paire). L'échantillon de paires ainsi sélectionné est envoyé aux examinateurs et, selon leurs décisions, les diverses paires de l'échantillon sont classées comme appariées ou non. À l'aide de ces classifications, nous estimons le nombre total de fausses correspondances ou de correspondances manquées dans chaque strate. Les totaux estimés servent à calculer les taux de fausses correspondances ou de correspondances manquées pour chaque strate et pour le couplage global.

Nous avons récemment décidé d'un estimateur de variance pour nos taux d'erreurs à partir d'un projet de recherche mené au printemps 2023 (Millar et Loewen, 2023). Une étape suivante déterminée au moment de choisir un estimateur de variance approprié consistait à estimer les intervalles de confiance. Deux facteurs font qu'il est difficile de créer des intervalles de confiance de nos taux d'erreur :

- L'un est la méthode d'échantillonnage que nous utilisons (échantillonnage systématique stratifié). Des plans d'échantillonnage complexes nécessitent des méthodes complexes pour construire des intervalles de confiance, car des méthodes simples présentent souvent des problèmes de couverture, contenant le paramètre réel de population à un taux inférieur à la valeur nominale. Aucun consensus n'existe sur la meilleure méthode à utiliser pour construire des intervalles de confiance pour des échantillons tirés d'un échantillon systématique stratifié.
- L'autre facteur est que nos taux d'erreurs (taux de fausses correspondances et de correspondances manquées) sont souvent proches de 0 (en particulier le taux de fausses correspondances). Il est bien connu que les intervalles de confiance simples comme Wald offrent une couverture particulièrement mauvaise pour des proportions proches de 0 ou 1 (Franco, Little, Louis et Slud, 2019 ainsi que Neusy et Mantel, 2016). Une méthode plus complexe est nécessaire pour créer des intervalles de confiance de sorte que la couverture de l'intervalle soit au moins au niveau nominal. Les strates dont la variance estimée est de 0 posent également un problème, car certains intervalles de confiance n'auraient pas de longueur dans ce cas, lors de l'utilisation d'un calcul d'intervalle de confiance de type Wald.

Le problème auquel nous faisons face est de trouver un intervalle de confiance approprié à utiliser pour indiquer la qualité des taux d'erreurs de couplage dans l'ECDS.

Progrès :

Les intervalles de confiance ont été mis à l'essai en fonction de la couverture du paramètre réel de population et de la longueur de l'intervalle de confiance. En fin de compte, l'intervalle Agresti-Coull a été choisi parce qu'il répondait à l'exigence de couverture, qu'il n'était pas considérablement plus long que tout autre intervalle répondant également à l'exigence de couverture, et qu'il était le plus simple de ces intervalles. Cela correspond à ce que Brown, Cai et DasGupta (2001) ont déclaré : l'intervalle Agresti-Coull fonctionne bien lorsque la taille de l'échantillon est supérieure à 40.

De plus, un rapport (Loewen et Jin, 2025a) et deux présentations ont été produits (Loewen et Jin, 2025b et 2025c). Loewen et Jin (2025a) décrivent en détail le contexte de nos calculs de taux d'erreur, les travaux menés pour trouver le meilleur intervalle de confiance et les résultats de notre étude. La première présentation (Loewen et Jin, 2025b) a été donnée au Comité d'examen scientifique. Elle décrit en détail le processus de recherche d'une bonne méthode d'intervalle de confiance. La deuxième présentation (Loewen et Jin, 2025c) a été faite dans le cadre d'un séminaire de division et décrit en détail notre travail et les résultats de notre étude.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Rylee Loewen (rylee.loewen@statcan.gc.ca).

Bibliographie

Brown, L., Cai, T. et DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101-117.

Franco, C., Little, R., Louis, T. et Slud, E. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 334-364.

Loewen, R. et Jin, N. (2025a). Confidence intervals for Social Data Linkage Environment Error Rates. Document interne, Statistique Canada.

Loewen, R. et Jin, N. (2025b). Confidence intervals for Social Data Linkage Environment Error Rates. Diapositives pour présentation au Comité d'examen scientifique, 23 mai 2025, Statistique Canada.

Loewen, R. et Jin, N. (2025c). Confidence intervals for Social Data Linkage Environment Error Rates. Diapositives pour le séminaire divisionnaire, 12 juin 2025, Statistique Canada.

Millar, G. et Loewen, R. (2023). Variance estimation for record linkage error-rates obtained via clerical review of systematic samples of linked pairs. Document interne, Statistique Canada (Document Overview: Variance estimation for record linkage error.docx).

Neusy, E. et Mantel, H. (2016). Confidence intervals for proportions estimated from complex survey data. *Recueil de la Section des méthodes d'enquête*, Société Statistique du Canada, juin 2016.

1.3 Estimation sur petits domaines

PROJET : Inférence hiérarchique bayésienne pour l'estimation sur petits domaines à l'aide de lois *a priori* non informatives et informatives

Le modèle de Fay-Herriot (FH) au niveau du domaine est souvent utilisé lorsque les estimateurs directs de paramètres de population sont instables en raison de la petite taille des échantillons. L'idée derrière le modèle de FH et les autres modèles d'estimation sur petits domaines (EPD) est d'emprunter de la force à d'autres domaines. Toutefois, lorsque le nombre de domaines est petit, les modèles de FH et d'autres modèles d'EPD ne donnent habituellement pas de bons résultats. L'idée de ce projet est d'emprunter de l'information non seulement d'autres domaines, mais également au fil du temps à l'aide d'une approche hiérarchique bayésienne (HB).

La modélisation HB est très populaire dans l'estimation sur petits domaines, et la spécification de la loi *a priori* est très importante dans cette approche. Dans le cadre de ce projet, nous étudions le rendement des estimateurs HB sur petits domaines à l'aide de lois *a priori* non informatives et informatives pour les paramètres de régression et les composantes de variance. Nous appliquons les modèles bayésiens de You et Chapman (2006) et de You (2021) aux données de l'Enquête sur la population active (EPA) du Canada et nous évaluons l'impact des lois *a priori* sur les estimateurs HB. Nous étudions l'impact de lois *a priori* informatives et non informatives, ainsi que correctes et incorrectes, 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 étudié des spécifications de lois *a priori* pour les paramètres de régression de modèles au niveau du domaine pour l'estimation sur petits domaines. Plus précisément, nous avons examiné l'utilisation de lois *a priori* non informatives et informatives en les appliquant à l'EPA et dans le cadre d'une étude par simulation. Un document de recherche (You et Bosa, 2025) a été produit et soumis à *Techniques d'enquête*. Ce document a été révisé en fonction des suggestions du rédacteur associé et des examinateurs. La version finale de l'étude a été acceptée par la revue et sera publiée dans son numéro de décembre 2025.

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. *Techniques d'enquête*, 47(2), 389-399. Article 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. À paraître dans le numéro de décembre 2025 de *Techniques d'enquête*.

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

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 estimateurs de variance fondés sur le plan peuvent être considérés comme des estimateurs de variances d'échantillonnage lissées, mais ils sont généralement instables pour des petits échantillons. Pour résoudre ce problème, les estimations de la variance fondées sur le plan sont habituellement lissées, souvent au moyen d'un modèle de lissage log-linéaire.

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; iii) elles produisent des prédictions qui demeurent toujours dans la fourchette des valeurs observées. Nous envisageons une procédure bootstrap pour l'estimation de l'erreur quadratique moyenne de prédiction.

Progrès :

Auparavant, nous avons élaboré et évalué, au moyen d'études empiriques, une version non paramétrique du meilleur prédicteur empirique (MPE) lorsque des forêts aléatoires sont utilisées pour remplacer des modèles entièrement paramétriques. La variable prédictive que nous proposons utilise 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 (mais sans l'éliminer) l'effort de modélisation.

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 sera présenté au Comité consultatif des méthodes statistiques de Statistique Canada en juin 2025, et nous élaborons actuellement un programme en R mettant en œuvre la méthodologie proposée.

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 estimer l'erreur quadratique moyenne d'estimations sur petits domaines qui ont été réconciliées

Le modèle d'estimation sur petits domaines (EPD) 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'estimation de l'emploi total pour une sous-division des provinces, les estimations directes sont très négativement corrélées dû au calage des poids. Pour appuyer les enquêtes dans la même situation que l'EPA, ce projet met l'accent sur les extensions du modèle de base de FH au cas de forte dépendance. En particulier, il vise à obtenir une bonne estimation de la matrice de variance-covariance d'échantillonnage des estimations directes, qui est présumée être connue dans le modèle de FH. Pour atteindre cet objectif, 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 (lissés) et d'une matrice variance-covariance des estimations directes agrégées (estimations provinciales dans le cas de l'EPA). Comme les estimations de totaux sont de bons candidats à l'application d'une méthode itérative du quotient, les estimations sur petits domaines réconciliées et non réconciliées sont étudiées.

Progrès :

Les mathématiques du problème de rétro-ingénierie et de sa solution ont été rendues explicites, ce qui a donné une matrice de variance-covariance de travail clairement définie. Les logiciels d'EPD existants, comme G-Est, ne s'appliquent qu'au modèle de FH de base. Pour tirer le meilleur parti des logiciels existants, en tenant compte de la matrice de variance-covariance de travail présumée, on a obtenu des formules d'erreur quadratique moyenne (EQM) d'estimateurs réconciliés et non réconciliés résultant de l'ajustement du modèle de base. Cette méthodologie a été appliquée pour la première diffusion officielle des estimations sur petits domaines de l'EPA. Les résultats ont été présentés lors du Symposium international sur les questions de méthodologie de 2024 (Verret et Walker, 2024).

Le modèle de FH étendu aux erreurs d'échantillonnage dépendantes a ensuite été envisagé pour obtenir des estimateurs plus efficaces. Des formules d'EQM correspondantes ont été élaborées pour les estimateurs réconciliés et non réconciliés. Le modèle étendu a été programmé, ainsi que de nombreuses caractéristiques de G-Est (sélection de variables rétrospectives, diagnostics, détection des valeurs aberrantes). Un diagnostic de colinéarité et le premier diagnostic de Lesage, Beaumont et Bocci (2022) ont également été généralisés au modèle étendu. Les avantages du modèle étendu par rapport au modèle de base ont été étudiés dans le contexte de l'EPA.

Pour obtenir plus de renseignements, veuillez communiquer avec :
François Verret (francois.verret@statcan.gc.ca) ou
Braedan Walker (braedan.walker@statcan.gc.ca).

Bibliographie

Lesage, É., Beaumont, J.-F. et Bocci, C. (2022). Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot. *Techniques d'enquête*, 47(2), 303-323. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00001-fra.pdf.

Verret, F. et Walker, B. (2024). Rétro-ingénierie d'un processus de réconciliation hypothétique pour estimer l'erreur quadratique moyenne (EQM) des estimations sur petits domaines réconcilié. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

2 Méthodes et applications de la science des données

PROJET : Grands modèles de langage comme outil de post-traitement de moteurs de reconnaissance optique de caractères

Dans le cadre de ce projet, nous examinons les défis et les progrès de la technologie de reconnaissance optique de caractères (ROC), particulièrement dans le contexte de la reconnaissance de l'écriture manuscrite, en explorant la façon dont les grands modèles de langage (GML) peuvent améliorer les données de sortie de ROC grâce au post-traitement. Malgré les améliorations notables apportées aux systèmes de ROC, les textes manuscrits continuent de présenter des défis importants en raison de la variabilité des styles d'écriture manuscrite, ce qui entraîne des erreurs ayant une incidence sur les tâches en aval, comme la synthèse de texte et la reconnaissance d'entités nommées (REN).

Progrès :

Dans le cadre de cette recherche, nous avons évalué le rendement de divers outils de ROC en matière de reconnaissance de textes manuscrits et nous avons constaté que le système de reconnaissance optique de caractères par transformateur (TrOCR) fournissait les meilleurs résultats par rapport aux outils populaires de ROC en source ouverte, comme PaddleOCR et EasyOCR, tandis que des solutions exclusives comme GPT-4o et Document Intelligence ont également fourni les meilleurs résultats en présentant les taux d'erreurs de caractères les plus faibles. L'étude a démontré comment de grands modèles de langage pouvaient améliorer grandement la qualité des données de sortie de ROC en corrigeant les incohérences grammaticales et les fautes d'orthographe, GPT-4o étant un outil post-traitement réduisant substantiellement ces erreurs trouvées dans le texte extrait à l'aide de Tesseract (moteur de ROC gratuit à source ouverte) sur des extractions d'échantillons de l'ensemble de données d'écriture manuscrite de l'IAM.

De plus, nous avons exploré et conclu que les cadres d'ingénierie rédactique (requête ou invite de commande) automatique, comme DSPy, se sont révélés plus efficaces que les stratégies de rédactique manuelle pour générer des requêtes riches en contexte et détaillées. Même si la combinaison d'outils de ROC spécialisés et de modèles de langage s'est révélée très prometteuse pour améliorer l'exactitude de la reconnaissance, le défi de combinaisons d'erreurs complexes est demeuré, même pour les GML avancés, qui excellaient dans les corrections simples, mais éprouvaient des difficultés avec les substitutions de mots combinées et des problèmes plus complexes.

Nous avons reconnu la popularité de l'utilisation des GML Vision comme outils de ROC au cours des derniers mois. Par conséquent, nous avons également exploré le rendement d'outils en source ouverte comme Florence et Phi-3.5 de Microsoft et les sources fermées (GPT-4o, Document Intelligence) comme outils de reconnaissance de textes manuscrits (HTR) pour extraire du texte manuscrit et imprimé. Ces deux types d'outils fournissaient des niveaux acceptables de texte extrait. Nous espérons continuer d'explorer ces solutions, car des outils de ROC viables combinés à une optimisation de la rédactique font partie de nos travaux futurs.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Oladayo Ogunnoiki (oladayo.ogunnoiki@statcan.gc.ca) ou
Johan Fernandes (johan.fernandes@statcan.gc.ca).

PROJET : Détection des erreurs dans les valeurs unitaires du Commerce international

Ce travail consistait à terminer et à publier un article décrivant en détail la méthodologie sous-tendant un nouveau système fondé sur l'apprentissage automatique visant à vérifier et imputer des données sur le commerce, en production depuis octobre 2023. La Division du commerce et des comptes internationaux (DCCI) traite des millions d'enregistrements d'importations mensuels, y compris des variables comme les codes du Système harmonisé (SH), la valeur et la quantité; la valeur unitaire (VU) étant dérivée de ces deux dernières. En raison de l'accent mis sur la vérification de la valeur aux fins de droits de douane, les champs de quantité sont souvent mal déclarés, ce qui entraîne des erreurs dans les quantités agrégées et des problèmes liés aux valeurs unitaires qui, par le passé, ont demandé beaucoup de temps aux analystes pour les régler.

L'ancien système de découpage de VU, qui remplaçait les valeurs extrêmes par des donneurs autour de la médiane, comportait plusieurs limites, notamment une correction excessive, des erreurs persistantes non corrigées, une charge de travail manuel élevée et une utilisation retardée. Le nouveau système d'apprentissage automatique, passé en revue en 2022 et lancé en 2023, aborde ces problèmes et a été bien reçu par les analystes. L'officialisation de la documentation des méthodes et la publication des travaux accroîtront la transparence et faciliteront le partage des résultats pour stimuler la collaboration et le partage des connaissances, d'autant plus que l'utilisation de l'apprentissage automatique dans les statistiques officielles demeure un phénomène relativement nouveau.

Progrès :

Le projet a été mené à bien selon les étapes suivantes :

- Une version provisoire complète de l'article a été préparée (Hatko et Sall, 2024).
- L'examen par les pairs et la révision institutionnelle sont en cours.
- Les prochaines étapes comprennent l'examen du travail aux fins de confidentialité et de publication.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Stan Hatko (stan.hatko@statcan.gc.ca).

Bibliographie

Hatko, S. et Sall, A. (2024). International Trade Unit Value Error Detection and Correction. Rapport interne, Statistique Canada.

PROJET : Mesures pour le texte généré par l'IA

Ce projet porte sur la façon dont la qualité des données de sortie de la génération automatique de texte (GAT) est évaluée en fonction de différentes tâches, méthodes et cas d'utilisation. À mesure que les technologies de GAT, en particulier les grands modèles de langage (GML), se répandent, leurs données de sortie exigent des méthodes d'évaluation robustes, interprétables, multidimensionnelles et propres aux tâches. Toutefois, il n'existe actuellement aucun cadre normalisé de conception de pipelines d'évaluation. Pour aborder ce problème, le projet examine les techniques d'évaluation existantes, présente une taxonomie conceptuelle fondée sur la décomposition fonctionnelle et propose un cadre de conception d'évaluation réutilisable. L'accent est principalement mis sur les tâches unimodales de génération de texte à texte (par exemple résumé, traduction automatique, dialogue), dans le but de fournir à la fois de la clarté analytique et des outils pratiques pour la conception de futures évaluations.

Progrès :

Le projet de recherche est finalisé. Nous avons examiné la documentation d'évaluation de la GAT, en déterminant les limites critiques des pratiques actuelles (biais de surface des mesures fondées sur des règles, incohérence des protocoles d'évaluation humaine) et en synthétisant les résultats dans un cadre structuré et extensible. La revue de la littérature a été effectuée à l'aide d'un protocole de type PRISMA, afin d'analyser 385 documents de recherche, en ciblant principalement la documentation récente, tout en réexaminant les travaux fondamentaux. Cet examen couvre des méthodes d'évaluation fondées sur des règles, des humains et des GML, en mettant l'accent sur la nature multidimensionnelle de la qualité de la GAT. Nous proposons une taxonomie modulaire faisant la distinction entre les composantes d'évaluation, comme les fonctions de notation, la transformation des données d'entrée et les cibles d'évaluation. Nous avons conçu un cadre pratique de conception de pipelines d'évaluation (Istrate et Robatian, 2025), comprenant un flux de travail d'évaluation par étapes, des listes de vérification propres à chaque étape et des recommandations de pratiques exemplaires pour améliorer la convivialité et la reproductibilité, tout en proposant des outils et artefacts conviviaux pour les professionnels et n'exigeant pas de connaissances théoriques approfondies. L'objectif est de commencer à expérimenter ce cadre dans de futurs projets comprenant des textes générés par des GML, en intégrant continuellement les leçons apprises et les pratiques exemplaires, et en améliorant les artefacts du cadre.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Alexandre Istrate (alexandre.istrate@statcan.gc.ca) ou
Damoon Robatian (damoon.robatian@statcan.gc.ca).

Bibliographie

Istrate, A. et Robatian, D. (2025). Evaluation of AI-Generated Text: Survey, Taxonomy and Practical Framework for Designing NLG Evaluation Pipelines. Rapport interne, Statistique Canada.

PROJET : Quantification de l'incertitude dans la classification des grands modèles de langage avec une prédiction conforme

Le présent projet vise à améliorer la fiabilité et l'exactitude de grands modèles de langage (GML) utilisés par Statistique Canada pour classer les tâches. En utilisant les techniques de prédiction conforme, nous cherchons à introduire des mesures de confiance quantifiables pour les prédictions faites par ces modèles avec des garanties théoriques. Cette recherche aide à améliorer la gestion des risques dans les processus de prise de décisions reposant sur des GML.

Progrès :

Notre mise en œuvre de la prédiction conforme pour les classificateurs de GML a démontré un compromis optimal entre une plus grande exactitude et l'automatisation pour les tâches de classification de GML, au moyen d'une couverture garantie de 90 %. Les GML ont atteint une exactitude de référence de 77 %. Toutefois, en appliquant la prédiction conforme aux prédictions de modèle et en automatisant le codage et la classification aux GML lorsqu'ils produisaient des prédictions très certaines (ensembles de prédictions de taille 1), 70,2 % des données étaient automatiquement classées avec une exactitude supérieure de 89,8 %. Les données restantes pouvaient également être codées automatiquement ou envoyées en vérification manuelle avec interaction humaine. Les résultats ont surpassé les approches de référence ne s'accompagnant pas de garanties théoriques. Cette technique fournit une quantification de l'incertitude statistiquement valide, traitant automatiquement des prédictions de grande fiabilité tout en acheminant intelligemment les cas ambigus aux fins d'examen manuel. Comme prochaine étape, cette approche pourrait être appliquée à la classification par code du Système de classification des industries de l'Amérique du Nord (SCIAN) en fonction des descriptions des entreprises.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Rafik Chemli (rafik.chemli@statcan.gc.ca).

PROJET : Données synthétiques pour les statistiques officielles

Ce projet porte sur les défis importants auxquels Statistique Canada doit faire face en raison de la baisse des taux de réponse aux enquêtes, du coût élevé de l'acquisition de données et de la nécessité croissante de statistiques détaillées sur de sous-populations rares. En raison de ces facteurs, il est difficile de produire des estimations statistiques solides et exactes. La solution proposée consiste à tirer parti de modèles probabilistes de diffusion de réduction du bruit de pointe pour générer de grands volumes de données tabulaires synthétiques reproduisant étroitement les répartitions de données du monde réel. Reconnaissant que les données synthétiques introduisent des biais, le projet intègre un processus de suppression de biais, inspiré par l'inférence par prédiction et l'estimateur par différence généralisé.

Progrès :

Le projet a été mené à bien selon les étapes suivantes.

- Des modèles de diffusion tabulaires ont été entraînés et appliqués à trois ensembles de données tabulaires à source ouverte.
- Les données synthétiques générées ont été analysées et comparées aux données d'origine, présentant de fortes similitudes en matière de distribution, tout en produisant des points de données entièrement uniques.

- Les estimations par échantillonnage non probabiliste fondées sur des données synthétiques ont produit des estimations présentant un biais relatif plus faible et des niveaux de couverture semblables ou légèrement réduits par rapport aux niveaux de référence.
- Pour le contexte d'échantillonnage probabiliste fondé sur le plan, un estimateur mathématique à deux degrés de suppression de biais a été proposé. Toutefois, cet estimateur n'est pas sans biais, et l'estimateur de variance n'est pas sans biais, et n'a également pas de forme explicite, ce qui rend cette approche non valide pour l'estimation fondée sur le plan.
- Nous concluons que les modèles de diffusion sont de puissants outils d'IA générative qui peuvent être utilisés pour la diffusion de microdonnées synthétiques. Les expériences semblent indiquer qu'ils maintiennent les propriétés statistiques de la répartition réelle des données sous-jacentes. Toutefois, trouver un estimateur approprié dans le contexte d'un plan de sondage probabiliste pour l'estimation de paramètres de population peut être hors de portée.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Nicholas Denis (nicholas.denis2@statcan.gc.ca).

PROJET : Guide de rédactique pour une utilisation efficace de grands modèles de langage

Le Guide de rédactique met l'accent sur l'élaboration d'un guide complet sur la rédactique afin d'aider les scientifiques de données, les chercheurs et les analystes à utiliser efficacement les grands modèles de langage (GML). L'objectif est de démystifier la rédactique et de fournir des pratiques exemplaires fondées sur des renseignements empiriques et les besoins opérationnels au sein de la communauté de la science des données fédérale. Ce guide sert de ressource pratique pour améliorer la fiabilité, la transparence et l'efficacité du déroulement des opérations axées sur les GML soutenant les programmes, l'élaboration de politiques et la prestation de services publics fédéraux.

Ce guide aborde les concepts fondamentaux, l'anatomie et la structure des requêtes, et présente des techniques de rédactique clés, comme la rédactique d'instructions, la rédactique de rôles et la rédactique « en quelques coups » (few-shot).

Progrès :

Le document a été créé et le contenu a été structuré en sections clés, notamment : 1) Principes fondamentaux de la rédactique; 2) Structure de la rédactique; 3) Techniques de rédactique (instructions, rôles, en quelques coups); 4) Pratiques exemplaires et embûches. Chaque section est appuyée par des exemples pratiques et des illustrations visuelles pour faciliter la compréhension et l'utilisation. Des commentaires sont actuellement recueillis auprès de multiples groupes internes, y compris des équipes techniques, d'IA responsable, et de capacité en matière d'IA et d'élaboration de programmes. Ce guide sera révisé et mis à jour au cours des prochains mois en fonction de ces commentaires, afin d'en assurer l'applicabilité générale et l'harmonisation avec les priorités organisationnelles.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Saptarshi Dutta Gupta (saptarshi.dutta-gupta@statcan.gc.ca).

PROJET : Création d'un ensemble de données structurées à partir de données non structurées à l'aide de grands modèles de langage

Ce projet porte sur l'utilisation de grands modèles de langage (GML) pour extraire des données structurées de documents PDF non structurés. Il évalue deux approches principales : 1) peaufiner des GML aux fins d'extraction directe de données; 2) utiliser la génération améliorée par récupération d'information (GARI) pour récupérer des fragments de texte pertinents avant l'extraction structurée. L'objectif est de créer des ensembles de données de grande qualité qui permettent d'autres applications d'apprentissage automatique et d'analyses statistiques.

Cette recherche vise à comparer le rendement de méthodes peaufinées et fondées sur la GARI à l'aide de l'ensemble de données CORD-19 (Wang, Lo, Chandrasekhar, Reas, Yang, Burdick, Eide, Funk, Katsis, Kinney, Li, Liu, Merrill, Mooney, Murdick, Rishi, Sheehan, Shen, Stilson, Wade, Wang, Wang, Wilhelm, Xie, Raymond, Weld, Etzioni et Kohlmeier, 2020), un corpus de plus de 1 000 000 d'articles scientifiques liés à la COVID-19. Des exemples de requête (par exemple « Quelle était la proportion de tous les patients asymptomatiques atteints de la COVID-19? ») sont utilisés pour évaluer l'exactitude de l'extraction par rapport à la réalité de terrain connue.

Progrès :

Méthodologie : L'ensemble de données CORD-19 a été utilisé pour élaborer et mettre à l'essai des approches peaufinées et fondées sur la GARI pour l'extraction structurée de données. Les questions de réalité de terrain ont été conçues pour permettre l'évaluation du rendement. Des ajustements ont été effectués à l'aide de modèles comme gpt-4o-mini et gpt-4.1-mini, tandis que des pipelines de GARI personnalisés ont été mis en œuvre en parallèle.

Résultats : L'évaluation montre que les modèles ajustés produisent des réponses plus cohérentes et quantitatives axées sur des résultats précis. En revanche, les données de sortie de la GARI fournissent des réponses plus générales et plus descriptives, combinant parfois des détails qualitatifs et quantitatifs. Bien que les deux méthodes fournissent des réponses comparables dans de nombreux cas, la variance de la qualité de la réponse dépend du type de requête. L'étape suivante consiste à mettre à l'essai le cadre au sein d'un ensemble plus diversifié de sources non structurées afin de valider le caractère généralisable et la robustesse des résultats.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Saptarshi Dutta Gupta (saptarshi.dutta.gupta@statcan.gc.ca).

Bibliographie

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O. et Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *ArXiv preprint*, https://doi.org/10.48550/arXiv.2004.10706.

PROJET : Cadre de protection des renseignements personnels d'entrée et de sortie – Couplage d'enregistrements assurant la protection des renseignements personnels

Le projet de Cadre de protection des renseignements personnels d'entrée et de sortie vise à concevoir et à mettre en œuvre des pipelines de protection des renseignements personnels robustes de bout en bout aux fins d'analyse et de couplage d'enregistrements de données de nature délicate en matière de protection des renseignements personnels. L'objectif central de cette initiative est de permettre à deux parties ou plus d'analyser et de coupler en collaboration des ensembles de données (comme des dossiers personnels ou organisationnels) sans divulguer de renseignements de nature délicate au niveau de la personne. Le projet est axé sur la protection de la vie privée à la fois à l'étape de l'entrée, de l'intégration (veiller à ce que les données soient protégées pendant le transfert et le calcul) et à l'étape de sortie (veiller à ce qu'aucune divulgation n'ait lieu dans les résultats publiés), en tirant parti des technologies d'amélioration de la confidentialité (TAC), en particulier le chiffrement homomorphe (HE), et les méthodes traditionnelles de contrôle de la divulgation statistique (CDS).

Deux principaux volets technologiques sont explorés : un pipeline fondé sur la cryptographie utilisant le chiffrement homomorphe et le calcul multipartite sécurisé, et un pipeline fondé sur des enclaves déployant des environnements matériels sécurisés. Le principal cas d'utilisation démontrant ces technologies est le couplage d'enregistrements assurant la protection des renseignements personnels, dans le cadre duquel des résumés statistiques peuvent être calculés à partir de l'ensemble de données couplé de façon sécurisée. Ce cadre est très pertinent pour des contextes comme les soins de santé, les finances ou les statistiques officielles, dans le cadre desquels les organisations doivent se conformer à des règlements rigoureux en matière de protection des renseignements personnels.

Progrès :

Des progrès importants ont été réalisés l'an dernier en ce qui concerne la conception, la mise en œuvre du prototype et la validation expérimentale du pipeline fondé sur la cryptographie. Un protocole de couplage d'enregistrements assurant la protection des renseignements personnels de bout en bout a été élaboré, qui combine des techniques de chiffrement homomorphe avancées et des mesures de protection en matière de protection des renseignements personnels des produits.

Un prototype de démonstration a été mis en œuvre pour permettre à un client et à un serveur de déterminer en toute sécurité les chevauchements d'enregistrements (correspondances exactes ou floues) et de calculer des résumés statistiques chiffrés (par exemple, la moyenne par strate) sans que l'une ou l'autre des parties ne découvre les données de nature délicate de l'autre partie. Des sous-protocoles critiques veillent à ce que les tailles d'intersections inférieures à un seuil établi soient supprimées pour éviter les atteintes à la vie privée et à préserver la sécurité des produits sommaires, comme l'exigent les lignes directrices sur le contrôle de la divulgation. L'évaluation expérimentale d'ensembles de données synthétiques a démontré la faisabilité du système pour des ensembles de données de taille moyenne, des produits statistiques exacts et un rendement de calcul raisonnable. De plus, ce cadre est conçu pour être modulaire et extensible pour des applications futures et des données de sortie de base. De la documentation et un rapport technique (Shukla, Rossiter et Santos, 2024), ainsi que la mise en œuvre de la démonstration, sont disponibles pour un examen plus approfondi et une analyse comparative.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Benjamin Santos (benjamin.santos@statcan.gc.ca).

Bibliographie

Shukla, A., Rossiter, E. et Santos B. (2024). Input/Output Privacy Framework: A Privacy-Preserving Record Linkage System. Rapport interne, Statistique Canada.

PROJET : Diffusion de données synthétiques avec garanties de confidentialité différentielle – Solution de rechange pour les fichiers de microdonnées à grande diffusion

Ce projet porte sur la production de données synthétiques avec garanties de confidentialité différentielle comme solution de rechange robuste aux fichiers de microdonnées à grande diffusion (FMGD) traditionnels. L'objectif est de permettre aux chercheurs, aux scientifiques des données et au public d'effectuer des analyses utiles de données synthétiques réalistes, tout en protégeant rigoureusement la confidentialité des personnes. Cette initiative vise à comparer l'efficacité de données synthétiques avec des garanties de confidentialité différentielle par rapport aux techniques classiques de contrôle de la divulgation actuellement utilisées avec les FMGD. Bien que les données synthétiques soient reconnues comme un mécanisme plus sécuritaire de partage de données, des recherches récentes indiquent que seule la confidentialité différentielle offre des protections solides et démontrables contre la réidentification sur les données synthétiques et potentiellement contre les menaces émergentes, comme celles que représentent les grands modèles de langage.

Ce cadre met l'accent sur les FMGD canadiens largement utilisés, diffusés par Statistique Canada dans le cadre de l'Initiative de démocratisation des données. En tirant parti de modèles génératifs avancés, y compris des modèles de diffusion et des réseaux antagonistes génératifs, ainsi que de méthodologies statistiques et d'évaluation de la protection des renseignements personnels robustes, le projet vise à produire des ensembles de données synthétiques protégés, à confidentialité différentielle ayant une validité analytique élevée. L'une des composantes cruciales est l'élaboration d'un progiciel en source ouverte Python, inspiré de la trousse d'outils *Folktables* (Ding, Hardt, Miller et Schmidt, 2021), visant à fournir des données synthétiques pour étalonner les modèles d'apprentissage automatique et faciliter la recherche reproductible.

Progrès :

Le projet a permis de mener à bien une revue de littérature approfondie, a déterminé les principaux FMGD à expérimenter et a examiné des méthodes de pointe pour des données à confidentialité différentielle et des données synthétiques génératives. L'équipe a lancé la mise en œuvre de nouveaux modèles génératifs (y compris des modèles de diffusion et des réseaux antagonistes génératifs) intégrés à des mécanismes de confidentialité différentielle, de concert avec le développement parallèle d'une trousse d'outils Python en source ouverte pour fournir des ensembles de données synthétiques à partir de FMGD à des fins d'étalonnage et d'analyse. Les premières étapes des évaluations mettent l'accent sur les principales mesures d'utilité et de confidentialité, en examinant le compromis entre l'utilité analytique et la confidentialité, tout en relevant le défi constant consistant à veiller à ce que les données synthétiques générées maintiennent la validité pour des tâches typiques d'apprentissage automatique démographique et socioéconomique.

Une comparaison juste de la confidentialité et de l'utilité entre des FMGD et des données synthétiques nécessitera l'entraînement de générateurs de données synthétiques à l'aide de microdonnées dépersonnalisées, une étape qui sera effectuée une fois que l'accès sera sécurisé. Les produits livrables à

venir comprennent un rapport technique et un progiciel Python ouvert à l'usage du public et à des fins de référence.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Benjamin Santos (benjamin.santos@statcan.gc.ca).

Bibliographie

Ding, F., Hardt, M., Miller, J. et Schmidt, L. (2021). Retiring Adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

PROJET : Contrôle de la divulgation : limites et autres solutions

Dans son rôle d'organisme statistique national, Statistique Canada agit à titre d'intermédiaire pour recueillir ou agréger des données provenant de diverses sources et les compiler en tableaux sommaires ainsi qu'en ensembles de données plus détaillés. La fourniture de ces tableaux comprend toujours un risque de divulgation de renseignements personnels ou stratégiques de nature délicate. Pour éviter les fuites de renseignements de nature délicate, l'organisme déploie des programmes (comme G-Confid) visant à déterminer et à supprimer les cellules de nature délicate dans un tableau de contingence ou dans des ensembles de données plus détaillés comme des fichiers de microdonnées à grande diffusion (FMGD) ou les fichiers de microdonnées du recensement. L'objectif de ce projet était de cerner les vulnérabilités de ces méthodes, étant donné que le paysage de calcul a changé. Cela est non seulement attribuable à l'arrivée de grands modèles de langage et d'outils d'entrée de gamme facile d'accès, mais également à l'augmentation de la puissance de calcul, rendant certains calculs inversés moins difficiles.

Progrès :

Ce projet est terminé et un rapport a été soumis. Le projet portait principalement sur les tableaux de l'Entrepôt commun de données de sortie et la façon d'accéder aux valeurs supprimées. Nous avons résumé diverses méthodes d'attaque fondées sur l'expertise de l'adversaire. De plus, nous avons présenté des recommandations pour corriger ces vulnérabilités. En ce qui concerne les FMGD, nous avons relevé certaines attaques initiales, mais il faudrait obtenir l'autorisation pour effectuer un moissonnage du Web pour comparer des renseignements personnels sur des sites publics comme Facebook et LinkedIn. Par conséquent, nous n'avons fourni qu'une méthodologie (aucun résultat) pour les ensembles de données de FMGD.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Abhishek Shukla (abhishek.shukla@statcan.gc.ca).

3 Problèmes d'estimation dans les enquêtes

PROJET : Calcul des degrés de liberté théoriques pour les intervalles de confiance de Wilson modifiés dans le contexte du questionnaire détaillé du recensement de la population canadienne

Pour les utilisateurs des données des recensements de la population au Canada, les intervalles de confiance peuvent être utilisés pour mesurer la précision d'une estimation et pour la prise de décision.

Pour produire ces intervalles, on suppose une certaine distribution, dans notre cas, la loi de Student. Cette distribution prend comme paramètre un degré de liberté qui, normalement, est estimé par le nombre de répliques de poids dans l'échantillon. Cette approximation peut mener à de la sous-couverture des intervalles de confiance, habituellement observée dans les petits domaines. L'objectif du projet est de mettre en œuvre la formule théorique qui a été élaborée dans l'article de Toupin et Martin (2025), c'est-à-dire de trouver une approximation ou un estimateur à partir des développements théoriques, qui permettrait la mise en œuvre dans un système automatisé utilisé dans les différentes diffusions du programme du recensement.

Progrès :

Dans la dernière année, des développements algébriques de plusieurs approximations ont été faits dans le but d'évaluer leur performance. Un environnement de simulations Monte Carlo a été élaboré dans le logiciel SAS pour évaluer les différentes approximations. On a trouvé que plusieurs d'entre-elles amélioraient la couverture des intervalles de confiance pour l'estimation des chiffres à partir du questionnaire détaillé, en comparaison de la valeur fixe présentement mise en œuvre dans le système, mais rarement suffisamment pour atteindre le taux nominal visé. L'approximation la plus performante offre des gains trop limités pour justifier la complexité de son intégration dans le système de diffusion.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Samer Farfour (samer.farfour@statcan.gc.ca).

Bibliographie

Toupin, M.-H. et Martin, V. (2025). Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application à l'estimation du questionnaire détaillé du recensement canadien. *Techniques d'enquête* (en cours de révision).

PROJET : Estimation bootstrap de la variance pour des estimateurs par calage avec contraintes sur l'étendue des poids

Les estimations d'enquête sont souvent calées pour qu'elles correspondent à des totaux connus de la population, appelés totaux de contrôle. L'estimateur par la régression généralisée est un estimateur par calage fréquemment utilisé qui suppose une relation linéaire entre la variable d'enquête et les variables auxiliaires. La théorie qui sous-tend l'estimateur par la régression généralisée est 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 par la régression généralisée. Cela suggère l'utilisation de l'estimateur de variance de l'estimateur par la régression généralisée, ou une adaptation de ce dernier, pour d'autres estimateurs par calage. C'est ce type d'estimateur de variance qui est généralement mis en œuvre en pratique.

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

Ce projet a débuté au cours de l'année 2023-2024 où nous avons élaboré un estimateur de la variance bootstrap qui tient correctement compte des bornes sur les poids et des totaux de contrôle.

Progrès :

Nous avons élaboré et terminé une nouvelle étude par simulation dans laquelle nous comparons quatre approches bootstrap pour estimer la variance d'un estimateur par calage ridge. Nous avons examiné deux populations différentes et quatre tailles d'échantillon. Les résultats de cette étude montrent l'importance d'utiliser des contraintes ajustées et de correctement tenir compte des totaux de contrôle pour obtenir la version bootstrap de l'estimateur par calage ridge, et ainsi des estimateurs approximativement sans biais de la variance de l'estimateur par calage ridge.

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 : Étude sur les effets de méthodes de sélection des variables et du réglage de paramètres sur la méthode Cubist avec et sans prétraitement

Au sein de la Division des méthodes de la statistique économique (DMSE), et en particulier du Programme intégré de la statistique des entreprises (PISE), des méthodes d'imputation classiques sont actuellement utilisées pour imputer la non-réponse partielle et totale. Malgré la simplicité des méthodes individuelles, comme l'imputation par donneur ou par quotient, des dizaines, voire des centaines, de ces méthodes sont superposées pour créer des stratégies complexes de vérification et d'imputation qui deviennent difficiles à mettre en œuvre, à tenir à jour et à dépanner. En 2023-2024, dans le cadre du Programme d'accélération des processus méthodologiques (MethAx2023-24), nous avons mené des recherches exploratoires dans le cadre desquelles le rendement de méthodes de régression linéaire, de forêt aléatoire et Cubist a été examiné à la suite des résultats d'études par simulation menées par Dagdoug, Goga et Haziza (2023), où les auteurs ont déclaré, entre autres résultats, que la méthode Cubist (Kuhn et Johnson, 2016 et Kuhn, 2022) fonctionnait bien dans diverses situations et que la méthode de la forêt aléatoire fonctionnait bien dans des environnements à dimensions élevées. Nos recherches sur MethAx2023-24 ont également indiqué que les méthodes Cubist permettaient d'obtenir de bons résultats parmi les méthodes susmentionnées. Dans le cadre de cette recherche, nous avons comparé le rendement de la méthode d'imputation Cubist selon différents scénarios afin de déterminer les processus qui pourraient améliorer leur rendement.

Progrès :

Pour comprendre le rendement de la méthode d'imputation Cubist, nous avons mené une étude visant à comprendre les effets des méthodes de sélection de caractéristiques, comme l'élimination de variables de variance proches de zéro, de variables qui sont une combinaison linéaire d'autres variables, de variables hautement corrélées et de variables sujettes à la multicolinéarité, ce qui a donné quatre scénarios différents. Pour chaque scénario, nous avons étudié l'impact du réglage des paramètres avec et sans sélection de caractéristiques, avec et sans tolérance à l'amélioration. Cela a donné 16 scénarios différents (configurations). De plus, nous avons adapté la méthode Cubist à l'aide de deux fonctions R différentes, à savoir caret et tidymodels, ce qui a donné 32 configurations différentes. De plus, pour comprendre si le prétraitement des données améliore le rendement, nous avons appliqué ces scénarios

à des ensembles de données non normalisés et normalisés, où toutes les variables sont normalisées, et à des ensembles de données ne contenant que les variables importantes produites par l'algorithme Cubist ajusté précédemment.

Nous avons sélectionné quelques variables cibles à partir de trois enquêtes économiques différentes de Statistique Canada de différentes tailles. Les prédicteurs sont des variables de la base de sondage de l'année en cours et toutes les variables de l'année précédente. L'ensemble de données définitif contenait les répondants de l'année en cours et de l'année précédente. Les résultats ont été obtenus dans le cadre d'une validation croisée à 10 groupes. Notre programme R reproductible produit en sortie les valeurs prédites, l'erreur quadratique moyenne, le biais, la variance, l'erreur absolue moyenne et médiane, le pourcentage d'erreur absolue moyenne, le coefficient de détermination, le temps de calcul, le nombre de comités et de voisins, le cas échéant, selon la configuration et le groupe pour différents ensembles de données prétraités. Nos résultats indiquent que procéder à certaines activités de nettoyage des données à l'aide des méthodes de sélection des variables énumérées ci-dessus, en combinaison avec le réglage des paramètres (mais pas la variation des tolérances), et/ou aux activités de prétraitement des données susmentionnées ont amélioré le rendement des modèles Cubist pour certaines variables.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Ahalya Sivathayalan (ahalya.sivathayalan@statcan.gc.ca).

Bibliographie

Dagdoug, M., Goga, C. et Haziza, D. (2023). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology*, 11, 141-188.

Kuhn, M. et Johnson, K. (2016). *Applied Predictive Modelling*. Springer, 69-71.

Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-93, https://cran.r-project.org/package=caret.

PROJET : Méthodes bayésiennes non paramétriques pour des données d'enquête

L'application de méthodes non paramétriques bayésiennes est un domaine de recherche important pour les organismes statistiques. Dans le cadre de notre travail, nous étudions l'application de techniques fondées sur des processus de Dirichlet à des données d'enquête complexes. Nous proposons une nouvelle approche qui combine des méthodes existantes, utilisées par Statistique Canada, pour la dérivation de poids d'enquête pour des plans d'échantillonnage probabilistes stratifiés à plusieurs degrés avec probabilités inégales et la souplesse des méthodes bayésiennes non paramétriques fondées sur des processus de Dirichlet.

Nous avons développé des logiciels et appliqué de tels modèles à certains des ensembles de données recueillis au cours du sixième cycle de l'Enquête canadienne sur les mesures de la santé et avons mené plusieurs études de simulation pour montrer la robustesse de nos modèles dans divers scénarios ainsi que l'applicabilité de ces modèles à une production éventuelle de statistiques officielles.

Progrès :

Le projet est terminé. Les résultats obtenus indiquent que la méthode proposée présente un potentiel important de recherches supplémentaires. Des logiciels ont été élaborés et appliqués à des modèles pratiques. Un article (Volkov, 2025) a été publié dans The Survey_Statistician.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Oleksii Volkov (oleksii.volkov@statcan.gc.ca).

Bibliographie

Volkov, O. (2025). Bayesian nonparametric methods for survey data. *The Survey Statistician*, 91, 34-43, https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_January_N91.pdf.

PROJET : Sous-échantillonnage pour le suivi des cas de non-réponse dans les enquêtes sociales

Dans le contexte des défis que posent la baisse des taux de réponse et le coût élevé du suivi facilité par un intervieweur, il est important de tenir compte de l'ampleur du suivi des cas de non-réponse requis et de la façon de l'utiliser efficacement. Sélectionner uniquement un sous-échantillon d'un plus grand échantillon principal pour réaliser un suivi plus approfondi (complet) des cas de non-réponse, tout en effectuant un suivi moindre (réduit) pour les autres, offre l'occasion de réduire les sommes dépensées pour le suivi facilité par un intervieweur, puis de réallouer ce budget, ou, du moins une partie de celui-ci, à un suivi plus approfondi des cas de non-réponse de la partie de suivi complète de l'échantillon. Cela peut aider à gérer le potentiel accru de biais associé à de faibles taux de réponse ou peut viser à le faire sans compromettre la qualité des estimations. Ce type de plan d'enquête nécessite des ajustements du processus de pondération; l'estimateur multimodal concomitant (EMC) simultané a été élaboré expressément pour ce contexte (Mather, Boulet et Brennan, 2024). Ce projet vise à approfondir les travaux théoriques antérieurs en permettant d'étudier ses propriétés de façon empirique au moyen d'études pilotes et en évaluant les applications potentielles au moyen de simulations.

Progrès :

L'approche consistant à sélectionner un sous-échantillon aléatoire pour un suivi complet, puis à appliquer l'EMC a été étudiée au moyen d'un essai empirique et de simulations (Boulet et Mather, 2025). Un projet pilote a été mené dans le cadre de l'Enquête sociale canadienne. Une vague habituelle de cette enquête comporte un échantillon de 20 000 unités, toutes admissibles au suivi des cas de non-réponse par téléphone; dans le projet pilote, le nombre d'unités admissibles à ce suivi a été réduit à 17 000, mais la taille globale de l'échantillon a été augmentée pour passer à 34 000. Comme pour une vague habituelle, toutes les unités étaient admissibles au suivi des cas de non-réponse au moyen de rappels par la poste, de courriels et de messages texte. De par sa conception, le coût global relatif à la collecte était à peu près le même qu'une vague habituelle. Il en a résulté une augmentation de la précision par rapport à une vague habituelle (taille d'échantillon réelle plus grande, erreur-type plus faible pour les variables communes et augmentation du nombre de domaines répondant aux lignes directrices en matière de diffusion). De plus, malgré les signes indiquant que l'approche du sous-échantillonnage aléatoire et de l'application de l'EMC introduisent un certain biais, l'ampleur de ce biais potentiel est faible et est jugée comme se situant dans des intervalles acceptables, particulièrement en ce qui concerne les gains de précision.

De plus, une série de feuilles de travail de planification a été élaborée pour modéliser les coûts et les rendements de divers contextes d'enquête. Selon les hypothèses provenant de résultats empiriques et de simulations, le modèle tient compte des relations entre les taux de rendement attendus pour les sous-échantillons avec suivi complet et réduit, les coûts relatifs associés à un suivi complet et réduit et l'augmentation de l'effet de plan attribuable au sous-échantillonnage et à l'application de l'EMC. Les résultats portent à croire que cette approche pourrait être efficace dans certains contextes, afin d'accroître la précision en conservant les mêmes coûts de collecte ou de réduire les coûts de collecte pour un même niveau de précision. Cette approche semble particulièrement applicable aux enquêtes présentant des différences modérées, mais non extrêmes, tant en ce qui concerne le coût que le taux de réponse entre un suivi complet et réduit. C'est le cas pour de nombreuses enquêtes de Statistique Canada, comme celles dont le suivi réduit est effectué par courrier et par courriel et dont le suivi complet inclut des tentatives d'appel. Le modèle et les feuilles de travail de planification connexes sont prêts à être utilisés pour évaluer l'applicabilité de l'approche du sous-échantillonnage et de l'EMC à des enquêtes particulières pour lesquelles on souhaiterait mettre en œuvre cette idée.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Cilanne Boulet (cilanne.boulet@statcan.gc.ca) ou
Anne Mather (anne.mather@statcan.gc.ca).

Bibliographie

Boulet, C. et Mather, A. (2025). Methodology Acceleration Initiative Research Project: Subsampling for Non-Response Follow-Up in Social Surveys. Document interne, Statistique Canada.

Mather, A., Boulet, C. et Brennan, A. (2024). An estimator for concurrent use of full and reduced collection effort on random subsamples. Article présenté au Comité consultatif sur les méthodes statistiques, 78, Statistique Canada.

PROJET : Panels Web probabilistes à Statistique Canada

En 2020, Statistique Canada a commencé à utiliser des panels Web probabilistes comme autre méthode de collecte de statistiques officielles. Dans le cadre d'un panel Web, les répondants à de grandes enquêtes sociales probabilistes sont invités à fournir leurs coordonnées pour participer à de futures petites enquêtes (« enquêtes par recrutement »). Les panels Web peuvent servir à recueillir rapidement des renseignements et sont donc un outil utile pour fournir des données en temps opportun. Toutefois, les faibles taux de réponse cumulés associés à cette méthode limitent la précision des estimations des panels Web et soulèvent des préoccupations au sujet d'un biais potentiel. Ce projet a permis d'examiner cinq ans d'expérience de panels à Statistique Canada pour obtenir des renseignements sur le recrutement, la collecte et les estimations des panels.

Progrès :

La première composante du projet consistait à examiner des façons de recruter efficacement des participants aux panels et, par la suite, de recueillir des données à l'aide d'enquêtes par panel. La façon dont les questions de recrutement sont présentées peut entraîner des taux de participation très différents. Un test des présentations différentes des questions a montré que la présence d'une question de consentement explicite avait une grande incidence; 26 % des répondants fournissaient leurs coordonnées (courriel ou numéro de téléphone ou les deux) lorsqu'on leur demandait explicitement, sur

un premier écran, de consentir à participer à de futurs panels avant d'avoir l'occasion de fournir leurs coordonnées, comparativement à 88 % des répondants qui fournissaient leurs coordonnées sans la question de consentement. De plus, la mine de renseignements auxiliaires disponibles dans le cadre de l'enquête par recrutement peut être utilisée pour gérer activement les opérations de collecte par panel, en prédisant la probabilité de réponse et en utilisant cette information pour cibler les efforts de suivi. Lors de son utilisation, cette approche a permis de réduire la fourchette de taux de réponse entre les groupes ayant initialement des taux de réponse plus faibles (par exemple les groupes d'âge plus jeunes ou ceux ayant des niveaux de scolarité plus faibles) et les groupes ayant des taux de réponse plus élevés. Pour obtenir de plus amples renseignements sur ces travaux, consultez MacIsaac, Boulet et Thomas (2024).

Les renseignements auxiliaires provenant des enquêtes par recrutement peuvent également servir à élaborer de riches modèles qui peuvent être utilisés dans le processus de pondération pour les panels. La deuxième composante de ce projet portait sur l'efficacité de cette approche pour réduire le biais potentiel associé aux faibles taux de réponse obtenus par panel. Les estimations des variables de l'enquête par recrutement ont été calculées à l'aide de poids d'enquête par recrutement et de poids de panel Web et ont été comparées; les différences indiquent la possibilité d'un biais résiduel qui n'a pas été corrigé par le processus de pondération du panel Web. Cette enquête a révélé plus de différences significatives que ce à quoi on s'attendrait si l'estimateur par panel Web corrigeait entièrement le biais résultant du processus de réponse par panel Web. Il s'est avéré que les questions liées à certains sujets, comme la politique et le vote, le sentiment d'appartenance et la consommation de médias, présentaient les différences les plus significatives entre les estimations des panels Web et les estimations des enquêtes par recrutement. Pour obtenir de plus amples renseignements sur ces travaux, consultez Mather et Boulet (2024) et Mather, Boulet et Ra (2024).

Les résultats de ce projet ont été présentés au Symposium international de 2024 sur les questions de méthodologie dont le thème est « L'avenir des statistiques officielles ».

Pour obtenir plus de renseignements, veuillez communiquer avec :
Cilanne Boulet (cilanne.boulet@statcan.gc.ca).

Bibliographie

MacIsaac, K., Boulet, C. et Thomas, M. (2024). Recrutement et collecte de panels Web à Statistique Canada. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Mather, A. et Boulet, C. (2024). *Analysis of Residual Bias on Web Panels*. Document de travail de la Direction de la méthodologie (SSMD-2024-002E), Statistique Canada.

Mather, A., Boulet, C. et Ra, Y. (2024). Évaluation des biais liés aux panels Web probabilistes de Statistique Canada. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

PROJET : Évolution de la quantité et de la qualité des données recueillies pour l'Enquête sur les dépenses des ménages au fil des ans

Les taux de réponse aux enquêtes sociales ont considérablement diminué au cours des 25 dernières années, tendance encore accélérée par la pandémie de COVID-19 et l'abandon des interviews en

personne. Alors que cette baisse est bien documentée, les initiatives récentes sont principalement axées sur la compréhension et la gestion des taux de réponse. Toutefois, l'incidence du désengagement croissant du public à l'égard des programmes statistiques peut s'étendre au-delà d'une simple participation plus faible; cela pourrait également avoir une incidence sur la quantité et la qualité des renseignements fournis par ceux qui répondent toujours et compromettre la validité et l'exactitude des résultats publiés, si les mesures de sécurité méthodologiques ne sont pas suffisamment robustes.

Contrairement à la non-réponse globale, cet aspect du désengagement est plus difficile à mesurer, ce qui nécessite une enquête plus approfondie. La présente recherche vise à évaluer les variations de quantité et de qualité des données fournies dans les enquêtes-ménages, en mettant l'accent sur l'Enquête sur les dépenses des ménages (EDM). En comparant les indicateurs clés au fil du temps, la présente étude vise à établir un cadre que d'autres enquêtes auprès des ménages peuvent adopter, ce qui permettrait en fin de compte d'obtenir des renseignements plus généraux sur les défis en évolution de la collecte et du traitement des données d'enquête.

Progrès :

La quantité et la qualité des renseignements fournis par les répondants ont été évaluées à l'aide de divers indicateurs provenant de multiples cycles de l'Enquête sur les dépenses des ménages (2015 à 2023). Dans la mesure du possible, les indicateurs ont été ventilés par mode de collecte, afin d'évaluer l'incidence des changements dans la stratégie de collecte.

Cette étude n'a révélé aucune tendance claire à la hausse ou à la baisse de la quantité et de la qualité des données au cours des cinq derniers cycles de l'EDM. Bien que le désengagement croissant des répondants ait joué un rôle dans la non-réponse aux enquêtes, cela n'a manifestement pas eu d'incidence sur la non-réponse partielle ou l'exhaustivité et l'exactitude des réponses. Dans l'ensemble, les indicateurs indiquent que les répondants ont maintenu des tendances constantes en matière de réponses au fil du temps. Toutefois, ces constatations peuvent s'appliquer plus particulièrement à l'EDM; d'autres enquêtes sociales devraient mener leurs propres analyses, en utilisant cette étude comme modèle, afin de déterminer si le désengagement des répondants a eu une incidence sur la qualité de leurs propres données.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Yves Lafortune (yves.lafortune@statcan.gc.ca).

Bibliographie

Lafortune, Y. et Mayer, E. (2025). The Evolution of the Quantity and Quality of the Data Collected for the Survey of Household Spending Through the Years. Document de recherche, Statistique Canada.

4 Confidentialité et accès aux données

Les études de Statistique Canada portant sur la confidentialité ont continué de mettre l'accent sur l'élaboration de nouvelles méthodes et idées offrant d'autres formes d'accès, tout en continuant de veiller à ce que les renseignements personnels et commerciaux ne soient 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 comme moyen de contribuer au renforcement des capacités en matière de détermination et de traitement des risques de divulgation (voir la section 5.5).

PROJET : Évaluation de la confidentialité des estimations sur petits domaines

Comme il est indiqué dans le rapport précédent, Statistique Canada ne dispose pas de directives officielles sur les règles de confidentialité pour la diffusion d'estimations sur petits domaines et aucune étude officielle n'a encore été menée à ce sujet.

Progrès :

Une étude (Tang, 2024) décrivant le processus de simulation et traitant des justifications des règles de confidentialité proposées a été présentée au Symposium de 2024 sur les questions de méthodologie et sera publiée dans le cadre des actes du Symposium.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Cissy Tang (cissy.tang@statcan.gc.ca).

Bibliographie

Tang, C. (2024). Statistical disclosure control analysis for small area estimation. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

PROJET : Données synthétiques

Il est essentiel de travailler à accroître les options à la disposition des utilisateurs de données. La création de données synthétiques est une façon 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 :

Compte tenu du succès de la diffusion de la Base de données synthétiques pour le modèle de microsimulation dynamique PASSAGES, une version provisoire d'une mise à jour des lignes directrices internes pour la création de fichiers de données synthétiques a été préparée (Gauvin, 2025). Cette version remplacera la documentation existante utilisée pour guider les développeurs lors de la création de données synthétiques.

Gauvin (2024) a présenté son travail d'élaboration des données synthétiques pour le modèle PASSAGES à l'assemblée annuelle de 2024 de la Société statistique du Canada.

Yu (2024) a présenté un examen de l'évaluation du risque de divulgation de données synthétiques dans le cadre du Symposium international de 2024 sur les questions de méthodologie.

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. (2025). Practical Guidelines for the Creation of Synthetic Data Files. Document interne, Statistique Canada.

Gauvin, H. (2024). Creating a synthetic version of a longitudinal and structured file: Challenges and lessons. *Recueil de la Section des méthodes d'enquête, Société Statistique du Canada*, St. Johns, Terre-Neuve, https://ssc.ca/sites/default/files/imce/gauvin_ssc2024.pdf.

Yu, Z. (2024). Évaluation des risques liés à la divulgation de données synthétiques. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

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

La suppression de cellules complémentaires (SCC) est une méthode typique pour supprimer de façon confidentielle les cellules sensibles lors de la diffusion de données quantitatives tabulaires. Cette méthodologie est bien élaborée et appuyée par la solution G-Confid de Statistique Canada, dans le cadre de laquelle des solutions de suppression optimales sont obtenues pour veiller à ce que les modèles de suppression soient valides et réduire au minimum la quantité de renseignements supprimés.

Progrès :

Statistique Canada a construit avec succès une solution bêta de G-Confid fondée sur Python. Cela comprend une solution d'arrondissement additive optimale, un outil de calcul des mesures de sensibilité, la création d'une tendance optimale de SCC et un programme de vérification pour s'assurer que les tendances sont valides. Les solutions fondées sur le code source ouvert disponibles dans le cadre de la trousse PuLP Python ont été étudiées en tant que remplacements potentiels et appliquées avec prudence aux cas les plus complexes à Statistique Canada.

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

PROJET : Cadre fondé sur le risque de divulgation et l'utilité des données pour comparer les mécanismes de contrôle de la divulgation statistique

En tant qu'organisme national de statistique (ONS), Statistique Canada a été chargé de trouver des façons de diffuser les données à des niveaux plus détaillés. Offrir des analyses statistiques sur de petites sous-populations à des niveaux géographiques plus désagrégés suscite un intérêt particulier. Il s'agit de l'initiative du Plan d'action sur les données désagrégées (PADD). Toutefois, la diffusion des données doit toujours se faire en respectant les dispositions en matière de confidentialité de la *Loi sur la statistique*. Une composante clé de la diffusion de produits statistiques conformes aux règles sur la confidentialité consiste à appliquer des mesures de contrôle de la divulgation statistique (CDS) aux produits statistiques avant leur diffusion.

Progrès :

Pour mieux répondre aux besoins des utilisateurs, des recherches sont en cours pour analyser les diverses méthodes de CDS, dans le but de mesurer l'utilité et le risque de divulgation qui leur sont associés. Un cadre de compromis entre l'utilité et le risque de divulgation est proposé pour comparer les méthodes de CDS présentées. En appliquant ce cadre, un responsable de la diffusion de données peut choisir une méthode de CDS qui équilibre de façon optimale le risque de divulgation et l'utilité dans une situation

donnée. Un document de travail interne est en cours d'élaboration et sa publication externe sera envisagée.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Joshua Miller (joshua.miller@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

La confidentialité différentielle est un cadre de contrôle de la divulgation statistique (CDS) visant à limiter le niveau de fuite de renseignements privés encourue en fonction de la contribution d'une personne à une base de données privée après la diffusion des résultats statistiques. Ce cadre a d'abord été étudié dans Dwork, McSherry, Nissim et Smith (2006). Un cadre de CDS respectant les exigences de confidentialité différentielle a été proposé par des universitaires et des organismes gouvernementaux. Plus précisément, des méthodes se prêtant au cadre ont été utilisées comme solutions remplaçant des méthodes traditionnelles de CDS, comme l'arrondissement et la suppression de cellules. Par exemple, le Bureau du recensement des États-Unis a mis en œuvre des algorithmes de confidentialité différentielle comme méthode de publication des résultats de son Recensement de 2020, comme l'illustrent Abowd, Ashmead, Cumings-Menon, Garfinkel, Heineck, Heiss, Johns, Kifer, Leclerc, Machanavajjhala, Moran, Sexton, Spence et Zhuravlev (2022). Il est donc de la plus haute importance que Statistique Canada comprenne pleinement le cadre de confidentialité différentielle et investisse dans des domaines où cela pourrait se traduire par de grandes améliorations de son écosystème de diffusion de données.

Même si l'on sait que la confidentialité différentielle offre de solides garanties en matière de protection des renseignements personnels, elle peut faire l'objet de critiques. L'une des critiques courantes de la confidentialité différentielle est l'impact des méthodes sur l'utilité des données publiées. De plus, il peut être difficile d'établir un compromis approprié entre l'utilité et le risque de divulgation. Dans le contexte de la confidentialité différentielle, cela se traduit par la détermination d'un budget approprié pour la protection des renseignements personnels. De plus, si elle n'est pas correctement mise en œuvre, la confidentialité différentielle peut ne pas protéger les cellules sensibles. Par conséquent, une évaluation de l'utilité des données respectant la confidentialité différentielle ainsi qu'une étude plus approfondie des défis de la confidentialité différentielle sont justifiées. Cela comprend également de comparer des méthodes respectant les notions de confidentialité différentielle avec les méthodes traditionnelles de CDS.

Progrès :

Un article détaillé résumant les défis pratiques auxquels fait face l'adoption d'un cadre de confidentialité différentielle (Miller, 2025) a été rédigé. Cet article comprend un résumé de la confidentialité différentielle et une analyse de ses propriétés de base. Une comparaison des différentes définitions de la confidentialité différentielle est également incluse. De plus, une explication intuitive des garanties qu'un cadre de confidentialité différentielle fournit est présentée. Les méthodes traditionnelles de CDS, comme l'arrondissement aléatoire, sont également examinées dans l'optique de la confidentialité différentielle. L'un des principaux obstacles pratiques auxquels font face les responsables de la diffusion lorsqu'ils envisagent l'adoption de la confidentialité différentielle est le choix d'un paramètre de confidentialité approprié. Cette question est traitée en détail dans le document et des suggestions sont proposées.

Un algorithme de CDS respectant les principes de confidentialité différentielle a été mis à l'essai sur les données du Recensement canadien de 2021 pour démontrer à quoi pourrait ressembler l'adoption de la confidentialité différentielle en pratique. L'algorithme mis en œuvre était fondé sur l'algorithme TopDown du Bureau du recensement des États-Unis. L'algorithme TopDown a d'abord été appliqué aux données du recensement des États-Unis de 2020, comme il en est question dans Abowd et coll. (2022). L'une des principales constatations de l'étude de cas du recensement canadien était que l'utilité des produits statistiques pouvait être considérablement compromise si l'objectif du responsable de la diffusion était d'offrir des garanties importantes en matière de protection de la vie privée.

Le document de Miller sur la confidentialité différentielle (2025) fait l'objet d'un examen par les pairs et sera officialisé au cours de l'exercice 2025-2026 en tant que document de travail de la direction. Les résultats de ces recherches ont également été communiqués sous forme de séminaire sur la recherche méthodologique.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Joshua Miller (joshua.miller@statcan.gc.ca).

Bibliographie

Abowd, J.M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M. et Zhuravlev, P. (2022). The 2020 Census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review*, (Numéro spécial 2).

Dwork, C., McSherry, F., Nissim, K. et Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3), 17-51.

Miller, J. (2025). A Privacy-Utility Based Study of Differential Privacy. Rapport interne, Statistique Canada.

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 a pour objectif de maintenir une expertise de haut niveau et d'offrir des consultations sur les séries chronologiques à l'ensemble de l'organisme. Le centre fournit des consultations et des conseils sur les problèmes liés aux séries chronologiques, étudie les problèmes pour lesquels aucune solution n'est actuellement connue ou satisfaisante et élabore et tient à jour des outils permettant d'appliquer des solutions à des problèmes réels de séries chronologiques.

Ces projets peuvent être répartis en quatre sous-sujets mettant l'accent sur les aspects suivants :

- Consultation et formation en séries chronologiques;
- Soutien et amélioration du système et des outils de traitement de séries chronologiques;
- Modélisation et prévision 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 sur les séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques est responsable de l'élaboration et de la prestation de formations sur les méthodes de séries chronologiques, y compris la désaisonnalisation, l'étalonnage, le rapprochement et la modélisation de séries chronologiques, aux participants de Statistique Canada ainsi qu'à ceux d'autres organismes. De plus, il fournit des conseils et des consultations sur les projets de séries chronologiques en général pour les programmes de l'ensemble de Statistique Canada.

Le Centre a offert des cours sur les composantes de séries chronologiques, la désaisonnalisation, l'étalonnage et le rapprochement au cours de l'année aux participants internes et externes par l'entremise du centre de formation de Statistique Canada (Statistique Canada, 2024). En particulier, le cours d'introduction sur les composantes de séries chronologiques et la désaisonnalisation a été présenté aux membres du Bureau de la statistique du Laos. Le Centre a également participé à des activités de sensibilisation et à de la formation ponctuelle auprès d'autres groupes de Statistique Canada sur des sujets liés aux séries chronologiques (série de séminaires de la Direction de la méthodologie pour les recrues, cours sur les navigateurs de données et démonstrations du tableau de bord de la désaisonnalisation).

Le Centre a en outre fourni des consultations dans le cadre de divers programmes internes (désaisonnalisation, modélisation de séries chronologiques, rétropolation, prévision immédiate, prévision, estimation de tendances, calendarisation, etc.) En particulier, le Centre a fourni un soutien en matière de séries chronologiques au Système de comptabilité nationale dans divers domaines, y compris les programmes sur le produit intérieur brut trimestriel et mensuel, la balance des paiements, le commerce international de services et les opérations sur titres. 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. Les travaux relatifs à la production de lignes directrices sur la continuité des séries chronologiques se sont poursuivis et ont été intégrés à un projet plus vaste de mise à jour et d'expansion de lignes directrices sur les méthodes de séries chronologiques pour les programmes de Statistique Canada. De récents travaux comprenaient des discussions sur les lignes directrices proposées avec le Bureau de gestion de projets de l'organisme, afin de déterminer la façon d'intégrer ces lignes directrices au cadre de gestion de projet.

De plus, pour appuyer divers programmes internes, le Centre a consulté et échangé à l'externe sur des sujets liés aux séries chronologiques (stratégie de désaisonnalisation pendant la pandémie, rétropolation, déflation, outils logiciels, etc.) avec de multiples organismes publics fédéraux et provinciaux ainsi que des organismes nationaux de statistique.

Soutien et amélioration du système et des outils de traitement des 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-Séries, aux fins d'étalonnage et de réconciliation (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).

Le travail restant sur une nouvelle version libre de G-Séries en R est terminé, avec l'ajout de la fonctionnalité d'équilibrage, des fonctions d'utilité et de la documentation bilingue élargie. La nouvelle version de G-Séries est élaborée et tenue à jour dans l'environnement GitLab interne de Statistique Canada en respectant des pratiques exemplaires d'intégration et de prestation continues (CI/CD). Elle a fait l'objet d'essais alpha et bêta, ainsi que d'un examen du code et d'une évaluation de la vulnérabilité des applications de cybersécurité. La nouvelle version de G-Séries sera diffusée officiellement en tant que progiciel R sur GitHub et CRAN en mai 2025.

Le système de traitement des séries chronologiques est une application personnalisable reposant sur SAS et permettant d'appliquer des techniques de séries chronologiques, dont des techniques de désaisonnalisation, d'étalonnage et de réconciliation, largement utilisées au sein de Statistique Canada dans la production d'estimations désaisonnalisées pour des programmes infra-annuels (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-Séries, le système de traitement des séries chronologiques sera remanié en R. Le travail devrait commencer l'an prochain et offrira la souplesse nécessaire pour intégrer les outils et les nouvelles techniques disponibles à partir de logiciels libres.

Le tableau de bord de la désaisonnalisation est maintenant hébergé sur le GitLab interne de Statistique Canada et reconfiguré pour faciliter son installation. Parmi les autres améliorations mineures apportées au tableau de bord, mentionnons la réduction des dépendances et du temps de chargement. Le tableau de bord a été déployé cette année pour deux autres programmes sur la main-d'œuvre, dans le cadre desquels une révision mineure a été apportée pour intégrer un effet de calendrier propre au programme. Le Centre a également offert de la formation à des analystes spécialisés.

Modélisation et prévision de séries chronologiques

Le Centre a terminé un projet sur la production d'indications précoces de ruptures structurelles à l'aide de modèles d'espace d'états. Un outil R a été créé pour détecter et modéliser les chocs (changements soudains) dans les séries chronologiques, soit dans des équations de mesure (valeurs aberrantes additives), soit dans des équations d'état (effets plus permanents).

Dans l'Enquête sur la population active, certaines estimations sont calées en fonction des chiffres de population. La proportion de résidents temporaires au sein de la population varie selon les saisons (en partie en raison des étudiants étrangers et des travailleurs temporaires). Pour stabiliser le calage, une méthodologie simple est actuellement utilisée pour lisser ces fluctuations (moyenne mobile sur 12 mois). Le Centre a lancé une étude visant à déterminer si les techniques de désaisonnalisation et d'estimation de tendance-cycle pourraient fournir une méthodologie améliorée, particulièrement en période de changements rapides dans ce segment de la population. L'étude est en cours.

La période de 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 pour 2023 et 2024, on peut commencer à évaluer l'impact de la pandémie sur les tendances saisonnières, la tendance-cycle et la volatilité. Le Centre a entamé un examen préliminaire à l'aide de séries chronologiques représentatives de divers programmes statistiques (main-d'œuvre, commerce,

fabrication, permis de construction et investissement en construction de bâtiments, tourisme, Indice des prix à la consommation), afin de déterminer si ces séries sont revenues à des tendances prépandémie ou sont passées à un nouveau régime postpandémie. Cette enquête continuera d'être affinée à mesure que d'autres données postpandémie deviendront disponibles.

Le Centre a également contribué à un projet de prévision immédiate dans le cadre de l'initiative d'accélération méthodologique de Statistique Canada, initiative lancée pour répondre à la nécessité d'améliorer considérablement la capacité de l'organisme à mettre en œuvre efficacement et rapidement des solutions modernes aux défis liés aux données. Le Centre a élaboré, mis à l'essai et validé un scénario de prévision immédiate de macro-estimations pour l'Enquête sur les marchandises vendues au détail, à l'aide des modèles ARIMA et de réseaux neuronaux. De plus, un scénario de micromodélisation, au niveau de l'entreprise, a été élaboré, mis à l'essai et validé.

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 (IPP).

Cette sous-section a élaboré et mis à l'essai un protocole d'échantillonnage pour le contrôle de la qualité des classificateurs d'apprentissage automatique utilisés pour les problèmes de classification à grande échelle à multiples catégories et la collecte longitudinale. Au cours des dernières années, l'IPC a intégré des données de lecteurs optiques aux points de vente, ce qui permet de fournir chaque mois un recensement des opérations de quelques grands détaillants. Les descriptions textuelles des produits doivent être classées dans la structure d'agrégation hiérarchique de l'IPC de plus de 700 catégories de produits. Un classificateur de machines à vecteurs de soutien linéaires a été entraîné pour effectuer cette classification. Toutefois, il fait trop d'erreurs pour être acceptable pour l'IPC, nécessitant un examen supplémentaire par des annotateurs humains. Le coût de l'annotation humaine est l'étape limitant l'expansion de l'utilisation des données de lecteurs optiques. Trois stratégies d'échantillonnage ont été élaborées pour sélectionner les cas devant faire l'objet d'une annotation humaine et fournir des estimations des taux d'erreurs de classification erronée avec des intervalles de confiance de Wilson modifiés. Le Centre a collaboré avec des scientifiques des données de la Division des prix à la consommation pour mettre en œuvre ces stratégies d'échantillonnage en Python et les mettre à l'essai sur des cas simulés de ventes d'aliments, dans le but de rendre le code public pour d'autres organismes nationaux de statistique et d'autres chercheurs. Les résultats de la simulation ont privilégié une solution hybride entre deux stratégies : 1) un échantillon aléatoire simple stratifié pour de nouveaux produits chaque mois, afin d'estimer les taux de classification erronée et de surveiller le rendement du modèle; 2) un échantillon par probabilité proportionnelle à la taille parmi les unités non examinées des mois précédents, ciblant les cas les plus susceptibles de causer un biais de classification erronée dans l'estimation de l'IPC, selon une mesure d'influence dérivée des indices de prix (Spackman, Francis et Goussev, 2025). Les résultats de la stratégie et de la simulation ont été présentés lors d'une réunion d'avril 2025 de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) à Genève, en Suisse, sur les méthodes d'indice des prix à la consommation.

Le Centre a également mené des consultations sur l'élaboration d'un ensemble d'indicateurs de la qualité pour l'IPC. La première étape de ce projet a donné lieu à un ensemble d'indicateurs quantitatifs et à une mesure agrégée pour des problèmes potentiels de collecte et de traitement. De plus, un ensemble

préliminaire d'indicateurs qualitatifs portant sur les six dimensions de la qualité de Statistique Canada a été élaboré (Francis, Carrillo-Garcia, Yélou et Rassart, 2025).

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

Bibliographie

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. R package version 3.0.2, https://StatCan.github.io/gensol-gseries/fr/.

Francis, J., Carrillo-Garcia, I., Yélou, C. et Rassart, E. (2025). Developing CPI Quality Indicators: Project Update. Document interne, Statistique Canada.

Spackman, W., Francis, J. et Goussev, S. (2025). Optimal use of machine learning at scale: Designing quality control of machine learning classification and mitigating misclassification error. *Proceedings of the 17th Meeting of the Group of Experts on Consumer Price Indices*, CEE-ONU, Genève, Suisse. Disponible à l'adresse : https://unece.org/sites/default/files/2025-04/Spackman%20et%20al%20%282025%29%20-%20QC%20for%20mitigating%20misclassification%20bias%20-%20paper.pdf.

Statistique Canada (2024). Formation. Disponible à : https://www.statcan.gc.ca/fr/afc/formation.

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

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 pour les statistiques économiques a été réorganisée en tant que section appelée le Centre de ressources sur l'innovation et les outils de statistiques économiques. Ce nouveau centre est responsable, entre autres, du soutien et du développement de trois systèmes généralisés : G-Sam (système généralisé d'échantillonnage), Banff (système généralisé de vérification des données statistiques) et G-Est (système généralisé d'estimation).

Progrès :

Les projets du Centre peuvent généralement être classés en soutien (pour les utilisateurs du système), recherche et développement.

Le Centre a traité un volume de cas de soutien typique pour G-Sam, Banff et G-Est. Ceux-ci comprenaient des cas de soutien pour les deux versions actuelles des systèmes fondées sur SAS et du soutien pour la migration vers la version Python de Banff. Même si certains cas ont été résolus en quelques échanges de courriels, plusieurs 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. Un résumé de ces cas est fourni ci-dessous.

Équipe du système Banff :

- Poursuite de la collaboration avec l'Enquête sur les postes vacants et les salaires pour mettre en œuvre des changements à son étape de vérification et d'imputation. De plus, l'équipe a fourni une version à code source ouvert du processus, appelant la version Python de Banff, en excluant les étapes personnalisées de SAS.
- Collaboration avec la section des impôts sur une stratégie d'imputation pour le projet de remaniement du système de traitement des données du fichier T1, à l'aide de la version Python de Banff.
- Consultation auprès de plusieurs équipes d'enquête, tant du côté économique que des ménages, pour étudier l'utilisation de la nouvelle version Python. Cela a inclus des démonstrations du processeur Banff et des recommandations sur des méthodes à code source ouvert qui pourraient être intégrées au système Banff.

Équipe de G-Sam :

- Soutien à l'Enquête sur les ventes d'aliments et de boissons prêts à servir, à l'Enquête sur les infrastructures publiques essentielles du Canada (IPEC) et à l'Enquête canadienne sur le commerce interprovincial en matière d'optimisation de la répartition selon diverses contraintes (taille d'échantillon, précision et coordination), incluant des commentaires supplémentaires sur la rotation et la restratification.
- Présentation de la fonctionnalité de répartition de G-Sam à l'équipe de l'IPEC, en soulignant ses caractéristiques et ses applications potentielles.
- Présentation d'une explication détaillée de l'algorithme de stratification de Lavallée-Hidiroglou au ministère de la Statistique de Singapour, y compris son fondement théorique et son application pratique dans le plan de sondage.
- Résolution de problèmes numériques liés à la répartition pour l'Enquête sur les fruits et légumes, afin d'assurer la mise en œuvre efficace de stratégies d'échantillonnage.

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 (SOS) 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 du système. Cela a inclus des présentations au Comité directeur sur les systèmes généralisés, des séminaires de la direction, des réunions du Conseil de planification du secteur et un article externe (Gray et Pierre, 2025). L'équipe du Centre a également rencontré des représentants de l'Office fédéral allemand de la statistique (Destatis) pour discuter de la transition de Banff vers les sources ouvertes et pour amorcer un projet de collaboration sur l'élaboration d'un cadre et d'un outil d'évaluation des méthodes d'imputation.

Le projet de modernisation de Banff a été mené à bien en respectant la portée, le budget et le calendrier définis. Banff est un système modulaire de vérification des données statistiques (VDS) permettant de détecter et de traiter les erreurs de déclaration et les cas de non-réponse. En janvier 2025, la version fondée sur Python a été diffusée, ce qui a permis d'améliorer la flexibilité, d'appuyer des méthodes d'imputation avancées et de faciliter la collaboration internationale. Banff comprend neuf procédures remplissant diverses fonctions de VDS, notamment la détection de valeurs aberrantes, la localisation des erreurs, l'imputation et l'ajustement proportionnel, ainsi que des méthodes de vérification et d'imputation des données sous contraintes de relations linéaires. Le processeur Banff est un outil axé sur

les métadonnées qui exécute la vérification des données en production, en appelant des procédures Banff intégrées parallèlement à des modules personnalisés dans le cadre d'un enchaînement de processus précisé par l'utilisateur. Les utilisateurs sont invités à partager des modules personnalisés compatibles avec Banff dans le répertoire des modules d'extension Banff, ouvert à tous les utilisateurs, afin de favoriser la collaboration et de réduire la répétition inutile. Cette version en Python harmonise le système avec le Modèle générique de vérification des données statistiques (CEE-ONU, 2019). Une présentation de premier plan portant sur le nouveau système a eu lieu à la réunion d'experts sur la vérification des données statistiques de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) (Gray, 2024). Pour promouvoir le système, l'équipe du système Banff a présenté un séminaire de lancement (Gray et Seffal, 2025a) et a communiqué les leçons apprises au cours de la semaine d'apprentissage de la Direction des méthodes statistiques modernes et de la science des données (Gray et Seffal, 2025b).

La version en R de G-Sam a été entièrement reconstruite, en conservant les méthodes de base de la version SAS tout en améliorant considérablement le rendement et en élargissant les fonctionnalités. Les nouvelles fonctions de répartition et de sélection gèrent un éventail plus large de problèmes et surpassent leurs équivalents en SAS avec moins de code et moins de données d'entrée. Les dépendances sont minimales et les structures d'entrées ont été remaniées pour améliorer la convivialité. Pour faciliter la transition, l'équipe élabore actuellement des outils et des conseils en matière de migration. Les données de sortie demeurent en grande partie cohérentes; elles présentent seulement des différences mineures en raison du changement dans les moteurs d'optimisation. Une version *bêta* devrait être diffusée en juin 2025, la diffusion de production étant prévue pour septembre 2025.

La version en R de G-Est offrira des fonctions de pondération et d'estimation de la variance pour des plans d'enquête complexes et comprendra la plupart des mêmes caractéristiques que la version SAS. Elle inclura la méthode du bootstrap, le calage, les corrections pour la non-réponse, l'estimation sur petits domaines et la variance due à l'imputation (par l'entremise du module du Système d'estimation de la variance due à la non-réponse et à l'imputation, SEVANI). Comme pour G-Sam, la mise en œuvre a été entièrement reconçue, optimisée pour R avec des dépendances externes minimales et la possibilité d'ajouter des méthodes à l'avenir. Les versions alpha du calage et de l'estimation de la variance en une étape sont terminées; les essais initiaux montrent des améliorations de rendement importantes par rapport à la version SAS. Pour l'estimation sur petits domaines, l'équipe de transition examine la pertinence d'un progiciel existant, tandis que l'élaboration des modules bootstrap et SEVANI devrait commencer ce printemps. Une version *bêta* devrait être diffusée en décembre 2025, la diffusion de production étant prévue pour décembre 2026.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Fritz Pierre (fritz.pierre@statcan.gc.ca).

Bibliographie

Commission Économique des Nations Unies pour l'Europe (CEE-ONU) (2019). *Generic Statistical Data Editing Model* (GSDEM), https://statswiki.unece.org/display/sde/GSDEM.

Gray, D. (2024). Building the new Banff: An open-source data editing system based on GSDEM concepts. UNECE Expert Meeting on Statistical Data Editing, 7 au 9 octobre 2024, Vienne.

Gray, D. et Pierre, F. (2025). De SAS vers les sources libres : conversion des systèmes généralisés de Statistique Canada. À paraître dans le prochain numéro de *Convergence*, une revue de l'Association des statisticiennes et statisticiens du Québec.

Gray, D. et Seffal, M. (2025a). The New Banff: A Modern Edit and Imputation Platform for Everyone. Présentation interne, Statistique Canada.

Gray, D. et Seffal, M. (2025b). From SAS to Open-Source: What we learned from the Banff project. Présentation interne, Modern Statistical Methods and Data Science Branch Learning Week, Statistique Canada.

5.3 Centre de ressources en couplage d'enregistrements

Les objectifs du Centre de ressources sur le 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. Nous facilitons également la diffusion de renseignements sur les méthodes, les logiciels et les politiques de couplage d'enregistrements, ainsi que sur l'analyse de données couplées aux parties intéressées à l'intérieur et à l'extérieur de Statistique Canada.

Progrès :

Nous avons offert du soutien à l'équipe de développement de G-coup, le système de couplage d'enregistrements mis au point à Statistique Canada. Le CRCE a également offert un soutien aux utilisateurs internes et externes de G-coup qui ont demandé de l'aide, fourni des commentaires ou soumis des suggestions au moyen de requêtes à G-coup_info.

Au cours de l'année, l'essentiel des travaux méthodologiques a porté sur la maintenance, le développement et l'accompagnement des utilisateurs de la version 3.5 de G-coup sur les serveurs SAS en environnement infonuagique. Un volume typique de cas d'assistance pour G-coup a été traité par l'équipe du projet. La plupart d'entre eux ont été résolus par des suggestions sur la manière d'appliquer le système en termes pratiques, mais plusieurs ont nécessité un engagement plus important.

Le travail de développement a été axé sur une mise à jour des outils pour la création de données synthétiques pour le couplage d'enregistrements à des fins de test et de formation. Le travail a été réalisé en R et en Python dans le cadre du passage à des outils à code source ouvert au sein de l'organisme.

Le CRCE a également travaillé sur divers autres couplages probabilistes dans l'Environnement de couplage des données sociales ECDS. Ces couplages nous ont aidés à analyser les performances du logiciel et les solutions à fournir. Le travail sur ces projets a abouti à des approches plus systématiques pour définir et ajuster les couplages d'enregistrements sur les serveurs SAS basés sur l'infonuagique. Des travaux ont également été entrepris pour réajuster les poids de sondage afin de compenser les biais introduits par les liens manqués, y compris des travaux pour des clients externes sur des données de recensement liées à plusieurs fichiers administratifs longitudinaux.

Les membres de l'équipe ont offert des cours formels avec le Centre de formation de Statistique Canada et ont préparé du matériel pour un atelier dans le cadre de la conférence du Réseau canadien des centres de données de recherche en 2025. Le CRCE a aussi mis à jour de la documentation de référence et de stratégie pour les utilisateurs des techniques de couplages probabilistes.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Abdelnasser Saïdi (abdelnasser.saidi@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 données d'enquête, les données de recensement et les données administratives, sont offerts aux employés de Statistique Canada et d'autres ministères, ainsi qu'aux analystes et chercheurs du milieu universitaire et des centres de données de recherche.

Progrès :

Consultations

Des services de consultation ont été fournis à la demande de clients internes et externes. Entre le 1er avril 2024 et le 31 mars 2025, le CRAD a répondu à environ 50 demandes. Les questions ont varié en complexité et ont porté sur des sujets comme la régression logistique au moyen des données d'enquête, l'estimation de la variance au moyen des poids bootstrap, les intervalles de confiance pour de petites proportions, la comparaison de sous-populations dépendantes, les tests du khi carré, le calcul des degrés de liberté pour des données d'enquête et la régression par quantile. Le CRAD a également aidé des clients à mettre en œuvre des méthodes statistiques avec les logiciels R, SAS, SUDAAN et STATA. De plus, le CRAD a examiné des documents analytiques pour des revues scientifiques, des conférences et des publications internes.

Services de formation

Le CRAD a présenté, en anglais, le cours interne 0438A « Analyse statistique des données d'enquête – Module 1 ». Ce cours de six jours allie théorie et pratique. Des exercices et des exemples ont été présentés en utilisant du code R, SUDAAN et SAS.

Le CRAD a fait une présentation sur l'analyse de données d'enquête complexe, en français et en anglais, lors de l'Atelier d'interprétation de données organisé par Statistique Canada. Le CRAD a présenté de nouveau, en anglais, les séances sur la régression linéaire au moyen de données d'enquête complexes, dans le cadre du cours sur la modélisation statistique à Statistique Canada. Le CRAD a également organisé un séminaire pour les recrues sur l'analyse des données d'une enquête complexe.

En collaboration avec la Division de l'analyse de la santé (DAS), le CRAD a élaboré et présenté un cours de 12 heures intitulé « Survivre à la transition en R pour les analystes d'enquêtes ». Cette formation pratique a été proposée aux analystes d'enquête de la DAS et comprenait des exemples pratiques et des exercices

axés sur la préparation préalable des données, l'analyse des données à l'aide du progiciel « survey » de R et la visualisation des données.

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 2019, 2020 et 2022 du Sondage auprès des fonctionnaires fédéraux ont été utilisées pour mesurer des variables latentes, comme les facteurs de risque psychologiques 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). Les modèles de mesure ont été élaborés au moyen de l'analyse factorielle et de la modélisation par équations structurelles, comme l'ont expliqué Blais, Mach, Michaud et Simard (2020), et Blais, Michaud, Simard, Mach et Houle (2021). Cette année, le CRAD a estimé la variance des scores factoriels, permettant le calcul des coefficients de variation et des intervalles de confiance pour tous les domaines présentés dans le tableau de bord.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Pierre-Olivier Julien (pierre-olivier.julien@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. *Rapports sur la santé*, 32, 12.

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 de stratégies de prévention de la divulgation.

Anonymisation

Le groupe de soutien à la confidentialité a continué d'offrir son expertise dans la compréhension et le développement d'idées liées à la dépersonnalisation et à l'anonymisation. Statistique Canada continue d'améliorer ses propres stratégies internes pour s'assurer que les renseignements internes sont dépersonnalisés dans la mesure du possible afin de réduire le plus possible les risques de divulgation.

Statistique Canada a contribué à un ensemble de fiches de renseignements qui seront publiées par le Commissaire à l'information et à la protection de la vie privée de l'Ontario.

Le Centre de ressources pour la confidentialité et l'accès aux données continue d'offrir gouvernance, orientation et conseils stratégiques en matière de production d'ensembles de données ouvertes. Cette

année, plus de 15 nouveaux fichiers de microdonnées à grande diffusion ont été examinés et rendus accessibles sur le site Web de Statistique Canada.

Consultations externes

Statistique Canada a continué d'offrir son expertise à plusieurs groupes à l'étranger et au Canada. À l'échelle internationale, Statistique Canada a continué de participer à une étude pilote entre le Canada (StatCan), la France (Institut national du cancer) et les États-Unis (National Cancer Institute) visant à étudier la capacité d'échanger des données sur le cancer entre les pays.

À l'échelle nationale, nous continuons de travailler avec Santé Canada en fournissant des tableaux non divulgateurs des chiffres quotidiens de mortalité afin de contribuer à son projet de Cote air santé. Plutôt que des techniques d'arrondissement classiques, nous proposons des méthodes d'ajout de bruit visant à accroître l'utilité pour le même niveau de risque de divulgation.

Statistique Canada a offert son expertise en matière de stratégies de contrôle de la divulgation auprès de Service correctionnel Canada, à des analystes de l'Enquête nationale sur la santé des Inuits, à l'Agence de la santé publique du Canada pour sa diffusion d'information de la Base de données en ligne des effets indésirables de Canada Vigilance et à la Banque du Canada pour son Enquête sur les attentes des consommateurs au Canada.

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

5.6 Activités de soutien et de recherche en intelligence artificielle

Le Programme de recherche et développement en méthodologie (PRDM) de Statistique Canada a appuyé de nombreuses activités du Centre de recherche et d'excellence en intelligence artificielle et de la Division des méthodes d'intelligence artificielle. Le soutien du PRDM a permis à de nombreuses personnes d'effectuer des recherches approfondies, de créer des prototypes, d'établir des collectivités, de créer des centres d'expertise et d'élaborer des lignes directrices pouvant profiter à l'organisme.

Activités, mandats et produits

Le financement a permis au Centre d'expertise du traitement automatique de la langue naturelle (TALN) de progresser dans son mandat de centraliser les ressources pour le partage des connaissances et le renforcement des capacités en matière d'analyse de textes à l'aide de l'apprentissage automatique et de créer, de tenir à jour et de promouvoir des pratiques exemplaires et des lignes directrices en matière d'analyse de textes. Les activités du Centre consistent à procéder à des examens, à fournir des services de consultation et de l'orientation aux spécialistes du TALN au sein de Statistique Canada, ainsi qu'à créer la liste des projets de TALN réalisés et en cours au sein de l'organisme.

Une autre communauté financée par le PRDM traite des technologies d'amélioration de la protection de la vie privée (TPVP). Ce soutien a permis le lancement de la communauté de pratique, qui est toujours en cours. En cette ère de protection de la vie privée, le fonctionnaire ayant besoin de TPVP sera en mesure de chercher et de trouver des renseignements sur les techniques les plus récentes, de communiquer avec l'équipe de TPVP de Statistique Canada et, éventuellement, de trouver des conférences et des articles à mesure que le site évolue. Le soutien du PRDM a également permis le lancement de l'équipe de TPVP

responsable, qui élaborera des lignes directrices et créera un comité d'examen pour tout projet de TPVP interne ou externe (travaux en cours).

Outre le soutien apporté aux communautés et aux centres, ce financement a permis de poursuivre une étude approfondie comprenant la mise à profit de modèles probabilistes de diffusion de débruitage pour générer de grands volumes de données tabulaires synthétiques qui reproduisent fidèlement les répartitions de données du monde réel. Le projet intègre un processus de réduction du biais, inspiré de l'inférence alimentée par des prédictions et de l'estimateur par la différence généralisée. La solution proposée est maintenant évaluée à des fins de mise en œuvre, ce qui pourrait permettre de gagner du temps, de réduire les efforts et d'économiser de l'argent lors de la réalisation d'enquêtes par échantillonnage.

Le deuxième projet de recherche prometteur financé par le comité est une étude approfondie visant à créer des mesures d'évaluation pour l'intelligence artificielle (IA) générative, c'est-à-dire un cadre pour les grands modèles de langage. Contrairement aux problèmes classiques, comme la classification ou la régression, avec des mesures établies (comme la racine de l'erreur quadratique moyenne, REQM, ou la cote F1), l'IA générative manque de méthodes d'évaluation normalisées; celles qui sont utilisées en pratique varient en fonction de tâches particulières et présentent des problèmes de fiabilité. Un cadre d'évaluation a été créé pour améliorer la convivialité et la reproductibilité, tout en offrant des outils et des artefacts conviviaux qui n'exigent pas de connaissances théoriques approfondies pour les praticiens.

D'autres recherches ont été rendues possibles par le PRDM, qui soulignent son importance en matière de soutien à la recherche en IA, à la modernisation de la méthodologie et à la communauté gouvernementale de l'IA en général.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Marie-Eve Bedard (marie-eve.bedard@statcan.gc.ca) ou
Nabila Ould-Brahim (nabila.ould-brahim@statcan.gc.ca).

5.7 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 Centre de ressources en conception de questionnaires est effectuée selon le principe du recouvrement des coûts, la section est fréquemment sollicitée, 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. Même si la plupart de ces examens portaient sur des questionnaires de Statistique Canada, plusieurs ont été effectués pour des enquêtes menées par d'autres organismes gouvernementaux, comme Travaux publics et Services gouvernementaux Canada, Régie de l'énergie du Canada, Services publics et Approvisionnement Canada et d'autres.

Le groupe a également contribué à diverses initiatives de consultation de l'organisme.

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

5.8 Centre de ressources en assurance de la qualité

Le Centre de ressources en assurance de la qualité (CRAQ) a pour mission de faire progresser la recherche et développement dans les méthodes statistiques visant à améliorer les processus d'assurance et de contrôle de la qualité. Notre principal objectif est d'élever les normes des opérations de collecte et de traitement des données d'enquête au sein de l'organisme. Pour atteindre cet objectif, nous explorons un éventail de méthodologies, en mettant particulièrement l'accent sur l'amélioration de la qualité des données sortantes.

Au cœur de nos efforts se trouve la prestation de services méthodologiques pour G-Code, un système généralisé mis au point à Statistique Canada pour la création de bases de données codées et la mise en œuvre d'algorithmes d'apprentissage automatique dans le traitement des données. Nos recherches portent sur un large éventail de pratiques d'assurance et de contrôle de la qualité pour lesquelles nous devons résoudre des questions d'efficacité et d'automatisation. Les résultats de nos recherches sont pertinents non seulement pour nos activités internes, mais également par leur grande applicabilité à différentes étapes des processus d'enquête.

Progrès :

L'équipe de soutien méthodologique a aidé l'équipe de développement de G-Code et a fait le suivi des commentaires des utilisateurs pour cibler des possibilités d'amélioration de G-Code. De plus, le CRAQ a apporté son soutien aux utilisateurs internes et externes de G-Code chaque fois que de l'aide, des commentaires ou des suggestions concernant G-Code étaient nécessaires.

Tout au long de l'année, le CRAQ s'est consacré à mettre en œuvre une nouvelle méthodologie appelée « Contrôle de la qualité par score » afin d'améliorer le contrôle de la qualité (CQ) des processus de codage de texte par apprentissage automatique (AA). Comme la technologie d'AA joue un rôle de plus en plus essentiel dans le traitement des données, il est primordial d'assurer la qualité des codes générés. Pour relever ce défi, Statistique Canada cherche activement à élaborer une stratégie pour déterminer les taux d'échantillonnage optimaux aux fins de CQ en utilisant des scores tirés du processus d'apprentissage automatique. Cette méthodologie favorise une approche responsable de classification des données tout en facilitant l'adoption plus large de l'apprentissage automatique. Notre objectif est d'appliquer cette méthode à des fins de CQ pour plusieurs classifications au sein d'enquêtes essentielles comme l'Enquête sur la population active (EPA), l'Enquête sur les postes vacants et les salaires, l'Enquête sur la santé dans

les collectivités canadiennes et le Registre statistique des entreprises (RSE). Un article exposant en détail cette méthodologie a notamment été présenté au Comité consultatif des méthodes statistiques (Oyarzun, Wile et Evans, 2023).

De plus, le CRAQ a aidé l'équipe de l'EPA à mettre à jour ses données pour refléter les plus récentes classifications des industries et des professions, ce qui a entraîné d'importants changements structurels. Traditionnellement, les catégories fractionnées étaient traitées au moyen d'un codage manuel et d'une répartition aléatoire. Un nouveau cadre hybride, combinant l'apprentissage automatique (fastText) et la programmation linéaire, a été élaboré pour améliorer l'efficacité tout en maintenant la cohérence avec les estimations traditionnelles. Ces travaux ont été présentés au Symposium de 2024 de Statistique Canada (Evans et Wile, 2024).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Javier Oyarzun (javier.oyarzun@statcan.gc.ca).

Bibliographie

Evans, J. et Wile, L. (2024). Filer sur la voie du FastText : Exploiter l'apprentissage automatique restreint par la programmation linéaire pour réviser les classifications. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Oyarzun, J., Wile, L. et Evans, J. (2023). Quality control by score. Article présenté au Comité consultatif sur les méthodes statistiques, octobre 2023, Statistique Canada.

5.9 Secrétariat de l'éthique des données

Le rôle du Secrétariat à l'éthique des données est de mettre en œuvre le Cadre de nécessité et de proportionnalité. Concrètement, le Secrétariat de l'éthique des données effectue des examens éthiques des nouvelles acquisitions de données par l'entremise d'enquêtes ou d'autres sources et des nouvelles utilisations de données, comme le couplage de microdonnées. Son travail a récemment été élargi pour inclure des examens liés à l'apprentissage automatique et à l'intelligence artificielle. L'objectif de ces examens éthiques est d'assurer une utilisation responsable des données tout au long de leur cycle de vie. Le Secrétariat de l'éthique des données soulève des considérations éthiques, tient des discussions avec les gestionnaires de programme et formule des recommandations au dirigeant principal de l'éthique des données, de la qualité et de l'intégrité scientifique. Le Secrétariat à l'éthique des données appuie également le Comité interne d'éthique des données et joue un rôle de renforcement des capacités.

Progrès :

En plus d'avoir mené plus de 150 examens éthiques au cours de la dernière année seulement, les membres du Secrétariat de l'éthique des données ont fait de nombreuses présentations pour informer les partenaires internes et des collègues d'autres ministères fédéraux et d'organismes internationaux de l'approche de Statistique Canada en matière d'évaluation de l'éthique des données. L'équipe recueille des renseignements pour se tenir au courant des sujets qui pourraient être jugés délicats par le public. Pour ce faire, elle procède à des revues de la littérature sur certains sujets ciblés et à des discussions informelles avec des partenaires internes, comme les Communications et le Centre de ressources en conception de

questionnaires, ainsi qu'avec des homologues d'autres ministères fédéraux ou de bureaux nationaux de la statistique à l'étranger.

Outre ses activités internes, l'équipe est très active à l'échelle internationale, jouant un rôle de chef de file au sein de l'équipe de travail de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) sur le leadership éthique. Le principal objectif de cette équipe de travail est de rédiger un ouvrage de référence sur l'éthique destiné aux organismes nationaux de statistique. Les travaux sur cet ouvrage de référence ont été réalisés en 2025.

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

5.10 Secrétariat de la qualité

Le Secrétariat de la qualité a entre autres pour mandat de concevoir et de gérer des études liées à la gestion de la qualité et de répondre aux demandes de renseignements ou d'assistance en matière de gestion de la qualité provenant des divers programmes de Statistique Canada ou d'autres organismes.

PROJET : Renforcement des capacités avec des partenaires internes, nationaux et internationaux

Le Secrétariat de la qualité a pour objectif de donner des conseils et de prendre des mesures de renforcement des capacités à l'interne, avec des partenaires nationaux (d'autres ministères ou organismes) et internationaux, principalement en présentant un aperçu général des pratiques de gestion de la qualité de Statistique Canada et des documents officiels liés à la qualité (le Cadre d'assurance de la qualité et les Lignes directrices concernant la qualité) et en offrant des services de soutien en gestion de la qualité.

Progrès :

Le Secrétariat de la qualité a entrepris de renforcer les capacités de nombreux partenaires au cours de la période visée. À l'interne, divers cours ont été offerts au personnel. En ce qui concerne les partenaires nationaux, une présentation sur les pratiques de gestion de la qualité relativement au Modèle générique du processus de production statistique a été donnée au Centre de gouvernance de l'information des Premières Nations.

Des discussions ont eu lieu au sein du Groupe de travail sur la qualité des données de la Communauté de pratique sur les données ministérielles à l'échelle du gouvernement du Canada. Ce groupe de travail, coprésidé par Statistique Canada, a publié en janvier 2024 une version abrégée du cadre de la qualité des données, appelée « Orientation sur la qualité des données ». Depuis son relancement à l'automne 2024, ce groupe vise à cerner les lacunes dans la gouvernance en matière de qualité en effectuant une analyse de l'environnement sur des sujets comme les métadonnées, les données ouvertes et la production de rapports sur la qualité.

À l'échelle internationale, le Secrétariat à la qualité a rencontré le Centre de statistique d'Abu Dhabi pour lui fournir des conseils sur les tableaux de bord sur la qualité des données, tout en poursuivant son engagement auprès du groupe d'experts des Nations Unies sur les cadres nationaux d'assurance de la qualité. Statistique Canada a coprésidé le sous-groupe sur les sources de données administratives et

d'autres sources de données, chargé de préparer un module d'assurance de la qualité lors de l'utilisation de sources de données administratives et d'autres sources de données pour produire des statistiques officielles. Ce module, publié au début de 2025, vise à fournir des orientations pratiques et concises ainsi que des pratiques exemplaires aux organismes statistiques, afin d'assurer la qualité des statistiques officielles quand d'autres sources de données sont utilisées pour la production de statistiques officielles. Il s'utilise en complément du Manuel des cadres nationaux d'assurance de la qualité des Nations Unies en statistique officielle (Nations Unies, 2019).

PROJET : Examens du programme triennal

Dans le cadre des examens triennaux des programmes de Statistique Canada, le Secrétariat de la qualité a conçu un questionnaire d'autoévaluation qui a été rempli par les sept programmes concernés par les examens de 2024-2025. Ce questionnaire a permis à chaque programme d'évaluer son degré de préparation en matière de qualité et ses pratiques exemplaires actuelles en ce qui concerne sa culture et les six dimensions de la qualité de Statistique Canada. Le Secrétariat de la qualité continuera de jouer un rôle clé dans ces examens à mesure que leur portée s'élargit en 2025-2026.

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

Bibliographie

Nations Unies (2019). United Nations National Quality Assurance Frameworks Manual for Official Statistics. Disponible à l'adresse : https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/.

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-enquête. 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 et de décembre 2024 (50-1 et 50-2) ont été publiés. Le numéro de juin 2024 est un numéro spécial comprenant des articles présentés à la 29e conférence Morris Hansen sur l'utilisation d'échantillons non probabilistes par Courtney Kennedy, Yan Li et Jean-François Beaumont. Ces trois documents font l'objet de discussions par des experts internationaux du domaine, suivies de réponses. Une introduction de Partha Lahiri, rédacteur invité de ce numéro spécial, précède les articles. En tout,

16 articles ont été publiés dans le numéro de décembre 2024, dont l'article sollicité Waksberg de 2024 de Richard Valliant intitulé « Plan d'échantillonnage à partir de modèles ».

En 2024, 62 articles ont été soumis à la revue. Le nombre moyen de jours entre la soumission et l'évaluation initiale était de 43 jours. Tous les articles soumis ont été évalués dans un délai de 125 jours, et 77 % d'entre eux l'ont été dans un délai de 90 jours. Parmi ces 62 articles soumis, 37 ont été rejetés, 17 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 9 juin 2025. D'avril 2024 à mars 2025, les pages de *Techniques d'enquête* ont été consultées 46 472 fois.

Le numéro de juin 2025 sera consacré aux célébrations du 50e anniversaire de *Techniques d'enquête*. Il comprendra deux articles avec discussions. 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 a pour titre : « Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage ». Ces deux articles sont suivis de discussions d'éminents statisticiens d'enquête et d'une réponse. Dans le cadre de ce numéro spécial, nous sommes heureux d'annoncer le lancement d'une nouvelle section de la revue consacrée aux entretiens. Ces articles contiendront généralement 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 comprendra deux entretiens, un entretien avec Ivan P. Fellegi et l'autre avec Geoffrey Hole, pour commencer cette nouvelle section en force. Il comprendra également sept articles sollicités rédigés par des spécialistes renommés de la statistique et de la méthodologie d'enquête.

Nous prévoyons également publier deux numéros spéciaux en 2026 ou 2027. Le premier mettra en vedette 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 pour ce numéro spécial. Le deuxième numéro spécial portera sur le thème suivant : « Façonner l'avenir de la statistique d'enquête à l'ère des données ». Maria Rosaria Ferrante et Natalie Shlomo sont les rédactrices invitées pour ce numéro spécial.

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 liés 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 au moyen de 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 dans un court laps de temps et axées sur l'apprentissage actif. Cette année, le travail pour remanier le deuxième cours d'analyse statistique en présence de données d'enquête s'est poursuivi. Nous avons aussi élaboré un cours d'introduction aux simulations. Ce cours a été offert dans les deux langues officielles. Nous avons également travaillé à mettre au goût du jour un ancien cours sur l'estimation de la variance.

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, la version définitive du cours sur l'estimation de la variance sera établie et le cours sera offert 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, notamment les plateformes DataCamp et Onyxia, 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 Symposium international de 2024 sur les questions de méthodologie de Statistique Canada, dont le thème était « Le futur des statistiques officielles », a eu lieu les 30 et 31 octobre et le 1er novembre 2024. Le Symposium a offert des séances plénières ainsi que des séances parallèles qui ont porté sur divers thèmes. Tout comme d'autres congrès dans le monde entier, le Symposium de 2024 a permis à des conférenciers de donner leurs présentations en présentiel. Les observateurs avaient le choix d'y assister en personne ou de se joindre aux séances sous forme virtuelle.

Progrès :

Le Symposium de 2024 comptait environ 325 participants provenant de la Direction des méthodes statistiques modernes et de la science des données, ainsi que 150 participants qui travaillaient au sein d'autres directions de l'organisme ou qui représentaient d'autres organismes. Au total, 9 pays étaient représentés parmi les participants.

Les articles pour les actes du Symposium ont été recueillis, vérifiés et corrigés et sont actuellement à l'étape de la traduction. La publication des actes est prévue pour l'été 2025.

Des informations sur le Symposium 2024 sont disponibles sur notre site Web à l'adresse suivante :
https://www.statcan.gc.ca/fr/conferences/symposium2024/index.

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

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.

Boulet, C. et Mather, A. (2025). Methodology Acceleration Initiative Research Project: Subsampling for Non-Response Follow-Up in Social Surveys. Document interne, Statistique Canada.

Dasylva, 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*, 41(2), 569-597, OnlineFirst, Numéro spécial, https://journals.sagepub.com/doi/10.1177/0282423X251329407.

Dasylva, 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. À paraître dans *Statistical Journal of the International Association for Official Statistics*.

Evans, J. et Wile, L. (2024). Filer sur la voie du FastText : Exploiter l'apprentissage automatique restreint par la programmation linéaire pour réviser les classifications. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Ferland, M. (2025). gseries: Improve the Coherence of Your Time Series Data. R package version 3.0.2, https://StatCan.github.io/gensol-gseries/fr/.

Francis, J., Carrillo-Garcia, I., Yélou, C. et Rassart, E. (2025). Developing CPI Quality Indicators: Project Update. Document interne, Statistique Canada.

Gauvin, H. (2025). Practical Guidelines for the Creation of Synthetic Data Files. Document interne, Statistique Canada.

Gauvin, H. (2024). Creating a synthetic version of a longitudinal and structured file: Challenges and lessons. *Recueil de la Section des méthodes d'enquête, Société Statistique du Canada,* St. Johns, Terre-Neuve, https://ssc.ca/sites/default/files/imce/gauvin_ssc2024.pdf.

Gray, D. (2024). Building the new Banff: An open-source data editing system based on GSDEM concepts. UNECE Expert Meeting on Statistical Data Editing, 7 au 9 octobre 2024, Vienne.

Gray, D. et Pierre, F. (2025). De SAS vers les sources libres : conversion des systèmes généralisés de Statistique Canada. À paraître dans le prochain numéro de *Convergence*, une revue de l'Association des statisticiennes et statisticiens du Québec.

Gray, D. et Seffal, M. (2025a). The New Banff: A Modern Edit and Imputation Platform for Everyone. Présentation interne, Statistique Canada.

Gray, D. et Seffal, M. (2025b). From SAS to Open-Source: What we learned from the Banff project. Présentation interne, Modern Statistical Methods and Data Science Branch Learning Week, Statistique Canada.

Hatko, S. et Sall, A. (2024). International Trade Unit Value Error Detection and Correction. Rapport interne, Statistique Canada.

Istrate, A. et Robatian, D. (2025). Evaluation of AI-Generated Text: Survey, Taxonomy and Practical Framework for Designing NLG Evaluation Pipelines. Rapport interne, Statistique Canada.

Lafortune, Y. et Mayer, E. (2025). The Evolution of the Quantity and Quality of the Data Collected for the Survey of Household Spending Through the Years. Document de recherche, Statistique Canada.

Loewen, R. et Jin, N. (2025a). Confidence intervals for Social Data Linkage Environment Error Rates. Document interne, Statistique Canada.

Loewen, R. et Jin, N. (2025b). Confidence intervals for Social Data Linkage Environment Error Rates. Diapositives pour présentation au Comité d'examen scientifique, 23 mai 2025, Statistique Canada.

Loewen, R. et Jin, N. (2025c). Confidence intervals for Social Data Linkage Environment Error Rates. Diapositives pour le séminaire divisionnaire, 12 juin 2025, Statistique Canada.

MacIsaac, K., Boulet, C. et Thomas, M. (2024). Recrutement et collecte de panels Web à Statistique Canada. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Mather, A. et Boulet, C. (2024). *Analysis of Residual Bias on Web Panels*. Document de travail de la Direction de la méthodologie (SSMD-2024-002E), Statistique Canada.

Mather, A., Boulet, C. et Ra, Y. (2024). Évaluation des biais liés aux panels Web probabilistes de Statistique Canada. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Miller, J. (2025). A Privacy-Utility Based Study of Differential Privacy. Rapport interne, Statistique Canada.

Shukla, A., Rossiter, E. et Santos B. (2024). Input/Output Privacy Framework: A Privacy-Preserving Record Linkage System. Rapport interne, Statistique Canada.

Spackman, W., Francis, J. et Goussev, S. (2025). Optimal use of machine learning at scale: Designing quality control of machine learning classification and mitigating misclassification error. *Proceedings of the 17th Meeting of the Group of Experts on Consumer Price Indices*, CEE-ONU, Genève, Suisse. Disponible à l'adresse : https://unece.org/sites/default/files/2025-04/Spackman%20et%20al%20%282025%29%20-%20QC%20for%20mitigating%20misclassification%20bias%20-%20paper.pdf.

Tang, C. (2024). Analyse du contrôle statistique de la divulgation pour les estimations sur petits domaines. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Toupin, M.-H. et Martin, V. (2025). Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application à l'estimation du questionnaire détaillé du recensement canadien. *Techniques d'enquête* (en cours de révision).

Verret, F. et Walker, B. (2024). Rétro-ingénierie d'un processus de réconciliation hypothétique pour estimer l'erreur quadratique moyenne (EQM) des estimations sur petits domaines réconcilié. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.

Volkov, O. (2025). Bayesian nonparametric methods for survey data. *The Survey Statistician*, 91, 34-43, https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_January_N91.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. À paraître dans le numéro de décembre 2025 de *Techniques d'enquête*.

Yu, Z. (2024). Évaluation des risques liés à la divulgation de données synthétiques. *Recueil : Symposium 2024, Le futur des statistiques officielles*, Statistique Canada, Ottawa, Canada.