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

Programme de recherche et développement en méthodologie : réalisations, 2023-2024

Date de diffusion : le 25 octobre 2024

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, 2023-2024

Le présent rapport fait la synthèse des réalisations en 2023-2024 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 comprend les 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 s'exerceraient pas complètement 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 les 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, 2023-2024

Table des matières

1 Intégration de données

1.1 Intégration de données d'échantillons probabilistes et non probabilistes

PROJET : Estimation paramétrique de probabilités de participation pour des échantillons non probabilistes

Statistique Canada et d'autres organismes nationaux de statistique étudient actuellement la possibilité d'utiliser des échantillons non probabilistes comme solution de rechange aux échantillons probabilistes. On sait toutefois très bien que l'utilisation d'un échantillon non probabiliste seulement peut produire des estimations présentant un biais significatif en raison de la nature inconnue du mécanisme de sélection sous-jacent. Pour que ce biais soit réduit, les données d'un échantillon non probabiliste peuvent être intégrées aux données d'un échantillon probabiliste à condition que les deux échantillons contiennent des variables auxiliaires communes.

Dans la présente recherche, nous nous intéressons à l'estimation paramétrique de la probabilité de participer à l'échantillon non probabiliste, qui est une quantité clé pour réduire le biais de participation. Plusieurs méthodes d'estimation de la probabilité de participation ont été proposées récemment (par exemple Elliott, 2009; Chen, Li et Wu, 2020; Wang, Valliant et Li, 2021). La méthode par la pseudo-vraisemblance de Chen, Li et Wu (2020), qui a attiré l'attention de nombreux chercheurs, s'est récemment révélée inefficace dans certains scénarios. Les méthodes d'Elliott (2009) et de Wang, Valliant et Li (2021) peuvent parfois être plus efficaces, mais elles ne sont pas équivalentes à la méthode du maximum de vraisemblance quand l'échantillon probabiliste est un recensement.

Le problème de l'estimation des probabilités de participation ressemble à celui de l'estimation des probabilités de réponse aux enquêtes, la principale différence étant la structure des renseignements disponibles. Le but de la présente recherche était de trouver des moyens raisonnés d'exploiter à la fois les données auxiliaires des échantillons probabiliste et non probabiliste.

Progrès :

Dans la première partie du projet, nous avons élaboré une amélioration de la méthode de Chen, Li et Wu (2020), fondée sur la théorie de la meilleure estimation linéaire sans biais, qui tire plus efficacement parti des données disponibles des échantillons probabiliste et non probabiliste. De plus, nous avons élaboré une méthode de vraisemblance de l'échantillon, dont l'esprit ressemble à celui de la méthode d'Elliott (2009), qui tient adéquatement compte du chevauchement entre les deux échantillons quand il est possible de l'identifier dans au moins un des échantillons. Nous avons utilisé la théorie de la meilleure prédiction linéaire sans biais pour traiter le scénario où le chevauchement est inconnu. Il est intéressant de constater que les deux méthodes que nous proposons coïncident en cas de chevauchement inconnu. Ensuite, nous avons montré que de nombreuses méthodes existantes peuvent être obtenues comme cas particulier d'une fonction d'estimation sans biais générale. Ces résultats théoriques ont été documentés et publiés dans le numéro de juin 2024 de *Techniques d'enquête* (Beaumont, Bosa, Brennan, Charlebois et Chu, 2024). Cet article est une réponse aux commentaires sur un article des mêmes auteurs intitulés « Traiter les échantillons non probabilistes par pondération de probabilité inverse avec une application aux données d'approche participative de Statistique Canada », également publié dans le numéro de juin 2024 de *Techniques d'enquête*.

Dans la deuxième partie du projet, nous avons mené des expériences de simulation qui ont confirmé les gains d'efficacité des méthodes que nous avons proposées. Nous comptons documenter nos conclusions au cours de l'exercice financier 2024-2025.

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

Bibliographie

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2024). Réponse des auteurs aux commentaires sur l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada » : De nouvelles avancées concernant les méthodes de vraisemblance pour l'estimation des probabilités de participation pour des échantillons non probabilistes. *Techniques d'enquête*, 50, 1, 139-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00001-fra.pdf.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

1.2 Couplage d'enregistrements

PROJET : Capture-recapture avec erreurs de couplage

En vue de prioriser l'utilisation de sources administratives, Statistique Canada vise à mesurer leur couverture au moyen d'opérations sur le terrain ou d'examens administratifs, ce qui peut être coûteux. La méthode de capture-recapture peut offrir une solution économique au moyen d'une comparaison avec une autre source selon des hypothèses types, comme l'indépendance des sources et leur couplage parfait. Il faut toutefois l'adapter quand le couplage est imparfait. Jusqu'à présent, ces améliorations nécessitaient des examens manuels (Ding et Fienberg, 1994; Di Consiglio et Tuoto, 2015; de Wolf, van der Laan et Zult, 2019) ou une forte hypothèse selon laquelle les variables de couplage sont conditionnellement indépendantes (Racinskij, Smith et van der Heijden, 2019).

Progrès :

Une nouvelle méthode d'estimation par capture-recapture avec erreurs de couplage, qui n'est pas touchée par les limites mentionnées plus haut, a été proposée. Elle fonctionne en modélisant le nombre de liens à partir d'un enregistrement et s'appuie sur le modèle statistique décrit précédemment par

Dasylva et Goussanou (2022). La méthodologie est décrite dans un article accepté aux fins de publication dans *Techniques d'enquête* (Dasylva, Goussanou et Nambeu, 2024).

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

Bibliographie

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Consulté sur https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., Goussanou, A. et Nambeu, C.-O. (2024). Modèles d'erreur de couplage pour l'estimation par capture-recapture sans vérifications manuelles. *Techniques d'enquête*, 50, 2. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024002/article/00007-fra.pdf.

de Wolf, P.-P., van der Laan, J. et Zult, D. (2019). Connection correction methods for linkage error in capture-recapture. *Journal of Official Statistics*, 35, 577-597.

Di Consiglio, L., et Tuoto, T. (2015). Coverage Evaluation on Probabilistically Linked Data. *Journal of Official Statistics*, 31, 415-429.

Ding, Y., et Fienberg, S.E. (1994). Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement. *Techniques d'enquête*, 20, 2, 155-165. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1994002/article/14422-fra.pdf.

Racinskij, V., Smith, P. et van der Heijden, P. (2019). Linkage free dual system estimation. Disponible à l'adresse https://arxiv.org/abs/1903.10894.

PROJET : Mesure de la précision du couplage quand de nombreux fichiers sont couplés à un univers

Pour effectuer une analyse statistique, il est parfois nécessaire de coupler de nombreux fichiers à un univers, qui est en fait la liste d'unités uniques d'une population cible. Toutefois, il se peut que les liens en résultant soient imparfaits, quand les renseignements d'identification sont partiels ou erronés. Dans de tels cas, il faut évaluer l'exactitude du couplage en fonction des tuples d'enregistrement de trois enregistrements ou plus. Des solutions fondées sur des examens manuels sont décrites par Binette, Baek, Engineer, Jones, Dasylva et Reiter (2024). Cette méthode peut cependant être coûteuse. Une solution de rechange pourrait reposer sur l'extension du modèle de Fellegi-Sunter par Sadinle et Fienberg (2013) ou sur des modèles de résolution d'entités bayésiennes proposés par Steorts, Hall et Fienberg (2016), Marchant, Kaplan, Elazar, Rubinstein et Steorts (2021) ou autres. Toutefois, ces modèles reposent sur une hypothèse forte d'indépendance conditionnelle (Fellegi et Sunter, 1969), ou des variantes de cette hypothèse.

Progrès :

On propose une nouvelle méthodologie, qui ne demande ni examens manuels ni hypothèse d'indépendance conditionnelle. Dans cette méthodologie, le rappel et la précision sont étendus aux tuples d'enregistrement avec trois enregistrements ou plus, et les mesures qui en résultent sont évaluées au

moyen de la modélisation du vecteur, comprenant le nombre de liens vers chaque fichier pour chaque enregistrement d'univers. Dans les faits, le modèle a une portée plus vaste que celui décrit précédemment par Dasylva et Goussanou (2022) et s'étend dans un mélange multivarié fini, où chaque composante est la convolution d'une distribution multinomiale et d'une distribution de Poisson composée indépendante et multivariée. En plus d'informer les utilisateurs des données couplées, la méthodologie comporte de nombreuses autres applications pratiques, comme l'évaluation de la fiabilité de la moyenne qui repose sur les données couplées, par exemple grâce à une mesure qui s'apparente à un coefficient de variation, c'est-à-dire une mesure semblable au coefficient de variation. Cet aspect est utile pour savoir si les erreurs de couplage peuvent être ignorées en toute sécurité ou si d'autres étapes sont nécessaires. La méthodologie fournit également une base pour l'évaluation d'un univers, qui repose elle-même sur le couplage imparfait de nombreux fichiers administratifs, comme la probabilité que de nombreux enregistrements représentent, en réalité, une même personne étant donné qu'ils sont classés comme tels. Pour plus de détails, voir Dasylva et Goussanou (2024a).

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

Bibliographie

Binette, O., Baek, Y., Engineer, S., Jones, C., Dasylva, A. et Reiter, J. (2024). How to evaluate entity resolution systems: An entity-centric framework with application to inventor name disambiguation. Disponible à l'adresse : https://arxiv.org/pdf/2404.05622.pdf.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Consulté sur https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., et Goussanou, A. (2024a). Measuring the linkage accuracy when many files are linked to a spine. Rapport interne, Statistique Canada.

Fellegi, I., et Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Marchant, N., Kaplan, A., Elazar, D., Rubinstein, B. et Steorts, R. (2021). d-blink: Distributed end-to-end bayesian entity resolution. *Journal of Computational and Graphical Statistics*, 30, 406-421.

Sadinle, M., et Fienberg, S.E. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108, 385-397.

Steorts, R., Hall, R. et Fienberg, S. (2016). A bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660-1672.

PROJET : Estimation des moyennes de petits domaines avec données couplées

Des techniques d'estimation sur petits domaines sont couramment utilisées pour produire des estimations pour les domaines où la taille d'échantillon est trop petite, par la modélisation de la variable d'intérêt à partir de l'information auxiliaire (Rao et Molina, 2015). En présence d'un modèle au niveau de

l'unité, il faut obtenir cette information pour chaque unité échantillonnée, ce qui comprend un couplage d'enregistrements si l'information auxiliaire se trouve dans un fichier distinct. En général, ce couplage peut représenter une source d'erreurs dont il faut tenir compte. Salvati, Fabrizi, Ranalli et Chambers (2021) ont proposé une solution pour l'application du modèle linéaire à erreurs emboîtées, où les variables auxiliaires sont obtenues à partir d'un registre, qui est distinct de la base de sondage et qui contient un indicateur d'inclusion dans l'échantillon rarement disponible dans la pratique. Ils exigent également des estimations manuelles des taux d'erreur de couplage, dont la réalisation peut être coûteuse.

Progrès :

Pour dépasser les limites indiquées plus haut, on propose une solution qui repose sur l'extension de la méthodologie de Salvati, Fabrizi, Ranalli et Chambers (2021), au moyen du modèle d'erreur de couplage décrit par Dasylva et Goussanou (2022). Cette méthodologie a été présentée aux Joint Statistical Meetings de 2023. Pour plus de détails, voir Dasylva (2024).

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

Bibliographie

Dasylva, A. (2024). Estimation of small area means with linked data. Rapport interne, Statistique Canada.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*, 5, 181-216. https://doi.org/10.1007/s42081-022-00153-3.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. New Jersey: Wiley.

Salvati, N., Fabrizi, E., Ranalli, M. et Chambers, R. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society B*, 83, 78-107.

PROJET : Sélection de seuil fondé sur un modèle pour couplages probabilistes en agriculture

L'importance croissante de l'utilisation de données administratives dans la production de statistiques officielles rend le problème du couplage d'enregistrements sans identificateur unique de plus en plus important. La méthode probabiliste proposée par Fellegi et Sunter (1969) est une démarche courante, qui est mise en œuvre dans le Système généralisé de couplage d'enregistrements de Statistique Canada, G-Coup. Cependant, le seuil de pondération, un paramètre sur lequel repose l'optimalité de la procédure, peut se révéler difficile à définir. Les solutions proposées jusqu'à maintenant pour établir le seuil présentent d'importantes limites, soit parce qu'elles s'appuient sur des hypothèses optimistes, soit parce qu'elles exigent des données d'entraînement ou un examen manuel, deux processus potentiellement coûteux. Cependant, un modèle a été élaboré qui permet d'estimer l'erreur de couplage à partir du nombre de liens d'un enregistrement donné, en tenant compte de toutes les interactions entre les variables de couplage tout en atténuant le besoin de tout processus manuel (Dasylva et Goussanou, 2022).

Progrès :

À partir du modèle d'erreur de Dasylva et Goussanou, différents algorithmes ont été évalués aux fins de sélection du seuil de couplage. Que ce soit la recherche exhaustive, la recherche binaire ou une procédure de partition récursive plus sophistiquée proposée par Dasylva et Chen (2022), chaque méthode offre des avantages différents pour ce qui est du temps d'exécution et de la qualité des mesures produites. La méthodologie a également été comparée à la méthode proposée par Belin et Rubin (1995) et à la démarche de la théorie des valeurs extrêmes (Sariyar, Borg et Pommerening, 2011) au moyen de données simulées pour montrer les résultats de l'utilisation du modèle quand la réalité sur le terrain était connue. Dans les deux cas, le modèle d'erreur de Dasylva et Goussanou a permis d'estimer la précision du couplage de manière plus exacte. Enfin, conformément à la méthode de Dasylva et Goussanou (2024b), des tests d'adéquation de l'ajustement du chi carré ont été appliqués à des paires de couplage de données agricoles pour vérifier que le modèle s'applique sur des données réelles. Pour tous les couplages testés, les données ont réussi les tests d'adéquation de l'ajustement, ce qui a montré que le nouveau modèle d'erreur et la méthode de sélection de seuils sont utiles dans une configuration pratique.

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

Bibliographie

Belin, T., et Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Dasylva, A., et Chen, W. (2022). Probabilistic record linkage through recursive partitioning without training data. Présentation à la réunion mensuelle du groupe Machine Learning de l'ONS-UNECE, avril 2022.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Disponible à l'adresse : https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., et Goussanou, A. (2024b). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*. Disponible à l'adresse : https://doi.org/10.1007/s42081-023-00228-9.

Fellegi, I., et Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Sariyar, M., Borg, A. et Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

PROJET : Analyse secondaire de données catégoriques couplées

Il est de plus en plus courant pour les analystes qui ont recours à des données de Statistique Canada d'utiliser le couplage d'enregistrements pour ajouter de la profondeur aux fichiers d'enquête. Par exemple, les enquêtes sociales sont fréquemment couplées à des données fiscales et sur l'immigration au moyen de l'Environnement de couplage de données sociales (ECDS). L'utilisation de données couplées

sert à réduire le fardeau de réponse et fournit aux analystes des valeurs plus exactes pour certaines variables (par exemple le revenu) que ce qui pourrait être obtenu directement des répondants.

Cependant, de nombreux couplages dans l'ECDS sont probabilistes, et l'erreur de couplage est un sous-produit inévitable du couplage d'enregistrements probabiliste. L'analyse de données couplées qui ne corrigent pas la présence d'erreurs de couplage peut entraîner des résultats biaisés.

La plupart des équipes utilisant les données couplées de Statistique Canada sont des utilisateurs « secondaires » des données. Cela signifie qu'ils ont accès aux variables analytiques dans leurs fichiers d'intérêt et qu'ils reçoivent un ensemble de clés de couplage servant à joindre les fichiers, mais qu'ils n'ont pas accès aux variables des microdonnées (par exemple noms, dates de naissance) qui ont servi à effectuer les couplages. Afin de protéger les renseignements de nature délicate de la population canadienne, les employés ou les chercheurs externes ayant accès aux variables analytiques ne sont généralement pas autorisés à accéder aux variables de couplage.

Récemment, Li-Chun Zhang et Tiziana Tuoto (2021) ont proposé une nouvelle méthode prometteuse pour intégrer l'erreur de couplage à l'analyse des données couplées. Lors du Symposium international sur les questions de méthodologie de 2022 de Statistique Canada, Li-Chun Zhang (2022) a présenté une méthode élaborée pour réaliser une régression logistique du point de vue d'un analyste secondaire de façon à corriger les faux appariements. Le principal objectif de ce projet était d'évaluer cette méthode et d'élaborer des programmes pour l'utiliser avec des fichiers couplés au moyen de l'ECDS.

Progrès :

L'algorithme de Zhang et Tuoto a été mis en œuvre en Python, et une étude par simulations a permis de comparer ses performances à celles de la régression logistique naïve. Des versions artificielles de couplages de l'ECDS ont été créées par échantillonnage à partir de l'ensemble couplé d'un couplage réel de l'ECDS. Les poids de couplage disponibles dans les fichiers de couplage artificiels ont été affinés au moyen de l'algorithme espérance-maximisation (une technique qui avait d'abord été proposée pour le couplage d'enregistrements par Winkler (2000) et utilisée pour aider à simuler le processus de l'ECDS pour estimer la proportion de faux appariements pour un couplage donné). Enfin, des variables analytiques catégoriques simulées ont été ajoutées à chaque couplage artificiel. Ces variables étaient inspirées par les variables d'une analyse des données couplées réalisée en interne par Statistique Canada. Le programme Python a été alimenté par les ensembles artificiels de données analytiques couplées (et les taux d'erreur estimés), et les performances de l'algorithme ont été évaluées.

Dans certaines simulations, la méthode de Zhang et Tuoto a donné une réduction du biais et une légère inflation de la variance, comparativement à ce qui est obtenu au moyen d'une régression logistique naïve. Ces résultats représentent probablement ce qui se produirait dans un cas « typique ».

Cependant, pour certains autres ensembles de simulations, dans lesquels des covariables très déséquilibrées étaient présentes dans les ensembles artificiels de données couplées, il y a eu des cas où l'algorithme de régression naïve standard convergeait tandis que la méthode de Zhang et Tuoto ne convergeait pas. On a essayé plusieurs correctifs standards et des partenaires à l'interne ont été consultés pour résoudre ce problème, mais aucune solution n'a été trouvée. Finalement, Zhang et Tuoto ont apporté une solution intéressante. Ils ont proposé d'utiliser certains regroupements de fonctions R

standards (comme *glm* ou *glmet*) afin de mettre en œuvre une méthode itérative par les moindres carrés pondérés pour estimer les coefficients de régression. Cette méthode est à l'étude à l'heure actuelle.

Enfin, notons que l'une des entrées (facultatives) (la variance estimée de l'estimateur du taux de faux appariement) de la fonction Python écrite pour mettre en œuvre la méthode de Zhang et Tuoto n'est pas un résultat standard produit par l'ECDS, mais qu'elle pourrait être produite au moyen d'un programme SAS développé dans le cadre d'un projet de recherche mené pendant l'exercice précédent (Loewen et Millar, 2023).

Les résultats de ce projet comprennent un programme Python pour la mise en œuvre de la méthode de Zhang et Tuoto avec des données de l'ECDS qui seront mises à la disposition des utilisateurs par l'intermédiaire du Centre de ressources en couplage d'enregistrements, un rapport de stage de Toukal (2023) et un exposé présenté à une conférence par Millar (2024).

Ce projet a été réalisé en collaboration avec Julia Toukal et Abdelnasser Saidi, avec l'aide de Li-Chun Zhang, Tiziana Tuoto, Abel Dasylva, Mark Stinner et Kenza Sallier.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Goldwyn Millar (goldwyn.millar@statcan.gc.ca).

Bibliographie

Loewen, R., et Millar, G. (2023). Variance estimation for record linkage error-rates obtained via clerical review of stratified systematic samples of linked pairs. Présentation au séminaire de méthodologie le 10 mai 2023, Statistique Canada, Ottawa.

Millar, G. (2024). Logistic regression on linked data from a secondary analyst perspective. Présentation à la CANSSI-CRT Workshop on Modern Methods in Survey Sampling, University of Ottawa, 8 au 10 juillet.

Toukal, J. (2023). Logistic regression in the context of record linkage. Rapport de stage pour l'École Nationale de la Statistique et de l'analyse de l'information.

Winkler, W. (2000). *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*. Bureau of the Census, Statistical Research Division, Statistical Research Report Series, No. RR2000/05.

Zhang, L.-C. (2022). Analyse secondaire des données catégoriques couplées. Présentation au Symposium international de 2022 sur les questions de méthodologie, Statistique Canada.

Zhang, L.-C., et Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society, Series A*, 184, 522-554.

PROJET : Méthodes de création de pochettes par apprentissage automatique pour le couplage d'enregistrements : application de l'algorithme de couverture séquentielle et de la méthode des pochettes fondée sur les jetons de grande valeur au projet de couplage d'enregistrements des visiteurs réalisé par l'Environnement de couplage de données sociales de Statistique Canada

Une étape cruciale du prétraitement dans le couplage d'enregistrements consiste à choisir une technique de création de pochettes qui garantit que le nombre de paires d'enregistrements à comparer est gérable sur le plan du calcul et en même temps, couvre le plus grand nombre possible de vrais appariements. Une technique de création de pochettes efficace vise ces deux objectifs.

Le projet cherche à évaluer deux méthodes de création de pochettes par apprentissage automatique au moyen d'un exemple de couplage d'enregistrements réalisé à Statistique Canada : 1) l'algorithme de couverture séquentielle (*Sequential Covering Algorithm*) présenté dans Michelson et Knoblock (2006); 2) la méthode des pochettes fondée sur les jetons de grande valeur (*High-Value Token-Blocking*) d'O'Hare, Jurek-Loughrey et De Campos (2021). Intuitivement, l'algorithme de couverture séquentielle apprend une intersection d'attributs, appelée une règle, qui couvre certains exemples positifs, et apprend une autre intersection d'attributs, en répétant cette étape jusqu'à ce qu'elle ne puisse plus découvrir de règle dont le rendement dépasse un seuil choisi pour la mesure appelée intégralité des paires. L'intégralité des paires mesure la couverture des vrais positifs, c'est-à-dire le nombre de vrais appariements se trouvant dans l'ensemble candidat de paires d'enregistrements à comparer selon les règles établies par rapport au nombre d'appariements de l'ensemble complet. La méthode des pochettes fondée sur les jetons de grande valeur repose sur la fréquence du terme – fréquence inverse du document bien connue (*Term Frequency-Inverse Document Frequency* [TF-IDF]) et sa capacité à reconnaître les jetons de grande valeur, c'est-à-dire un terme ou un mot dans un document. Les valeurs TF-IDF indiquent l'importance d'un jeton en fonction de sa similitude dans un ensemble de données, les valeurs les plus faibles indiquant des jetons insignifiants très fréquents et les valeurs plus élevées indiquant des jetons significatifs plus rares.

Progrès :

Les performances des algorithmes ont été évaluées sur des ensembles de données de petite, moyenne et grande taille avant qu'ils soient mis à l'essai sur un exemple de couplage d'enregistrements effectué à Statistique Canada. L'intégralité des paires et le ratio de réduction ont servi de paramètres aux fins de l'évaluation. Le ratio de réduction indique la mesure dans laquelle une méthode de création de pochettes réduit le nombre de paires candidates qu'il faut comparer pour trouver les enregistrements correspondants. L'évaluation a montré que les algorithmes donnaient de bons résultats pour les ensembles de données de petite et de moyenne taille, mais qu'ils nécessitaient l'utilisation de techniques comme le partitionnement, le traitement parallèle et l'échantillonnage pour traiter les grands ensembles de données. Les algorithmes ont ensuite été évalués dans le cadre d'un projet de couplage d'enregistrements de Statistique Canada qui vise à apparier les enregistrements d'un fichier contenant des renseignements sur les visiteurs au Canada à un autre fichier appelé Dépôt pour déterminer les appariements. Cette évaluation visait à déterminer si ces algorithmes présentaient de meilleures performances que la méthode de création de pochettes utilisée actuellement pour ce couplage. L'algorithme de couverture séquentielle a obtenu de bons résultats avec un ratio de réduction supérieur à 99 % et une intégralité des paires de 83 %, ce qui est semblable aux performances de la méthode actuelle de création de pochettes. Cependant, la méthode actuelle crée mille fois plus de paires candidates à comparer pour trouver les enregistrements correspondants. Concernant la méthode des pochettes fondée sur les jetons de grande valeur, la mise en œuvre de la méthode d'échantillonnage est en cours.

Quelques ajustements sont nécessaires pour rendre la méthode plus efficace en matière de temps d'exécution.

Les prochaines étapes seront les suivantes :

1) terminer la mise en œuvre de la méthode d'échantillonnage pour la méthode des pochettes fondée sur les jetons de grande valeur;
2) déterminer une stratégie d'utilisation de l'algorithme de couverture séquentielle quand un couplage d'enregistrements est effectué pour la première fois, puisqu'il nécessite des données étiquetées.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Ronald Jean Paul (ronald.jeanpaul@statcan.gc.ca) ou
Bassirou Diagne (bassirou.diagne@statcan.gc.ca).

Bibliographie

Michelson, M., et Knoblock, C.A. (2006). Learning blocking schemes for record linkage. *Proceedings of the Association for Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 21, 440-445. Disponible à l'adresse : https://www.aaai.org/Papers/AAAI/2006/AAAI06-070.pdf.

O'Hare, K., Jurek-Loughrey, A. et De Campos, C. (2021). High-value token-blocking: Efficient blocking method for record linkage. *ACM Transactions on Knowledge Discovery from Data*, 16, 1-17. Disponible à l'adresse : https://doi.org/10.1145/3450527.

1.3 Estimation sur petits domaines

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

Quand la taille d'échantillon des domaines est petite, les estimateurs directs des paramètres de population qui sont convergents par rapport au plan de sondage sont susceptibles d'être instables. Pour améliorer la précision des estimateurs directs, on utilise souvent le modèle de Fay-Herriot au niveau du domaine. Celui-ci comprend deux composantes : un modèle d'échantillonnage et un modèle de liaison. Ce dernier spécifie la relation entre les paramètres d'intérêt de la population et des variables auxiliaires disponibles au niveau des domaines. Dans sa forme originale, le modèle de Fay-Herriot suppose un modèle de liaison linéaire dont la variance des erreurs est constante. Il requiert également d'estimer la variance lissée due au plan de sondage des estimateurs directs, c'est-à-dire l'espérance par rapport au modèle de la variance due au plan de sondage des estimateurs directs. Les estimateurs de la variance due au plan de sondage pourraient être considérés comme des estimateurs de la variance lissée, mais ils sont généralement instables en présence d'échantillons de petite taille. Pour résoudre ce problème, on lisse habituellement les estimations de la variance due au plan de sondage, souvent au moyen d'un modèle de lissage log-linéaire.

Or, les hypothèses des modèles de Fay-Herriot et de lissage ne sont pas toujours satisfaites en pratique, et la correction adéquate des modèles peut être difficile et prendre beaucoup de temps. Dans ce contexte, il peut être souhaitable d'utiliser des méthodes non paramétriques, surtout quand le nombre de domaines est élevé, parce que ces méthodes dépendent moins fortement de la validité des hypothèses des modèles et qu'elles peuvent accélérer la production d'estimation sur petits domaines. Nous nous intéressons

particulièrement aux forêts aléatoires pour trois raisons : i) elles peuvent être facilement appliquées dans le cas d'un mélange de variables auxiliaires catégoriques et continues; ii) elles ne nécessitent pas de spécifier des interactions; iii) elles produisent des prédictions qui demeurent toujours dans la plage des valeurs observées. Nous proposons une procédure bootstrap pour estimer l'erreur de prédiction quadratique moyenne.

Progrès :

Au cours de l'exercice précédent, nous avons élaboré des versions non paramétriques du meilleur prédicteur empirique quand les forêts aléatoires sont utilisées pour remplacer les modèles entièrement paramétriques. En 2023-2024, nous avons raffiné notre méthodologie et nous avons évalué les propriétés des meilleurs prédicteurs empiriques proposés au moyen de données réelles et d'études par simulations. Le prédicteur retenu utilise les prédictions hors du sac (*out-of-bag*) comme variable auxiliaire dans un modèle linéaire de Fay-Herriot. Nos résultats montrent que les forêts aléatoires offrent une robustesse par rapport à une spécification erronée des modèles, augmentent l'efficacité des estimations sur petits domaines et simplifient (mais sans éliminer) l'effort de modélisation. Ce projet sera présenté à l'International Conference on Establishment Statistics à Glasgow en juin 2024 et à la conférence de l'INCASS à Ottawa en juillet 2024. En 2024-2025, nous prévoyons la rédaction d'un article qui résumera nos résultats et la conception d'un programme en R qui mettra en œuvre notre méthodologie.

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

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

Le modèle au niveau du domaine de Fay-Herriot est souvent utilisé quand les estimateurs directs des paramètres de population sont instables en raison de la petite taille de l'échantillon. Le modèle de Fay-Herriot et d'autres modèles d'estimation sur petits domaines (EPD) reposent sur l'idée d'emprunter de la puissance à d'autres domaines. Toutefois, si le nombre de domaines est petit, le modèle de Fay-Herriot et les autres modèles d'EPD donnent généralement de mauvais résultats. L'idée du projet est d'emprunter de la puissance à d'autres domaines, mais aussi dans le temps au moyen d'une méthode hiérarchique bayésienne (HB).

La modélisation hiérarchique bayésienne est très populaire dans l'estimation sur petits domaines, et la spécification *a priori* est très importante dans cette approche. Dans le cadre de ce projet, nous étudions les performances des estimateurs HB sur petits domaines au moyen de distributions *a priori* non informatives et informatives pour les paramètres de régression et les composantes de la 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 évaluons l'effet des distributions *a priori* sur les estimateurs HB. Nous étudions l'effet des distributions *a priori* informatives et non informatives correctes et incorrectes pour l'estimation HB sur petits domaines en utilisant à la fois une application de l'EPA et une étude par simulations.

Progrès :

Pour les composantes de la variance, nous avons étudié les spécifications de la distribution *a priori* en utilisant des distributions *a priori* gamma inverses et des distributions *a priori* uniformes pour les modèles

HB sur petits domaines. Nous avons mené une étude par simulations et appliqué les modèles aux données de l'EPA. Un document de recherche (You, 2023) a été publié dans *Statistics in Transition new series*.

Concernant les paramètres de régression, nous avons étudié l'utilisation de distributions *a priori* informatives et non informatives au moyen d'une application de l'EPA et d'une étude par simulations. Un document de recherche (You et Bosa, 2024) a été rédigé et soumis à une revue pour une éventuelle publication. Le document est en processus de révision.

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. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition New Series*, 24, 169-178.

You, Y., et Bosa, K. (2024). Performance of hierarchical Bayes small area estimators using non-informative and informative priors with LFS application. Soumis à *Techniques d'enquête* (en cours de révision).

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 : Évaluation de l'estimation de l'intervalle de confiance sur petits domaines

Dans le cadre de ce projet, nous étudions et évaluons l'estimation de l'intervalle de confiance pour l'estimation sur petits domaines au moyen des meilleurs prédicteurs linéaires sans biais empiriques (MPLSBE). En particulier, nous étudions l'intervalle de confiance de Wald et l'intervalle de confiance de Wilson modifié pour l'estimation de la proportion sur petits domaines au moyen d'une application aux données de l'Enquête sur la population active et d'une étude par simulations.

Progrès :

Les résultats de l'analyse des données réelles et de l'étude par simulations montrent que l'intervalle de confiance de Wald et celui de Wilson modifié donnent tous deux de très bons résultats dans l'estimation sur petits domaines quand les MPLSBE sont utilisés. Le projet est terminé et nous avons rédigé un rapport de recherche interne (You et Hidiroglou, 2024).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Yong You (yong.you@statcan.gc.ca).

Bibliographie

You, Y., et Hidiroglou, M. (2024). Empirical study of confidence intervals for small area proportion estimation with LFS application. Rapport interne, Statistique Canada.

2 Méthodes et applications de la science des données

PROJET : Couplage d'enregistrements multipartite préservant la confidentialité

Le couplage d'enregistrements est un processus qui combine plusieurs sources de données tabulaires qui rendent compte, au moins partiellement, de certaines entités identiques. Ce processus exige parfois la collaboration de plusieurs parties, qui ont toutes des préoccupations concernant la protection de la vie privée des personnes figurant dans leurs données et qui ne sont pas nécessairement liées par un degré élevé de confiance. C'est pourquoi les méthodes permettant d'effectuer des couplages d'enregistrements tout en préservant la confidentialité des renseignements personnels contenus dans les données, en ne divulguant aucun renseignement au-delà des résultats des calculs requis, constituent un domaine de recherche précieux. Dans le cadre de ce projet, nous nous sommes penchés sur une étude de cas où un organisme statistique national voulait collaborer avec deux autres organismes pour calculer des agrégats tout en préservant la confidentialité.

Progrès :

Notre protocole tire profit des fonctions pseudo-aléatoires programmables et inconscientes, et des calculs fondés sur des circuits sécurisés. Ce travail prolonge le protocole à deux parties décrit dans (Dugdale, Santos et Zanussi, 2023) et repose sur l'intersection d'ensembles privés, où deux parties, chacune en possession d'un ensemble de données privées, visent à obtenir les éléments communs (ou identificateurs) aux deux ensembles de données, c'est-à-dire à en calculer l'intersection. Le couplage d'enregistrements préservant la confidentialité est lié à l'intersection d'ensembles privés en ce sens qu'il vise à calculer des agrégations sur des attributs appartenant aux identificateurs communs trouvés après l'intersection. Ces agrégats peuvent être assez simples, comme le calcul de la cardinalité de l'ensemble lié, ou plus complexes, comme les sommes pondérées fondées sur les attributs.

Nous avons commencé nos recherches sur les extensions multipartites des protocoles d'intersection d'ensembles privés que nous pourrions étendre pour gérer les agrégations. Nous avons découvert que le circuit d'intersection d'ensembles privés (Chandran, Gupta et Shah, 2022) permet le calcul sur l'intersection de deux ensembles au moyen du calcul sécurisé multipartites. De plus, on a étendu ce protocole pour traiter les scénarios où plusieurs parties possèdent un ensemble de données (Chandran, Dasgupta, Gupta, Obbattu, Sekar et Shah, 2021). Nous avons configuré et compilé la mise en œuvre dans un conteneur Docker afin de calculer l'intersection sur trois ensembles de données. Nous avons ensuite décrit un protocole utilisant un circuit d'intersection d'ensembles privés secondaire qui permet de calculer des agrégations sur les attributs de l'intersection (Santos, 2023). Nous avons constaté que l'extension multipartite du couplage d'enregistrements préservant la confidentialité est plus complexe et plus rigide, ce qui signifie que la solution doit être adaptée au problème de l'étude, à savoir la nature des ensembles de données et les agrégations souhaitées. Enfin, étant donné que ces protocoles cryptographiques

reposent sur des bases de codes provenant de différents groupes de recherche, il est difficile de vérifier, modifier, améliorer, déployer et documenter leurs mises en œuvre.

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

Bibliographie

Chandran, N., Dasgupta, N., Gupta, D., Obbattu, S.L.B., Sekar, S. et Shah, A. (2021). Efficient linear multiparty psi and extensions to circuit/quorum psi. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 1182-1204.

Chandran, N., Gupta, D. et Shah, A. (2022). Circuit-PSI with linear complexity via relaxed batch OPPRF. *Proceedings on Privacy Enhancing Technologies*.

Dugdale, C., Santos, B. et Zanussi, Z. (2023). Practical Privacy-Aware Data Linkage and Statistical Aggregation based on Privacy Enhancing Techniques. Conference on New Techniques and Technologies for Official Statistics, Bruxelles, Belgique.

Santos, B. (2023). Multi-Party Privacy Preserving Record Linkage based on Circuit Private Set Intersection. Canadian Mathematical Society Winter Meeting, Montréal, Canada.

PROJET : Génération de données tabulaires synthétiques à partir de modèles de diffusion

L'un des principaux objectifs d'un organisme national de statistique est de diffuser des données à l'intention du public et des chercheurs tout en respectant les exigences en matière de protection des renseignements personnels, de la confidentialité et de la sécurité (Stratégie des données de Statistique Canada, 2022). Habituellement, les organismes nationaux de statistique appliquent des méthodes de contrôle de la divulgation statistique (CDS) aux données afin de réduire les risques de divulgation de renseignements de nature délicate au moment de leur diffusion. La diffusion de données synthétiques a été proposée comme solution de rechange au CDS : dans ce cas, un modèle entraîné transforme les données originales en données synthétiques.

Les modèles de diffusion sont des candidats appropriés pour la génération de données tabulaires synthétiques (Kotelnikov, Baranchuk, Rubachev et Babenko, 2022). Il faut cependant évaluer le compromis entre confidentialité des données et utilité des données pour valider ces modèles. Nos objectifs étaient : i) de mettre en œuvre ou de modifier les modèles de diffusion existants pour les données tabulaires; ii) d'effectuer des mesures d'évaluation; iii) de traiter les aspects relatifs à la protection des renseignements personnels et à la confidentialité; iv) d'évaluer le compromis entre confidentialité des données et utilité des données.

Progrès :

Premièrement, nous avons commencé à étudier des modèles de diffusion pour les données tabulaires ainsi que d'autres méthodes, notamment les regroupements de fonctions disponibles pour la production de données synthétiques. Ensuite, nous avons élaboré un module de prétraitement pour transformer les données tabulaires et leur donner une forme convenant à l'entraînement des générateurs. Nous avons adapté la mise en œuvre de TabDDPM (Kotelnikov et coll., 2022), qui est un modèle de diffusion pour la

génération de données tabulaires synthétiques. Nous avons réalisé des progrès considérables dans la mise en œuvre d'un tableau de bord permettant de manipuler et prétraiter facilement les données tabulaires, aux fins d'entraînement des générateurs de données synthétiques et dans la production de mesures de la qualité et de la protection des renseignements personnels.

Pendant la phase d'expérimentation, nous avons comparé les performances des générateurs de réseaux antagonistes génératifs (GAN) (Figueira et Vaz, 2022) par rapport à TabDDPM. Nous avons constaté que la qualité des données synthétiques générées par le modèle de diffusion était supérieure pour presque toutes les mesures. De plus, TabDDPM obtenait de très bons résultats en matière de protection de la confidentialité, bien que le mécanisme d'ajout de la protection de la confidentialité ne soit pas entièrement compris. À cet égard, un GAN de confidentialité différentielle permet de mieux contrôler le compromis entre confidentialité et utilité. Toutefois, les progrès récents dans les modèles de diffusion à confidentialité différentielle ainsi que les modèles améliorés (Truda, 2023; Zhang, Zhang, Srinivasan, Shen, Qin, Faloutsos et Karypis, 2023) ouvrent la voie à une étude plus approfondie de l'utilisabilité de ces générateurs aux fins de production de données tabulaires synthétiques tout en préservant la confidentialité et l'utilité. Notre prochaine étape consistera à mettre en œuvre et évaluer des modèles de diffusion à confidentialité différentielle.

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

Bibliographie

Figueira, A., et Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, 10(15), 2733. MDPI AG. Consulté sur https://www.mdpi.com/2227-7390/10/15/2733.

Kotelnikov, A., Baranchuk, D., Rubachev, I. et Babenko, A. (2022). TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv preprint arXiv:2209.15421.

Stratégie des données de Statistique Canada (2022). Disponible à l'adresse : https://www.statcan.gc.ca/fr/apercu/strategiedonnees.

Truda, G. (2023). Generating tabular datasets under differential privacy. arXiv preprint arXiv:2308.14784.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C. et Karypis, G. (2023). Mixed-type tabular data synthesis with score-based diffusion in latent space. arXiv preprint arXiv:2310.09656.

PROJET : Chiffrement fonctionnel

Les systèmes cryptographiques sont la norme pour la protection des données en transit et au repos. En plus des systèmes, qui suppriment l'utilité du cryptogramme jusqu'à ce qu'il soit décrypté, d'autres systèmes permettent d'utiliser les données alors qu'elles sont chiffrées. Le chiffrement fonctionnel est un système cryptographique qui permet d'appliquer des fonctions à des ensembles de données chiffrés sans que les données soient déchiffrées. Le résultat de la fonction appliquée est déchiffré, ce qui permet aux utilisateurs autorisés d'obtenir les résultats de la fonction en texte clair appliquées aux données chiffrées sans compromettre la sécurité. Des clés individuelles peuvent être générées pour différentes fonctions, de sorte que seules les parties autorisées puissent utiliser et observer les résultats de fonction en texte clair. Le présent projet vise à étudier différentes utilisations du chiffrement fonctionnel du point de vue

d'un organisme national de statistique. Par exemple, plutôt que d'exiger que l'accès soit accordé et approuvé manuellement en cas d'accès total à des ensembles de données chiffrées de nature délicate, l'accès peut être accordé pour des sous-ensembles de données en fonction des privilèges et des attributs de l'utilisateur.

Progrès :

À la suite d'une enquête sur le chiffrement fonctionnel, des essais ont été réalisés et ont permis d'étudier l'application de ces techniques en pratique. Dans un premier temps, l'utilisation du chiffrement fonctionnel pour fournir des vues partielles d'un ensemble de données chiffrées a été étudiée au moyen d'un ensemble de données ouvertes sur les statistiques du revenu d'emploi selon le principal domaine d'études et le plus haut niveau de scolarité. Avec l'ensemble de données chiffré, on a défini sept expressions booléennes de sorte que les utilisateurs ayant la combinaison appropriée de rôles puissent afficher certaines combinaisons de colonnes de l'ensemble de données en texte clair, tandis que le reste demeure en cryptogramme. Ces règles reposent sur les attributs assignés à un utilisateur, par exemple un utilisateur privilégié, les utilisateurs généraux étant seulement autorisés à afficher les données chiffrées par défaut. Cette méthode permet de fournir un accès partiel en temps réel à des ensembles de données chiffrés sans qu'il soit nécessaire d'approuver ou de rejeter les demandes individuelles d'accès complet.

Dans une deuxième expérience, le chiffrement fonctionnel a été utilisé sur un ensemble de données fictif de données de localisation de correspondances de métro. Dans le cas présent, le déplacement d'un utilisateur est suivi et chiffré sur son appareil mobile. Plutôt que de demander à l'utilisateur d'envoyer ses données complètes, l'utilisateur peut fournir des clés permettant de récupérer seulement le total des visites de lieux et de l'utiliser pour générer une carte thermique. À titre d'exemple, pendant une crise de santé publique, cette méthode fournirait des cartes thermiques qui préservent la confidentialité des individus tout en déterminant les points névralgiques. Les deux expériences ont montré un fort potentiel d'utilisation lors d'un travail avec des données chiffrées.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Julian Templeton (julian.templeton@statcan.gc.ca).

PROJET : Apprentissage automatique aux fins d'imputation

L'imputation est une étape essentielle dans tous les programmes d'enquête de Statistique Canada. En tirant parti de la puissance des méthodes modernes d'apprentissage supervisé dans le processus d'imputation, on pourrait améliorer la qualité des données et des estimations qui en découlent. Nous étudions l'utilisation de l'apprentissage automatique aux fins d'imputation et comparons ses performances à celles des méthodes classiques, particulièrement dans le contexte de l'appariement selon la moyenne prédictive. L'appariement selon la moyenne prédictive emploie habituellement la régression linéaire comme modèle sous-jacent et utilise les valeurs prédites comme variable d'appariement unique pour l'imputation par donneur. Étant donné que l'appariement selon la moyenne prédictive ne s'appuie pas directement sur les prédictions du modèle de régression, la méthode est généralement considérée comme robuste à l'égard d'une spécification erronée du modèle. Dans le cas présent, nous remplaçons la régression linéaire par un modèle d'apprentissage automatique et nous évaluons les performances au moyen d'une étude par simulations.

Progrès :

À partir d'un sous-ensemble de variables du fichier de microdonnées à grande diffusion du Recensement de 2016 (Statistique Canada, 2023), nous construisons un ensemble de données de population complet, duquel nous tirons des échantillons pour notre étude par simulations. Dans chaque échantillon, nous générons une proportion fixe de valeurs manquantes pour une seule variable selon un mécanisme de données manquantes au hasard et nous utilisons les variables restantes pour ajuster les modèles d'imputation. Nous ajustons plusieurs modèles d'apprentissage automatique, en sélectionnant le modèle présentant la plus grande puissance de prédiction par validation croisée aux fins d'utilisation dans notre étude par simulations. Nous mesurons les performances des estimateurs de la moyenne et de la médiane de la population en examinant le biais, l'erreur-type empirique et la couverture.

Selon ce plan d'étude, nous constatons que l'utilisation de modèles d'apprentissage automatique avec l'appariement selon la moyenne prédictive entraîne des améliorations marginales des biais, de la couverture et de l'efficacité. Bien que modestes, ces améliorations mettent en évidence le potentiel des méthodes d'apprentissage automatique pour améliorer les processus d'imputation. De plus, les résultats démontrent l'efficacité de l'appariement selon la moyenne prédictive comme méthode d'imputation. D'autres études seraient nécessaires pour évaluer pleinement l'efficacité des modèles d'apprentissage automatique aux fins d'imputation. Nous recommandons de mener d'autres études par simulations, qui s'intéresseront aux différents mécanismes de données manquantes, aux niveaux de données manquantes et aux différents types de données.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Alden Chen (alden.chen@statcan.gc.ca).

Bibliographie

Statistique Canada (2023). Fichier de microdonnées à grande diffusion (FMGD) du Recensement de 2016, Fichier hiérarchique. Disponible à l'adresse : https://www150.statcan.gc.ca/n1/fr/catalogue/98M0002X.

PROJET : L'utilisation d'autoencodeurs masqués pour l'imputation généralisée de données tabulaires

La nécessité de traiter et de résoudre la question des données manquantes est omniprésente à Statistique Canada. Les données manquantes peuvent prendre de nombreuses formes, comme la non-réponse aux enquêtes, et elles peuvent nuire à l'exactitude et à la fiabilité des inférences statistiques, ce qui nécessite souvent l'utilisation de techniques d'imputation et de démarches heuristiques. Les méthodes d'imputation peuvent présenter l'inconvénient de nécessiter des connaissances nombreuses sur le domaine ou de manquer de fidélité.

Un ensemble de paradigmes d'apprentissage automatique de pointe (langage masqué/modélisation d'images) ajuste des modèles pour reconstruire une portion (manquante) masquée de l'entrée échantillonnée aléatoirement, ce qui donne des modèles capables d'effectuer des tâches en aval (texte/image) arbitraires. Les autoencodeurs masqués forment une de ces méthodes. En partant de la constatation que ces paradigmes optimisent simplement l'imputation, ce projet applique des autoencodeurs masqués aux ensembles de données tabulaires, afin de produire un modèle d'imputation généralisée qui, une fois ajusté au moyen de ce paradigme, peut effectuer une imputation de très grande qualité sur des caractéristiques manquantes arbitraires d'un ensemble de données tabulaires donné. La

méthode est simple et de nature générale, et elle ne nécessite aucune connaissance du sujet de l'ensemble de données.

Progrès :

On a utilisé une architecture axée sur des transformeurs pour mettre en œuvre les modèles d'imputation par autoencodeurs masqués. Des essais approfondis par rapport à 15 bases de référence d'imputation établies, allant des techniques classiques à l'apprentissage automatique et aux méthodes fondées sur l'apprentissage profond, dans 25 ensembles de données tabulaires variés.

Les résultats ont constamment montré que les autoencodeurs masqués étaient la technique d'imputation la plus puissante comparativement aux 15 bases de référence testées. Dans des essais par rapport à des ensembles de données dont 30 % des valeurs étaient manquantes (échantillonnées indépendamment et aléatoirement), notre méthode a obtenu le classement le plus élevé pour ce qui est de l'exactitude des caractéristiques catégoriques, l'erreur absolue moyenne des caractéristiques numériques, et la distance de Wasserstein pour les caractéristiques catégoriques. Elle est arrivée deuxième au classement de la distance de Wasserstein pour les caractéristiques numériques. Parmi les techniques d'imputation mises à l'essai, les autoencodeurs masqués ont toujours été les plus robustes pour ce qui est de la quantité de données manquantes : bien que notre méthode s'adapte à des modèles utilisant 30 % de données manquantes, les autoencodeurs ont montré de bonnes performances quand ils sont évalués sur des ensembles présentant jusqu'à 50 % de valeurs manquantes et ont été constamment plus robustes que les autres bases de référence quand ils sont évalués sur des ensembles présentant jusqu'à 90 % de valeurs manquantes. Les résultats démontrent clairement que les autoencodeurs masqués pour l'imputation tabulaire généralisée sont une technique d'imputation viable, applicable aux données manquantes pour toutes les variables d'un ensemble de données tabulaires particulier et ne nécessitant pas de connaissances sur le sujet.

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

PROJET : Optimisation du temps et des efforts dans le traitement des données

Le projet de recherche sur l'optimisation du temps et des efforts est une recherche approfondie visant à améliorer l'efficacité du prétraitement des microdonnées dans les données d'enquête à Statistique Canada. Le projet s'intéresse à la robustesse du prétraitement des données pour les tâches complexes et les performances d'une méthode d'imputation dans différentes conditions. Cette recherche vise à optimiser la rapidité d'exécution et à réduire les efforts manuels dans l'Enquête mensuelle sur l'approvisionnement et l'écoulement de l'électricité.

Progrès :

Pendant la phase initiale de la recherche, nous avons évalué les modèles d'imputation pour la production mensuelle d'électricité du projet précédent à différents moments. Dans la deuxième phase, on a procédé à l'analyse détaillée des processus du Programme intégré de la statistique des entreprises de l'Enquête mensuelle sur l'approvisionnement et l'écoulement de l'électricité, en mettant l'accent sur le rôle crucial que jouent les rajustements manuels dans l'amélioration de la qualité des données. On a ensuite élaboré une méthodologie utilisant le test de Dickey-Fuller augmenté pour sélectionner des séries de données stables aux fins d'une imputation simple. La méthode évalue aussi si la modélisation de ces unités à

puissance prédictive élevée améliore la précision globale de l'estimation. La méthode simple de l'étude a montré des résultats prometteurs pour ce qui est de l'Enquête mensuelle sur l'approvisionnement et l'écoulement de l'électricité, qui a obtenu un faible pourcentage d'erreur absolue moyenne pour l'estimation, ainsi qu'une réduction considérable du volume de corrections manuelles appliquées. Les étapes suivantes comprennent l'application des connaissances tirées du projet à la nouvelle initiative d'accélération méthodologique, qui vise à élaborer des outils de surveillance et de réduction des opérations de validation manuelle des données. Ce projet pourrait réduire considérablement les coûts et les délais de validation, augmentant ainsi l'efficacité des programmes de statistiques économiques.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Soufiane Fadel (soufiane.fadel@statcan.gc.ca).

3 Problèmes d'estimation dans les enquêtes

PROJET : Un estimateur en cas d'utilisation concomitante d'un effort de collecte complet et réduit sur des sous-échantillons aléatoires

Cette recherche est motivée par deux tendances importantes des dernières années, constatées à Statistique Canada et dans d'autres organismes statistiques. Premièrement, les taux de réponse aux enquêtes sociales continuent de diminuer. Deuxièmement, l'utilisation de l'autodéclaration en ligne comme mode de collecte est en hausse, en raison de sa facilité et de son coût relativement bas. Pour s'adapter à ces défis, ce projet construit un estimateur (appelé estimateur multimodal concomitant) destiné aux activités de collecte qui utilisent un suivi des cas de non-réponse moins coûteux (par exemple par la poste et par courriel) pour l'ensemble de l'échantillon en même temps qu'un suivi des cas de non-réponse plus coûteux par intervieweurs pour un sous-ensemble de non-répondants. L'utilisation concomitante de méthodes de contact moins coûteuses et de méthodes par intervieweurs rend ce projet différent des travaux antérieurs dans le cadre desquels le suivi des cas de non-réponse était effectué après une première étape d'autodéclaration (par exemple Hansen et Hurwitz, 1946 et Neusy, Beaumont, Yung, Hidiroglou et Haziza, 2022) et de l'expérience de l'Enquête nationale auprès des ménages, dans laquelle la fraction d'échantillonnage initiale était de grande taille (Beaumont, Bocci et Hidiroglou, 2014).

Progrès :

Les travaux réalisés en 2023-2024 comprenaient l'élaboration de la théorie, une étude par simulations et une application expérimentale de la méthode à l'Enquête sociale canadienne. Les travaux théoriques ont démontré que l'estimateur multimodal concomitant est sans biais avec spécification correcte du modèle pour les propensions à répondre incluses dans l'estimateur et que l'estimateur multimodal concomitant peut tirer parti des variables d'enquête des répondants pour une certaine modélisation, ce qui permet des modèles plus riches que ceux disponibles pour de nombreux ajustements pour la non-réponse aux enquêtes auprès des ménages. De plus, les simulations ont montré que, pour certaines variables semblables à celles des enquêtes auprès des ménages de Statistique Canada, la variance est plus petite et la taille effective de l'échantillon plus élevée pour l'estimateur multimodal concomitant que pour d'autres estimateurs ayant un biais comparativement faible. Les résultats donnent à penser que ces méthodes pourraient servir à réaliser des économies substantielles sans compromettre indûment la qualité des estimations.

Les résultats du projet ont été présentés au Comité consultatif sur les méthodes statistiques en mai 2024 (Mather, Boulet et Brennan, 2024) et au Congrès annuel de la Société statistique du Canada en juin 2024.

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

Bibliographie

Beaumont, J.-F., Bocci, C. et Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Document présenté au Comité consultative sur les méthodes statistiques, 58, Statistique Canada.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Mather, A., Boulet, C. et Brennan, A. (2024). An Estimator for Concurrent Use of Full and Reduced Collection Effort on Random Subsamples. Document présenté au Comité consultative sur les méthodes statistiques, 78, Statistique Canada.

Neusy, E., Beaumont, J.-F., Yung, W., Hidiroglou, M. et Haziza, D. (2022). Suivi de la non-réponse aux enquêtes auprès des entreprises. *Techniques d'enquête*, 48, 1, 103-128. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022001/article/00006-fra.pdf.

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

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

Les procédures d'élagage exigent la spécification d'une constante de réglage, où les poids supérieurs à la constante sont réduits à la valeur de la constante et le poids excédentaire est redistribué entre les unités restantes. Souvent, les procédures utilisées en pratique sont ad hoc; il manque en effet à l'heure actuelle une méthode formelle et théorique pour traiter l'élagage de poids. Ce projet a les objectifs suivants : 1) pour les procédures propres aux variables, élaborer certains résultats théoriques semblables à ceux de Ma et Wang (2020) sur le choix de seuil optimal; 2) élaborer certaines procédures d'élagage des poids pour traiter plusieurs variables d'enquête; 3) effectuer une vaste étude par simulations pour comparer les performances de plusieurs procédures d'élagage des poids (voir Potter [1988, 1990], Potter et Zheng [2015], Haziza et Beaumont [2017]) pour ce qui est du biais et de l'efficacité et pour évaluer les performances des estimateurs de l'erreur quadratique moyenne pour ce qui est du biais; puis 4) appliquer et démontrer la méthodologie proposée sur un ensemble de données existant.

Progrès :

On a élaboré un grand programme de simulation en utilisant un mélange de R et de SAS, exploitant le système généralisé de calage de Statistique Canada, G-EST. La simulation comprend plus de dix procédures d'élagage et les compare en termes de biais et d'efficacité quand elles sont appliquées à plusieurs variables d'enquête présentant des niveaux variables de corrélation avec les poids. Les résultats préliminaires montrent que l'élagage, quand il est appliqué à des poids ayant peu ou pas de corrélation avec les variables de l'enquête, améliore l'efficacité sans compromettre le biais, comme on s'y attendait. À mesure que la corrélation entre les poids et les variables de l'enquête augmente, les estimateurs obtenus ont tendance à présenter un biais appréciable qui, à son tour, peut mener à une erreur quadratique moyenne plus grande que celle de l'estimateur sans élagage.

Les prochaines étapes immédiates du projet consistent à poursuivre l'étude par simulations en ajoutant des méthodes d'élagage et en intégrant des variables d'enquête catégoriques. Il nous reste également à traiter les autres objectifs.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Jody Krahn (jody.krahn@statcan.gc.ca) ou
David Haziza (dhaziza@uottawa.ca).

Bibliographie

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

Ma, X., et Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115, 532, 1851-1860.

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

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

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

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

Statistique Canada publie maintenant des intervalles de confiance afin d'exprimer la qualité des estimations. La longueur d'un intervalle de confiance peut témoigner de la qualité de l'estimation en autant que la couverture annoncée est respectée. Un paramètre qui joue un rôle important dans le calcul des intervalles de confiance est le nombre de degrés de liberté. En pratique, cette valeur est généralement déterminée à l'aide d'une règle approximative ou règle du pouce (en anglais, « rule of thumb »). Dans le cas de petits domaines, cette règle du pouce surestime souvent le nombre de degrés de liberté réel ce qui engendre une sous-couverture des intervalles de confiance.

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

L'objectif de ce projet pour l'année 2023-2024 était de finaliser l'écriture et la révision à l'interne de l'article de Toupin et Martin (2024) dans le but de soumettre celui-ci à une revue scientifique.

Progrès :

L'article de Toupin et Martin (2024) a été soumis à la revue *Techniques d'enquête*. Celui-ci est en cours de révision.

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

Bibliographie

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

Toupin, M.-H., et Martin, V. (2024). 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 : Exactitude de prédictions par apprentissage automatique

Afin de réduire les coûts et d'améliorer l'actualité des données, Statistique Canada étudie actuellement une méthode consistant à remplacer périodiquement la collecte de données d'enquête par des prédictions. Ces prédictions seraient obtenues au moyen de données auxiliaires actuelles et de modèles déterminés à partir de données historiques. L'estimation de paramètres de population finie, comme des totaux de population, serait ensuite réalisée par l'agrégation de ces prédictions individuelles. Par exemple, on peut prédire les rendements des cultures par télédétection, au moyen de variables agroclimatiques et des données historiques, y compris les réponses antérieures (Statistique Canada, 2020; National Academies of Sciences, Engineering, and Medicine, 2023, Chapter 8.3). De plus en plus, les méthodes d'apprentissage automatique facilitent ces prédictions (Chu, 2022), qui doivent comprendre des mesures d'exactitude. Cependant, les techniques classiques de validation, comme la validation croisée, sont inadéquates parce qu'elles mettent l'accent sur l'incertitude de la prédiction totale plutôt que sur les prédictions individuelles (Hastie, Tibshirani et Friedman, 2001, Chapter 7), conditionnellement aux covariables.

Progrès :

Pour répondre au besoin décrit ci-dessus, on a élaboré et évalué une méthodologie bootstrap au moyen de simulations préliminaires (Dasylva, Beaumont, Bosa et Maranda, 2023), en examinant différents modèles pour la réponse moyenne conditionnelle dans la population et pour l'estimation de cette moyenne conditionnelle dans la procédure bootstrap. Ces simulations ont montré qu'on peut estimer l'erreur quadratique moyenne avec un biais relatif important quand le modèle bootstrap diffère grandement du modèle de population.

D'autres simulations ont été réalisées : elles ont révélé que le biais relatif de l'estimateur de l'erreur quadratique moyenne est une fonction croissante du coefficient de détermination pour un modèle de population linéaire. Par conséquent, il a tendance à être faible quand le coefficient de détermination est modéré ou faible, comme on s'y attend en pratique, alors qu'il est important quand le coefficient de détermination est élevé. D'autres simulations seront effectuées dans la prochaine année pour mieux comprendre les propriétés de notre estimateur de l'erreur quadratique moyenne bootstrap.

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

Bibliographie

Chu, K. (2022). Utilisation de l'apprentissage automatique pour prédire le rendement des cultures. Statistique Canada. Disponible à l'adresse : https://www.statcan.gc.ca/fr/science-donnees/reseau/rendement-cultures.

Dasylva, A., Beaumont, J.-F., Bosa, K. et Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, mai 2023.

Hastie, T., Tibshirani, R. et Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 2st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. Disponible à l'adresse : https://doi.org/10.17226/26804.

Statistique Canada (2020). Modélisation intégrée du rendement des cultures au moyen de la télédétection, de données agroclimatiques et de données d'enquête. Disponible à l'adresse : https://www.statcan.gc.ca/fr/programmes-statistiques/document/5225_D1_T9_V1.

PROJET : Estimation bootstrap de la variance pour estimateurs par calage

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

de l'estimateur de la variance de l'estimateur RG pour d'autres estimateurs par calage; c'est l'estimateur de la variance généralement mis en œuvre en pratique.

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

Progrès :

Nous avons élaboré un estimateur de la variance bootstrap qui tient correctement compte des contraintes de borne. Nous avons réalisé une étude par simulations dans laquelle nous avons comparé les estimations de Horvitz-Thompson, les estimations RG et les estimations de calage ridge ainsi que leurs estimations de variance. Dans nos simulations, nous avons montré l'importance d'utiliser des estimations d'enquête pré-calibrées (par exemple les estimations de Horvitz-Thompson) plutôt que des totaux de population connus comme totaux de contrôle bootstrap.

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 : Population synthétique pour les simulations du remaniement de l'Enquête sur la population active

La méthodologie de l'Enquête sur la population active (EPA) du Canada fait l'objet d'un examen minutieux tous les dix ans. Aux fins de l'examen actuel, nous avons élaboré un système de simulation qui reproduit le processus d'enquête complexe de l'EPA – de l'échantillonnage à l'estimation – pour envisager d'autres méthodes ou déterminer les éléments à améliorer. Le système tire des échantillons d'une population synthétique conçue pour être une approximation raisonnable de la population canadienne en termes de démographie, de géographie, de structure des ménages et de renseignements de base sur le travail, et pour préserver des transitions réalistes de la main-d'œuvre au niveau des personnes pendant une période de six mois (durée pendant laquelle une unité reste dans l'échantillon dans le plan de l'EPA à panel rotatif). Nous y sommes parvenus en construisant une « population à panel rotatif », modélisée sur une période de six ans au moyen d'une combinaison de techniques de modélisation transversale et longitudinale. Un panel de population commence par un ensemble de « clones » d'une population de base (les répondants au questionnaire détaillé du recensement de mai 2016, passés par le modèle transversal), puis chaque unité « clone » est modélisée de façon longitudinale pour une période de six mois avant d'être réinitialisée. La population synthétique complète comprend six panels décalés en parallèle. Cette structure produit la série de six mois requise pour chaque répondant simulé de l'EPA, tout en atténuant la dérive qui se produit en cas de projection d'une population sur une période prolongée.

Progrès :

Nous avons terminé la composante longitudinale : Nous avions déjà décidé de la catégorie de modèle et de l'ensemble des prédicteurs potentiels. En 2023, nous avons produit une série de modèles candidats et les avons évalués par validation croisée et au moyen d'autres mesures afin de sélectionner notre modèle final.

Nous avons terminé la composante transversale : Nous avions déjà élaboré un plan visant à diviser la population de base en catégories (de façon similaire à la post-stratification) et à rajuster au minimum les renseignements sur le travail des personnes pour qu'ils correspondent aux répartitions de l'EPA pour ce mois dans chaque catégorie. En 2023, nous avons programmé un algorithme d'arborescence de classification adapté afin que les décisions de fractionnement puissent être contraintes par des conditions sur un ensemble de données externe. Au moyen de cette fonction, nous avons établi des conditions de fractionnement pour la population de base qui garantissaient que l'échantillon mensuel de l'EPA pour chaque catégorie serait suffisamment grand pour produire des estimations raisonnables.

Nous avons produit les données. Nous avons élaboré une stratégie pour produire d'abord les données du mois initial dans le modèle transversal, puis pour projeter la série chronologique de six mois dans le modèle longitudinal, en maintenant la cohérence entre l'emploi, l'industrie et la catégorie de travailleurs. Nous avons produit des fichiers de données de population synthétique mensuels pour la période de 2013 à 2018.

Nous avons utilisé la population synthétique pour effectuer des simulations afin d'évaluer le plan d'enquête et les méthodes d'estimation. Les résultats ont joué un rôle important dans la prise de décisions concernant la plupart des aspects du remaniement.

Davantage de détails se trouvent dans Brennan et Summers (2023a, 2023b).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Andrew Brennan (andrew.brennan@statcan.gc.ca) ou
Pauline Summers (pauline.summers@statcan.gc.ca).

Bibliographie

Brennan, A., et Summers, P. (2023a). Synthetic population modelling details. Rapport interne, Statistique Canada.

Brennan, A., et Summers, P. (2023b). Synthetic population overview. Rapport interne, Statistique Canada.

4 Confidentialité et accès aux données

Les travaux de recherche sur la confidentialité à Statistique Canada continuent de porter sur l'élaboration de nouvelles méthodes et idées qui offrent d'autres formes d'accès tout en continuant à garantir que les renseignements personnels des particuliers et des entreprises ne sont divulgués d'aucune façon. Des progrès ont été réalisés dans le cadre des projets décrits ci-dessous. L'équipe responsable du Centre de la confidentialité et de l'accès de Statistique Canada a également continué d'offrir des services de

consultation aux partenaires internes et externes afin d'aider à renforcer la capacité de détection et de traitement des risques de divulgation (voir la section 5.5).

PROJET : Évaluation de la confidentialité pour des estimations sur petits domaines

À l'heure actuelle, Statistique Canada ne dispose d'aucune directive officielle concernant les règles de confidentialité aux fins de diffusion des estimations sur petits domaines et aucune étude officielle n'a encore été menée à ce sujet. Depuis quelques années, les chercheurs des centres de données de recherche (CDR) demandent de plus en plus des lignes directrices détaillées sur la confidentialité afin de pouvoir publier en toute sécurité des estimations sur petits domaines. La présente analyse de la confidentialité a été appliquée à l'estimation sur petits domaines au niveau du domaine et repose sur des simulations dans R pour examiner l'effet de différentes tailles d'échantillon et de divers niveaux d'erreur du modèle sur le risque de divulgation.

Progrès :

On a créé des populations simulées dans lesquelles les échantillons sont sélectionnés. Les populations simulées contenaient une variable auxiliaire, une variable d'intérêt et de l'information sur le domaine. La force de la relation entre la variable auxiliaire et la variable d'intérêt a été contrôlée au moyen d'une variable « d'erreur » avec une composante aléatoire. Des échantillons aléatoires stratifiés ont été tirés, et des estimations sur petits domaines au niveau du domaine ont été calculées au moyen du regroupement de fonctions *sae* de R (Molina et Marhuenda, 2015). On a comparé le risque de divulgation des estimations sur petits domaines aux estimations directes de Horvitz-Thompson pour démontrer que les estimations sur petits domaines sont intrinsèquement moins risquées que les estimations directes, particulièrement quand les taux d'échantillonnage sont extrêmement bas. Nous avons ensuite analysé les résultats, puis proposé des lignes directrices exhaustives concernant la confidentialité aux fins de diffusion d'estimations sur petits domaines au niveau du domaine.

Nous avons presque terminé de rédiger un article qui décrit le processus de simulation et explique les justifications des règles de confidentialité proposées. Un résumé pour le Symposium international de 2024 sur les questions de méthodologie a été soumis et est en attente d'approbation.

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

Bibliographie

Molina, I., et Marhuenda, Y. (2015). sae: An R Package for small area estimation. *The R Journal*, 7, 81-98.

PROJET : Données synthétiques

Il est essentiel de chercher à offrir plus d'options aux utilisateurs de données. La création de données synthétiques constitue un moyen de traiter les problèmes de confidentialité des données personnelles tout en conservant la plus grande valeur analytique possible. Les données synthétiques peuvent être particulièrement utiles quand il s'agit de collaborer avec des intervenants externes qui n'ont pas nécessairement accès aux microdonnées confidentielles.

Progrès :

La base de données synthétiques pour le modèle de microsimulation dynamique PASSAGES est terminée. Les données, accompagnées du modèle, ont été diffusées le 23 avril 2024 (Statistique Canada, 2024). La population de départ synthétique représente la population canadienne au 31 décembre 2015. Les antécédents familiaux et de revenu de ces acteurs individuels remontent à 1966. La population de départ synthétique a été créée principalement au moyen de méthodes d'apprentissage automatique qui intégraient des données de recensement, des données fiscales et d'autres sources de données administratives. Afin de créer cette base de données synthétique, on a eu recours à une combinaison de techniques de manière à tenir compte de sa dimension hiérarchique (personne organisée en unités familiales) et de sa dimension temporelle (variables relatives à la trajectoire de vie ayant des corrélations longitudinales complexes).

Le centre est en train de mettre à jour des lignes directrices internes aux fins de création de fichiers de données synthétiques et de terminer l'examen de l'évaluation des risques de divulgation pour les données synthétiques.

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

Bibliographie

Statistique Canada (2024). *Le Quotidien* – Nouveau modèle de microsimulation de revenu de retraite maintenant offert. Disponible à l'adresse : https://www150.statcan.gc.ca/n1/daily-quotidien/240423/dq240423c-fra.htm.

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

La suppression de cellules complémentaires est une méthode standard de suppression de cellules sensibles confidentielles lors de la diffusion de variables de données quantitatives tabulaires. Cette méthodologie est bien élaborée et appuyée par la solution G-Confid de Statistique Canada, par laquelle on obtient des solutions de suppression optimales qui garantissent que les modèles de suppression sont valides et réduisent le plus possible la quantité de renseignements supprimés.

Statistique Canada s'assure que ses solutions logicielles restent diversifiées et cherche des solutions autres que les solutions standards qui reposent sur le système d'analyse statistique (SAS) utilisé jusqu'à maintenant. Le retrait de G-Confid de SAS représente un défi, car il faut trouver des solutions compatibles avec le solveur OPTMODEL de SAS. Les solutions « libres » disponibles par l'intermédiaire du progiciel Python PuLP ont été étudiées comme solutions de remplacement possibles. Les résultats ont été analysés et des résultats préliminaires ont été présentés aux Joint Statistical Meetings (JSM) en 2023 (Chen et Thomas, 2023).

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

Bibliographie

Chen, H., et Thomas, S. (2023). Assessing the Performance of the Open-Source Linear Programming Solver in Cell Suppression Problems. Dans *Proceedings of the Survey Research Methods Section,* American Statistical Association. Disponible à l'adresse : https://doi.org/10.5281/zenodo.10359791.

5 Soutien (centres de ressources)

5.1 Centre de recherche et d'analyse en séries chronologiques

L'objectif du Centre de recherche et d'analyse en séries chronologiques est de maintenir une expertise de haut niveau et d'offrir des services de consultation en matière de séries chronologiques dans l'ensemble de l'organisme. Le centre offre des services de consultation et des conseils sur les problèmes relatifs aux séries chronologiques, il étudie les problèmes pour lesquels il n'existe pas actuellement de solutions connues ou satisfaisantes, et il élabore et tient à jour des outils permettant d'appliquer des solutions à des problèmes réels de séries chronologiques.

Les projets peuvent être répartis en quatre sous-catégories, l'accent étant mis sur les thèmes suivants :

- consultation et formation sur les séries chronologiques;
- soutien et amélioration du système et des outils de traitement des séries chronologiques;
- modélisation et prévision des séries chronologiques;
- soutien méthodologique aux programmes d'indice des prix à la consommation et à la production.

Progrès :

Consultation et formation sur les séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques est chargé d'élaborer et de fournir une formation sur les méthodes en séries chronologiques, comprenant la désaisonnalisation, l'étalonnage, le rapprochement et la modélisation en séries chronologiques, à l'intention des participants de Statistique Canada et aussi d'autres organismes. De plus, le centre offre des conseils et des services de consultations sur des projets de séries chronologiques en général pour les programmes de l'ensemble de Statistique Canada.

Au cours de l'année, le centre a proposé des cours sur les composantes des séries chronologiques, la désaisonnalisation, le rapprochement, la modélisation et les prévisions aux participants internes et externes par l'entremise du Centre de formation de Statistique Canada (Statistique Canada, 2024). Dans le cadre de la formation pour les nouvelles recrues (série de séminaires de la Direction de la méthodologie pour les recrues et cours de navigation des données), les membres du centre ont également participé à des activités de sensibilisation et de formation auprès d'autres groupes de Statistique Canada sur des thèmes touchant les séries chronologiques et ils ont donné un cours d'introduction à R.

Le centre a également offert des services de consultation à plusieurs programmes internes (désaisonnalisation, modélisation de séries chronologiques, extrapolation rétrospective, prévisions immédiates, prévisions, estimation des tendances, calendarisation, etc.). En particulier, le centre a fourni un soutien en matière de séries chronologiques au Système de comptabilité nationale dans un certain

nombre de domaines, notamment le produit intérieur brut mensuel, les transports et l'investissement en logement. En outre, des représentants du centre assistent périodiquement à un forum hebdomadaire d'analystes pour maintenir une présence au sein de la communauté des analystes. Le centre mène régulièrement des consultations sur l'extrapolation rétrospective afin de préserver ou de rétablir la comparabilité au fil du temps, et il a travaillé à la production de lignes directrices sur la continuité des séries chronologiques pour les programmes de Statistique Canada. Il s'agit d'une initiative conjointe avec le Système de comptabilité nationale, dont les récents travaux ont consisté à discuter des lignes directrices proposées avec notre Bureau de gestion des projets de l'organisme afin de déterminer la manière dont les lignes directrices pourraient être intégrées au cadre de gestion des projets.

De plus, afin de soutenir divers programmes internes, le centre a mené des consultations et des échanges avec de nombreux organismes publics fédéraux et provinciaux, ainsi qu'avec des organismes nationaux de statistique (Banque du Canada, Institut de la statistique du Québec, Bureau de la statistique de l'Australie, US Census Bureau, BMO, BC Stats) sur des thèmes liés aux séries chronologiques (stratégie de désaisonnalisation durant la pandémie, extrapolation rétrospective, déflation, outils logiciels, etc.).

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 des séries chronologiques pour les programmes de Statistique Canada qui produisent des données désaisonnalisées, notamment : le système généralisé G-Séries, pour l'étalonnage et le ratissage, le rapprochement et l'équilibrage (Statistique Canada, 2016); le Système de traitement des séries chronologiques (Ferland, 2022); le tableau de bord de la désaisonnalisation (Verret, 2021).

Les travaux sur une version prototype de G-Séries dans R se sont poursuivis. Les fonctionnalités d'étalonnage et de ratissage ont été mises en œuvre, développées en regroupements et hébergées dans le GitLab interne de Statistique Canada à usage interne (regroupement de fonctions *rgseriespt*). La fonction d'équilibrage sera terminée d'ici l'automne 2024. Une fois le prototype terminé, il sera examiné en vue de sa diffusion officielle (à l'interne et à l'externe).

Le système de traitement des séries chronologiques est une application personnalisable qui repose sur SAS et qui permet d'appliquer des techniques de séries chronologiques, dont des techniques de désaisonnalisation, d'étalonnage et de rapprochement, largement utilisées dans la production d'estimations désaisonnalisées pour les sous-programmes annuels de Statistique Canada (essentiels à la mission 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 d'aborder les nouveaux besoins des programmes de l'organisme. Nous étudions une nouvelle version du système qui permettrait à plus long terme une plus grande flexibilité pour intégrer les outils et les nouvelles techniques disponibles à partir de logiciels libres, en particulier ceux de désaisonnalisation (regroupement de fonctions R fournissant une interface au logiciel X-13ARIMA-SEATS de l'US Census Bureau), *PyX13* (US Census Bureau, version bêta), *JDemetra+/RJDemetra* (Eurostat) et *rgseriespt* (regroupement de fonctions R pour G-Séries). Des discussions concernant les systèmes de traitement des séries chronologiques ont eu lieu avec l'Institut national de la statistique et des études économiques (France), l'US Census Bureau et le Bureau de la statistique de l'Australie.

Un certain nombre d'améliorations ont été apportées au tableau de bord sur la désaisonnalisation cette année. En particulier, de nouvelles fonctionnalités visant à augmenter l'efficacité ont été ajoutées, la

compatibilité avec différentes versions de R a été ajoutée et des bogues mineurs ont été résolus. Le tableau de bord est en voie d'être déployé pour un programme supplémentaire sur la main-d'œuvre. Une documentation complète a été produite.

Modélisation et prévision des séries chronologiques

L'augmentation de l'actualité des indicateurs statistiques est une priorité importante pour Statistique Canada, et l'une des options pour y arriver est de modéliser les séries chronologiques pour établir des prévisions immédiates des indicateurs économiques beaucoup plus tôt que le moment où le premier estimateur traditionnel est produit. Des travaux sur un projet de prévisions immédiates pour élaborer une méthode plus précise d'estimation des dépenses liées aux activités de rénovation (dans le cadre du programme mensuel d'investissement en construction de bâtiments, un indicateur économique clé qui quantifie l'état de l'investissement en construction de bâtiments dans l'économie) ont été présentés au Comité d'examen scientifique de la Direction des méthodes statistiques modernes et de la science des données (Patak et Plunkett, 2023). Des travaux sur la production d'indicateurs avancés concernant les statistiques de l'énergie et d'un cadre général pour les indicateurs avancés de la qualité ont été présentés à la 76^{e} réunion du Comité consultatif sur les méthodes statistiques de Statistique Canada (Le Moullec et Matthews, 2023; Matthews, 2022). Une séance sur invitation sur le thème de l'estimation en temps réel a été organisée au Congrès mondial de la statistique de l'Institut international de la statistique en juillet 2023.

Le Centre a poursuivi son étude des moyens de produire des indications précoces de ruptures structurelles au moyen de modèles espace-état.

Soutien méthodologique aux programmes d'indice des prix à la consommation et à la production.

Le Centre de recherche et d'analyse des séries chronologiques possède aussi une sous-section chargée de fournir un soutien méthodologique aux programmes d'indice des prix à la consommation et à la production (IPC et IPP).

Au cours des dernières années, l'indice des prix à la consommation a fait l'objet de nombreux changements concernant la collecte et la méthodologie dans le but de moderniser les pratiques et de s'adapter à l'évolution des conditions depuis la pandémie de COVID-19. La collecte classique en personne a été remplacée par une combinaison de collecte en ligne et de données d'autres sources, notamment des données de lecteurs optiques aux points de vente, des bases de données administratives, le moissonnage du Web et d'autres enquêtes. Les pondérations du panier proviennent maintenant de la série sur les dépenses de consommation finale des ménages plutôt que de l'Enquête sur les dépenses des ménages, afin de permettre des mises à jour plus fréquentes et actuelles du panier. Des couplages d'enregistrements ont créé une base de sondage partielle de l'IPC liant les entreprises aux ventes au niveau des produits pour certains agrégats de l'IPC. Le centre a examiné la méthodologie d'échantillonnage de l'IPC dans le contexte de ces changements. Plusieurs améliorations possibles ont été proposées, notamment la mise à jour de la répartition d'échantillons, le renouvellement de l'échantillon, l'échantillonnage probabiliste pour certains agrégats, et l'utilisation d'indicateurs de qualité adaptés à des caractéristiques uniques de l'IPC (Francis, 2023). Des méthodes fondées sur le plan et fondées sur un modèle pour l'estimation de la variance ont été étudiées. Le centre a également élaboré une stratégie et un outil dans R pouvant servir à évaluer l'échantillon des différentes composantes de l'IPC. La méthode a

été appliquée à la composante habillement et chaussures cette année et des améliorations ont été proposées.

De nombreux indices des prix à la production pour les services (IPPS) utilisent l'échantillonnage probabiliste des entreprises au premier degré. La plupart utilisent un échantillonnage de Poisson séquentiel avec probabilité proportionnelle à la taille (Ohlsson, 1998), stratifié par groupes d'industries. Les plans d'échantillonnage précédents attribuaient l'échantillon aux groupes d'industrie par des méthodes plus simples, comme la répartition proportionnelle au revenu ou à la taille, en l'absence de bonnes estimations de la variance. Chaque échantillon de l'IPPS a été conçu de façon indépendante avec des budgets, des tailles d'échantillon et une précision très variables. Ce projet a permis d'étudier l'optimisation de l'utilisation des ressources afin d'assurer une qualité uniforme dans l'ensemble des IPPS et de conserver le budget dans la mesure du possible. On a utilisé le bootstrap généralisé de Beaumont-Patak (Beaumont et Patak, 2012) et la linéarisation en séries de Taylor pour estimer les variances dues au premier degré pour les indices de prix de Lowe avec des échantillons de Poisson séquentiels avec probabilité proportionnelle à la taille. Des ajustements ont été apportés pour tenir compte de caractéristiques propres aux indices de prix, à savoir la non-réponse périodique, l'imputation parentale, la pondération mise à jour en fonction des prix, l'enchaînement et la non-unicité (Francis, 2024). Des fonctions R ont été développées. À partir de ces estimations de la variance, on a estimé des tailles d'échantillon pour atteindre les cibles de coefficient de variation pour les niveaux les plus bas publiés. Les résultats ont été mis à l'essai pour une grande enquête (Indice des prix des services du commerce de gros) ainsi qu'une petite enquête (Indice des prix des services d'architecture, de génie et de services connexes) et présentés à la Division des prix à la production.

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

Bibliographie

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80, 127-148.

Ferland, M. (2022). *Time Series Processing System – v3.08*. Document interne, Statistique Canada.

Francis, J. (2023). *Canadian CPI Sample Design Methodology Review*. Document interne, Statistique Canada.

Francis, J. (2024). *Design-Based Estimates and Variance Estimation for SPPIs with Sequential Poisson PPS Samples*. Document interne, Statistique Canada.

Le Moullec, J., et Matthews, S. (2023). On the Path to Real-Time Economic Indicators: A use case in producing model-based flash estimates for monthly electricity generation: Simpler is better! Présenté à la 76e réunion du Comité consultative des méthodes statistiques, Statistique Canada.

Matthews, S. (2022). *A framework for Advance Indicators at Statistics Canada*. Document interne, Statistique Canada.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14, 149-162.

Patak, Z., et Plunkett, K. (2023). Nowcasting monthly renovation activity expenditures. Présenté lors de la réunion du 28 avril du Scientific Review Committee of the Modern Statistical Methods and Data Science Branch. Document interne, Statistique Canada.

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

Statistique Canada (2024). Ateliers, formation et conférences. Disponible à l'adresse : https://www.statcan.gc.ca/fr/afc/formation.

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 Systèmes généralisés pour les statistiques économiques

L'équipe des systèmes généralisés pour les statistiques économiques est responsable du soutien et du développement de trois systèmes généralisés : G-Sam, le système généralisé d'échantillonnage; BANFF, le système généralisé de vérification et d'imputation; et G-Est, le système généralisé d'estimation.

Progrès :

Un volume habituel de cas de soutien a été traité par l'équipe de projet pour G-Sam, BANFF et G-Est. La plupart des cas ont été réglés au moyen de suggestions sur la façon d'appliquer les systèmes en termes pratiques, mais plusieurs cas ont nécessité une intervention plus poussée. Plus particulièrement, l'équipe de BANFF a réalisé un examen approfondi du processus de vérification et d'imputation pour l'Enquête sur les postes vacants et les salaires. L'équipe de BANFF a proposé des mises à jour du plan du processus d'enchaînement des opérations et continuera de collaborer avec l'équipe d'enquête pour mettre en œuvre et mettre à l'essai ces modifications. Par ailleurs, l'équipe de G-Sam a collaboré avec des méthodologistes travaillant sur l'Enquête canadienne sur le commerce interprovincial et l'Enquête canadienne sur la situation des entreprises afin d'optimiser la répartition par rapport aux cibles de précision des erreurs-types pour des estimations de proportions. Ce cas de soutien comportait de vastes consultations avec les clients, l'examen de la théorie pertinente et l'élaboration de plusieurs exemples de programmes.

L'équipe des systèmes généralisés a participé activement à plusieurs activités liées à l'initiative de diversification analytique de Statistique Canada, une initiative de transition des systèmes vers des solutions de rechange libres. Un membre de l'équipe a participé activement au groupe de travail de l'initiative et a présenté le sujet de discussion à la 77^{e} réunion du Comité consultatif sur les méthodes statistiques (Gray, 2023). Tout au long de l'année, la sous-section a informé le Comité directeur des systèmes généralisés de l'avancement et des plans de diversification pour G-Sam, BANFF et G-Est. De plus, les membres de l'équipe ont rencontré, à l'occasion, des représentants d'organismes nationaux de statistique étrangers pour discuter de la diversification analytique, notamment l'Institut national de la statistique et des études économiques (France) pour discuter de sa propre migration de SAS à R et avec le Bureau de la statistique de l'Australie pour faire la lumière sur l'expérience de Statistique Canada à ce jour.

La version 1.04 de G-Sam est sortie le 9 novembre 2023 et comportait plusieurs modifications importantes du module de répartition. Ces modifications comprenaient : l'introduction de contraintes probabilistes spécifiées par l'utilisateur sur la taille de l'échantillon et le nombre de répondants pour des domaines arbitraires; de meilleures approximations de la variance qui sont appropriées aux solveurs d'optimisation; et un nouveau fichier de sortie de diagnostic. Les calculs sous-jacents sont donnés dans les manuels de l'utilisateur de G-Sam (Stinner, 2024). L'équipe de G-Sam a l'intention de présenter la méthode lors d'une conférence (ou d'un événement similaire) quand la version R de G-Sam sortira l'année prochaine. La version 1.04 de G-Sam est la version SAS finale.

Des progrès importants ont été réalisés dans le cadre du projet de modernisation de BANFF, et une version de Python devrait sortir en décembre 2024. Plusieurs modifications structurelles proposées par (Gray, 2022) doivent être mises en œuvre dans la version publiée, notamment de nouveaux contrôles de processus, des blocs de traitement et des modules d'extension Python personnalisés. Ces améliorations harmonisent le système avec le Modèle général d'édition des données statistiques (CEE-ONU, 2019), et l'équipe de BANFF a été invitée à faire une présentation principale sur le nouveau système à la prochaine réunion d'experts de la Commission économique des Nations Unies pour l'Europe sur la vérification de données statistiques (octobre 2024).

Des prototypes de versions libres de G-Sam et de G-Est sont en cours d'élaboration. Pour chaque système, l'équipe a évalué des options pour adopter des progiciels libres disponibles qui remplissent des fonctions particulières (par exemple stratification, calage), mais elle a constaté que ces progiciels ne répondaient pas aux normes de rendement pour la production. Des plans de diversification pour les deux systèmes ont été présentés au Comité directeur des systèmes généralisés.

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

Bibliographie

Gray, D. (2022). Banff's Next Step: An Open-Source Data Editing System for Advanced Tools and Collaboration. UNECE Expert Meeting on Statistical Data Editing.

Gray, D. (2023). Statistics Canada's Analytical Diversification Initiative – impact on Statistical Generalized Systems. Présenté à la 77e réunion du Comité consultative des méthodes statistiques, Statistique Canada.

Stinner, M. (2024). *G-Sam user guide (ébauche)*. Document interne, Statistique Canada.

UNECE (2019). Generic Statistical Data Editing Model. Disponible à l'adresse : https://statswiki.unece.org/display/sde/GSDEM.

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 des 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

d'information 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, et participer aux réunions du Groupe de travail sur le couplage d'enregistrements de la Division de la gestion du cycle de vie des solutions de la technologie de l'information (DGCSTI) et de la Division des méthodes d'intégration statistique (DMIS). L'équipe du CRCE a rencontré les représentants de la DGCSTI toutes les deux semaines et a fait le suivi des procès-verbaux mentionnant les sources possibles, passées ou présentes, de corrections, de bogues ou d'améliorations pour G-coup. 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 de 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 une plus grande participation.

Le développement a consisté à normaliser et à intégrer des procédures de revue manuelle pour l'estimation des erreurs de couplage d'enregistrements et à élaborer des indicateurs de qualité tels que la spécificité et la sensibilité, ainsi qu'à s'assurer que l'interface utilisateur rendait les résultats faciles à interpréter.

Le CRCE a également travaillé sur divers autres couplages probabilistes dans l'Environnement de couplage des données sociales ECDS. Ces couplages nous ont aidé à analyser les performances du logiciel et les solutions à apporter. Les travaux sur ces projets ont donné lieu 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 sur la repondération pour compenser le biais introduit par les liens manqués, y compris des méthodes expérimentales pour créer des poids répétés dans le cas d'un couplage entre deux échantillons tirés de manière indépendante. Les membres de l'unité ont également effectué d'autres travaux théoriques et prototypes sur les indicateurs de qualité pour des modèles ajustés à des données couplées.

Les membres de l'équipe ont offert des cours formels avec le Centre de formation de Statistique Canada, ainsi que des séminaires pour les statisticiens récemment recrutés, un forum sur le couplage d'enregistrements et d'autres présentations spéciales aux analystes.

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 à ceux d'autres ministères, ainsi qu'aux analystes et aux chercheurs du milieu universitaire et des centres de données de recherche (CDR).

Progrès :

Consultations

Des services de consultation ont été offerts à la demande de clients internes et externes. Entre le 1er avril 2023 et le 31 mars 2024, le CRAD a répondu à 37 demandes. La complexité des questions était variable et elles portaient sur des sujets comme l'analyse d'échelles de Likert, l'interprétation de résultats de régression, la comparaison de médianes, la régression logistique, les quartiles et les estimations de ratios avec des données d'enquête, la spécification de degrés de liberté, et la comparaison de plusieurs cycles d'une enquête. Le CRAD a également aidé les clients à mettre en œuvre des méthodes dans les logiciels SUDAAN, SAS, STATA et R.

Services de formation

Le CRAD a remanié et présenté, en français, le cours interne 0438A « Analyse statistique des données d'enquête – Module 1 ». Un code R a été élaboré pour les exercices et les exemples, en plus du code SUDAAN et du code SAS. Ce cours de six jours mêle théorie et pratique.

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

Collaboration

Le CRAD a collaboré avec le Secrétariat du Conseil du Trésor (SCT) 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 du Sondage auprès des fonctionnaires fédéraux (SAFF) de 2022 ont été utilisées pour mesurer des variables latentes comme les facteurs de risque psychologiques, les comportements, etc. 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).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Fritz Pierre (fritz.pierre@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.-F., Mach, L. et Houle, S. (2021). Mesurer les facteurs psychosociaux en milieu de travail au sein du gouvernement fédéral. *Rapport sur la santé*, 32, 12. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/82-003-x/2021012/article/00001-fra.pdf.

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

Le groupe de 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.

Dépersonnalisation

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.

Consultations externes

Statistique Canada a offert son expertise à plusieurs groupes au pays et à l'étranger. À l'échelle internationale, Statistique Canada a partagé son outil d'ajustement tabulaire aléatoire avec Statistique Suède et le National Agricultural Statistics Service aux États-Unis. L'équipe a également consulté des groupes externes, notamment IBM et Gurobi, afin de trouver des solutions de rechange appropriées à l'utilisation d'*OPTMODEL* de SAS dans ses stratégies de suppression de cellules complémentaires. Le groupe sur la confidentialité a également travaillé en étroite collaboration avec ses collègues de la santé à l'élaboration de nouvelles méthodes de communication de données sur le cancer avec des homologues internationaux.

Au Canada, Statistique Canada a tenu plusieurs réunions avec la Banque du Canada pour donner des conseils sur ses stratégies de contrôle de la divulgation, l'Agence du revenu du Canada pour discuter des risques de réidentification, et l'Agence de la santé publique du Canada pour discuter des données synthétiques.

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

5.6 Activités de soutien et de recherche à la Division de la science des données et de l'innovation

Le Programme de recherche et développement en méthodologie (PRDM) de Statistique Canada a appuyé de nombreuses activités pour la Division de la science des données et de l'innovation. Le soutien du PRDM

a permis la prestation de nombreux services, une communauté de pratique, des recherches plus poussées, un centre d'expertise et des lignes directrices qui peuvent profiter à l'organisme.

Activités, mandats et produits

Le financement a permis à la communauté de pratique de l'apprentissage automatique de remplir son mandat consistant à accroître les connaissances et renforcer les capacités à l'échelle de Statistique Canada en matière d'apprentissage automatique. Les activités liées à la communauté de pratique sont très variées : organisation de séminaires techniques hebdomadaires sur l'apprentissage automatique, diffusion de bulletins et de balados hebdomadaires, défis de codage et « heures de bureau porte ouverte » hebdomadaires. Un autre centre a bénéficié du financement : le Centre d'expertise du traitement du langage naturel (TLN), qui a pour mandat de centraliser les ressources pour le partage de connaissances et le renforcement des capacités en analyse de textes au moyen de l'apprentissage automatique et de créer, tenir à jour et 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 consultations et de l'orientation aux spécialistes du TLN au sein de Statistique Canada, ainsi qu'à créer la liste des projets de TLN réalisés et en cours au sein de l'organisme.

Outre les communautés et les centres, on a élaboré des lignes directrices sur l'apprentissage automatique qui forment un document pérenne à l'intention des spécialistes. Ces lignes directrices portent sur l'intelligence artificielle (IA) et l'apprentissage automatique responsables, l'IA explicable, et l'équité dans l'apprentissage automatique.

Enfin, deux projets de recherche de premier plan ont été soutenus en vue d'approfondir les expérimentations, car leurs résultats étaient prometteurs pour l'organisme. Le premier traite de la production de données synthétiques dans le domaine de la santé : il évalue l'innocuité et l'efficacité des méthodes de production de données synthétiques sur les soins de santé, en s'intéressant particulièrement à l'équilibre entre utilité et confidentialité. Ce projet s'inscrit dans la ligne des initiatives de l'Institut canadien de recherches avancées et vise à moderniser le partage des données et la prise de décisions dans le secteur des soins de santé. Un atelier sur le projet se tiendra prochainement.

Le deuxième projet de recherche prometteur financé par le comité est une recherche visant à améliorer l'efficacité du prétraitement des microdonnées dans les données d'enquête à Statistique Canada. Le projet s'intéresse à la robustesse du prétraitement des données pour les tâches complexes et aux performances d'une méthode d'imputation dans différentes conditions. Cette recherche est conçue pour optimiser l'actualité et réduire les efforts manuels dans le cadre de l'Enquête mensuelle sur l'approvisionnement et l'écoulement de l'électricité (voir aussi le projet « Optimisation du temps et des efforts dans le traitement des données » à la section 2).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Marie-Eve Bedard (marie-eve.bedard@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 il 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.

Bien qu'une grande partie du travail du Centre de ressources en conception de questionnaires soit 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 offre aussi des cours sur la conception de questionnaires.

Progrès :

Le Centre de ressources en conception de questionnaires a effectué de nombreux examens de questionnaires d'enquête. Bien que 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 Transports Canada, Services publics et Approvisionnement Canada et d'autres.

Le groupe a également contribué à diverses initiatives de consultation de Statistique Canada.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Paul Kelly (paul.kelly@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 le 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, il faut étudier des méthodologies diverses, 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 opérations, mais aussi par leur grande applicabilité à différentes étapes des opérations 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 a concentré ses efforts sur l'élaboration d'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 de l'apprentissage automatique (AA). Comme la technologie d'AA joue un rôle de plus en plus essentiel, il est primordial d'assurer la qualité des codes générés. Pour répondre à cette nécessité, 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 permettra une méthode responsable de classification des données avec une mise en œuvre plus grande de l'apprentissage automatique. Notre objectif est d'utiliser cette méthode à des fins de CQ pour plusieurs classifications à l'intérieur d'enquêtes essentielles comme l'Enquête sur la population active (EPA), l'Enquête sur les postes vacants et les salaires (EPVS), l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et le Registre statistique des entreprises (RSE). Un article décrivant en détail la méthodologie a notamment été présenté au Comité consultatif sur les méthodes statistiques (Oyarzun, Wile et Evans, 2023).

L'équipe du CRAQ a également étudié une méthodologie de calage visant à établir une relation plus cohérente entre les scores d'apprentissage automatique et l'exactitude des données codées. Dans ce contexte, le calage consiste à affiner le modèle d'apprentissage automatique afin d'aligner son système de notation avec l'exactitude réelle des données codées. En affinant cette relation, nous cherchons à améliorer la fiabilité et la précision du processus de classification, ce qui, en fin de compte, améliorera les mesures globales d'assurance de la qualité dans nos opérations de traitement des données. Cette méthodologie de calage représente une étape cruciale vers l'optimisation du rendement des algorithmes d'apprentissage automatique dans les tâches de codage, puisqu'elle assure une correspondance étroite entre les sorties et les classifications souhaitées et qu'elle renforce la confiance dans les données finales.

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

Bibliographie

Oyarzun, J., Wile, L. et Evans, J. (2023). Quality Control by Score. Document présenté au Comité consultative sur les méthodes statistiques, octobre 2023, Statistique Canada.

5.9 Secrétariat de l'éthique des données

Le Secrétariat de l'éthique des données a pour rôle 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. Ces examens éthiques ont pour objectif de garantir 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 à l'agent principal de l'éthique des données et de l'intégrité scientifique. Le Secrétariat de 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 effectué environ 180 examens éthiques, les membres du Secrétariat de l'éthique des données ont donné de nombreuses présentations pour informer les partenaires internes et les collègues d'autres ministères fédéraux et d'organismes internationaux sur l'approche de Statistique Canada en matière d'éthique des données. L'équipe recueille des renseignements pour être à jour à propos de sujets jugés délicats par le public. Pour ce faire, elle procède à une revue de la littérature sur certains sujets ciblés, 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 CEE-ONU sur le leadership éthique. Le principal objectif de l'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 considérablement avancé en mars 2024 lors de l'atelier sur l'éthique dans les organismes statistiques modernes, où des séances de discussion sur chaque section du manuel de référence ont été organisées. L'ouvrage de référence devrait être terminé en 2025.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Martin Beaulieu (martin-j.beaulieu@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 avec des partenaires 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 a trait aux partenaires nationaux, le Secrétariat a présenté des exposés officiels sur les pratiques de gestion de la qualité à deux organismes, en plus de tenir un certain nombre d'ateliers et de séminaires. Le Secrétariat de la qualité a collaboré, par l'entremise de l'Initiative de formation en littératie des données de Statistique Canada, à l'élaboration d'un module de formation en ligne intitulé Adéquation de la qualité des données à l'utilisation prévue. Ce cours présente un cadre clair et facile à utiliser pour aider les apprenants et apprenantes à définir leurs besoins en données, à évaluer la pertinence des sources de données

potentielles selon la finalité prévue, et à décider si une source de données potentielle est en fait adaptée à l'utilisation prévue.

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 pour mandat, dans le cadre de la mise en œuvre de la Stratégie de données, de définir un cadre de qualité des données applicable à tous les organismes du gouvernement du Canada. Une ébauche du Cadre de la qualité des données est mise à la disposition des partenaires d'autres ministères fédéraux, et une version abrégée du cadre, appelée Orientation sur la qualité des données, a été approuvée et mise à la disposition du public en janvier 2024.

À l'échelle internationale, notre participation au Groupe d'experts des Nations Unies des cadres nationaux d'assurance de la qualité s'est accrue, puisque nous assumons le rôle de coprésident du sous-groupe sur les sources de données administratives et autres. Le but du sous-groupe est de préparer un module d'assurance de la qualité en cas d'utilisation de sources de données administratives et autres pour produire des statistiques officielles. Ce module 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 des sources de données administratives, d'autres sources de données ou de multiples 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). Le module fera l'objet d'une consultation mondiale au printemps 2024 et sera soumis à l'approbation de la Commission de statistique des Nations Unies vers la fin de 2024.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Martin Beaulieu (martin-j.beaulieu@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/.

PROJET : Indicateurs de qualité pour les statistiques tirées de données intégrées

Afin de fournir aux utilisateurs des indicateurs de qualité pour les programmes combinant des sources de données administratives, le Secrétariat de la qualité a travaillé à l'élaboration d'un indicateur composite qui combine des indicateurs de qualité liés à différentes étapes du traitement des données (couplage d'enregistrements, imputation, géocodage, etc.) en un seul indicateur. L'objectif est de donner un aperçu global de la qualité d'une estimation en tenant compte de plusieurs facteurs qui peuvent introduire des erreurs (Gagnon, Qian, Yeung, Lebrasseur et Beaulieu, 2022).

Progrès :

Ces indicateurs ont été utilisés pour d'autres tableaux du Programme de la statistique du logement canadien (PSLC). Le Secrétariat de la qualité continue de fournir un soutien sur le code et la méthode.

Certaines solutions ont été étudiées en collaboration avec l'équipe du PSLC afin d'améliorer la méthode et la cohérence des résultats.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

Bibliographie

Gagnon, R., Qian, W., Yeung, A., Lebrasseur, D. et Beaulieu, M. (2022). Développement d'un indicateur composite de qualité pour les produits statistiques dérivés de sources administratives. Statistique Canada, disponible à l'adresse : https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-fra.htm.

PROJET : Mise à jour du Cadre d'assurance de la qualité

Le Secrétariat de la qualité a entrepris un examen du Cadre d'assurance de la qualité (CAQ) de Statistique Canada. La version actuelle a été publiée en 2017. Bien que le contenu de la version actuelle soit toujours valide, l'évolution rapide des nouvelles sources de données et des nouvelles techniques utilisées dans la production de statistiques officielles a rendu cet examen pertinent. La version mise à jour soulignera l'importance de l'intendance des données, des principes d'éthique des données et de certaines considérations relatives aux nouvelles techniques utilisées. Le plan de mise à jour a été présenté au Comité consultatif sur les méthodes statistiques à l'automne 2022 (Beaulieu, Yung et Rancourt, 2022).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

Bibliographie

Beaulieu, M., Yung, W. et Rancourt, E. (2022). Data Quality and Official Statistics in a Modern World. Document présenté au Comité consultatif sur les méthodes statistiques, octobre 2022, Statistique Canada.

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. La 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 pertinentes 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/techniquesdenquete. Le comité de rédaction est formé 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 2023 (49-1 et 49-2) ont été publiés. Le numéro de juin 2023 compte onze articles, dont un article spécial à la mémoire du professeur Chris Skinner, lauréat du prix Waksberg 2019, écrit par Natalie Shlomo, accompagné d'un hommage de Danny Pfeffermann et des commentaires de J.N.K. Rao ainsi que ceux de Jae Kwang Kim et Haiying Wang. Dix-huit articles ont été publiés dans le numéro de décembre 2023, incluant l'article Waksberg 2023 de Ray Chambers, intitulé « Le principe de l'information manquante – Un paradigme d'analyse de données désordonnées d'enquête par sondage », ainsi qu'un article spécial de Pascal Ardilly, David Haziza, Pierre Lavallée et Yves Tillé sur les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle. Ce dernier article comprend cinq discussions par Guillaume Chauvet, Marc Christine, Françoise Dupont, Camelia Goga et Anne Ruiz-Gazen, et Carl-Erik Särndal, ainsi qu'une réponse des auteurs. Le numéro de décembre présente également quatre articles sollicités qui ont été présentés au Colloque francophone sur les sondages de 2021.

En 2023, 44 articles ont été soumis à la revue. Le nombre moyen de jours entre la soumission et l'évaluation initiale était de 55. Tous les articles soumis ont été évalués à l'intérieur d'un délai de 130 jours, à l'exception d'un article qui exceptionnellement a été évalué en 188 jours, et 80 % d'entre eux ont été évalués à l'intérieur d'un délai de 90 jours. Parmi les 44 articles soumis, 24 ont été rejetés, 11 ont été acceptés et 9 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 18 juillet 2024. D'avril 2023 à mars 2024, les pages de *Techniques d'enquête* ont été consultées 52 046 fois.

Le numéro de juin 2024 est consacré à trois articles présentés à la conférence Morris Hansen de 2022 sur l'utilisation d'échantillons non probabilistes par Courtney Kennedy, Yan Li et Jean-François Beaumont. Les trois articles font l'objet de discussion par des spécialistes internationaux du domaine, et les discussions sont suivies de réponses. Une introduction de Partha Lahiri, rédacteur invité pour ce numéro spécial, précède les articles. Le numéro de juin 2025 sera consacré aux célébrations du 50^{e} anniversaire de *Techniques d'enquête*. Il comprendra un article spécial avec discussion de Carl-Erik Särndal, intitulé « Le progrès dans la science des enquêtes: hier – aujourd'hui – demain », ainsi que des discussions avec d'éminents statisticiens d'enquête. Outre l'article de Särndal et sa discussion, le numéro contiendra également plusieurs articles sollicités rédigés par des spécialistes renommés de la statistique et de la méthodologie d'enquête.

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, dont le mandat principal demeure la formation en statistique au sein de la Direction et de l'organisme, a connu une autre année occupée et productive. Plusieurs cours ont été offerts cette année dont ceux reliés aux séries chronologiques, à la conception de questionnaires, à l'échantillonnage, au couplage d'enregistrements, à l'imputation, à la pondération, à l'estimation sur petits domaines, au bootstrap, à la modélisation, à l'introduction à l'apprentissage automatique (AA) supervisé, à l'équité et l'explicabilité en AA et à la programmation en R.

Pour ce qui est des nouvelles activités, le groupe a continué à concevoir et à accorder la priorité à des activités d'apprentissage qui peuvent être élaborées en temps opportun et axées sur l'apprentissage actif. Cette année, un nouveau cours sur l'analyse statistique avec des données d'enquête a été développé. Ce cours expose notamment comment faire des analyses statistiques sous un plan de sondage complexe. Le cours a été offert dans les deux langues officielles et d'autres séances sont prévues pour l'an prochain. Nous avons aussi développé un cours d'introduction à la programmation avec Python. Ce cours a également été offert dans les deux langues officielles.

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, un cours sur l'estimation de la variance sera développé et offert au cours de l'année 2024-25. Un atelier concernant la conception de simulations sera également développé et offert en 2024-25.

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 la plateforme DataCamp, 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 sera « Façonner l'avenir des statistiques officielles », aura lieu les 30 et 31 octobre et le 1er novembre, 2024. Le Symposium offrira des séances plénières ainsi que des séances parallèles qui porteront sur divers thèmes. Tout comme les autres congrès à travers le monde, le Symposium de cette année offrira des conférenciers qui livreront leurs présentations en présentiel. Les observateurs auront le choix d'y assister en personne ou de se joindre aux séances sous forme virtuelle.

Progrès :

Les membres du comité organisateur, du comité du programme et du comité de la logistique ont été nommés, et le titre et le format virtuel du symposium de cette année ont été confirmés. Le comité de la logistique a commencé à assurer la coordination avec les Services de conférences au sujet de la plateforme de la conférence et des services connexes, comme la prestation de soutien technique et l'interprétation simultanée.

Le comité du programme a identifié plusieurs thèmes et organisateurs pour les séances invitées, et a commencé à identifier les thèmes et les organisateurs des séances de contributions libres. Nos séances plénières incluront un discours du lauréat du prix Waksberg en 2024, Richard Valliant, et un autre de notre conférencier principal.

Tout renseignement pertinent sera apporté au site Web à l'adresse :
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

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2024). Réponse des auteurs aux commentaires sur l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada » : De nouvelles avancées concernant les méthodes de vraisemblance pour l'estimation des probabilités de participation pour des échantillons non probabilistes. *Techniques d'enquête*, 50, 1, 139-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00001-fra.pdf.

Brennan, A., et Summers, P. (2023a). Synthetic population modelling details. Rapport interne, Statistique Canada.

Brennan, A., et Summers, P. (2023b). Synthetic population overview. Rapport interne, Statistique Canada.

Chen, H., et Thomas, S. (2023). Assessing the Performance of the Open-Source Linear Programming Solver in Cell Suppression Problems. Dans *Proceedings of the Survey Research Methods Section,* American Statistical Association. Disponible à l'adresse : https://doi.org/10.5281/zenodo.10359791.

Dasylva, A. (2024). Estimation of small area means with linked data. Rapport interne, Statistique Canada.

Dasylva, A., Beaumont, J.-F., Bosa, K. et Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, mai 2023.

Dasylva, A., et Goussanou, A. (2024a). Measuring the linkage accuracy when many files are linked to a spine. Rapport interne, Statistique Canada.

Dasylva, A., et Goussanou, A. (2024b). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*. Disponible à l'adresse : https://doi.org/10.1007/s42081-023-00228-9.

Dasylva, A., Goussanou, A. et Nambeu, C.-O. (2024). Modèles d'erreur de couplage pour l'estimation par capture-recapture sans vérifications manuelles. *Techniques d'enquête*, 50, 2. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024002/article/00007-fra.pdf.

Francis, J. (2023). *Canadian CPI Sample Design Methodology Review*. Document interne, Statistique Canada.

Francis, J. (2024). *Design-Based Estimates and Variance Estimation for SPPIs with Sequential Poisson PPS Samples*. Document interne, Statistique Canada.

Gray, D. (2023). Statistics Canada's Analytical Diversification Initiative – impact on Statistical Generalized Systems. Présenté à la 77e réunion du Comité consultative des méthodes statistiques, Statistique Canada.

Le Moullec, J., et Matthews, S. (2023). On the Path to Real-Time Economic Indicators: A use case in producing model-based flash estimates for monthly electricity generation: Simpler is better! Présenté à la 76e réunion du Comité consultative sur les méthodes statistiques, Statistique Canada.

Mather, A., Boulet, C. et Brennan, A. (2024). An Estimator for Concurrent Use of Full and Reduced Collection Effort on Random Subsamples. Document présenté au Comité consultative sur les méthodes statistiques, 78, Statistique Canada.

Millar, G. (2024). Logistic regression on linked data from a secondary analyst perspective. Présentation à la CANSSI-CRT Workshop on Modern Methods in Survey Sampling, University of Ottawa, 8 au 10 juillet.

Miller, J. (2024). *Disclosure Risk of Parametric Regression Output*. Rapport interne, Statistique Canada.

Oyarzun, J., Wile, L. et Evans, J. (2023). Quality Control by Score. Document présenté au Comité consultative sur les méthodes statistiques, octobre 2023, Statistique Canada.

Patak, Z., et Plunkett, K. (2023). Nowcasting monthly renovation activity expenditures. Présenté lors de la réunion du 28 avril du Scientific Review Committee of the Modern Statistical Methods and Data Science Branch. Document interne, Statistique Canada.

Santos, B. (2023). Multi-Party Privacy Preserving Record Linkage based on Circuit Private Set Intersection. Canadian Mathematical Society Winter Meeting, Montréal, Canada.

Statistique Canada (2024). *Le Quotidien* – Nouveau modèle de microsimulation de revenu de retraite maintenant offert. Disponible à l'adresse : https://www150.statcan.gc.ca/n1/daily-quotidien/240423/dq240423c-fra.htm.

Stinner, M. (2024). *G-Sam user guide (ébauche)*. Document interne, Statistique Canada.

Toukal, J. (2023). Logistic regression in the context of record linkage. Rapport de stage pour l'École Nationale de la Statistique et de l'analyse de l'information.

Toupin, M.-H., et Martin, V. (2024). 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).

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition New Series*, 24, 169-178.

You, Y., et Bosa, K. (2024). Performance of hierarchical Bayes small area estimators using non-informative and informative priors with LFS application. Soumis à *Techniques d'enquête* (en cours de révision).

You, Y., et Hidiroglou, M. (2024). Empirical study of confidence intervals for small area proportion estimation with LFS application. Rapport interne, Statistique Canada.