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

Programme de recherche et développement en méthodologie : réalisations, 2021-2022

Date de diffusion : le 7 octobre 2022

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, 2021-2022

Le présent rapport fait la synthèse des réalisations en 2021-2022 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 en méthodes statistiques 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, communiquez avec :

Jean-François Beaumont
(613-863-9024, jean-francois.beaumont@statcan.gc.ca)

Wesley Yung
(613-404-2203, wesley.yung@statcan.gc.ca)

Programme de recherche et développement en méthodologie :

réalisations, 2021-2022

Table des matières

1 Intégration des données

1.1 Intégration des données d'échantillons probabilistes et d'échantillons non probabilistes

SOUS-PROJET : Traitement d'échantillons non probabilistes par pondération de probabilité inverse

Statistique Canada et d'autres organismes nationaux de statistique étudient de plus en plus la possibilité d'utiliser des échantillons non probabilistes en complément des échantillons probabilistes. Il est cependant bien connu que l'utilisation d'un échantillon non probabiliste seulement peut produire des estimations présentant un biais important 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 qui présente des variables auxiliaires communes à l'échantillon non probabiliste.

Dans le cadre de cette étude, nous nous sommes concentrés sur les méthodes de pondération de probabilité inverse, lesquelles consistent à modéliser la probabilité de participation à l'échantillon non probabiliste. Comme point de départ, nous avons examiné un modèle logistique ainsi que la méthode du pseudo maximum de vraisemblance de Chen, Li et Wu (2020). Nous avons proposé une procédure de sélection de variables basée sur un critère d'information d'Akaike modifié qui tient compte de la structure des données et du plan de sondage probabiliste. Nous avons également proposé une méthode simple fondée sur le rang pour former des strates *a posteriori* homogènes. De plus, nous avons étendu l'algorithme des arbres de classification et de régression (CART) à ce scénario d'intégration des données, tout en tenant compte, encore une fois, du plan de sondage probabiliste. Notre version modifiée de l'algorithme CART s'appelle nppCART. Enfin, nous avons proposé un estimateur de la variance bootstrap, qui reflète deux sources de variabilité : le plan de sondage probabiliste et le modèle de participation.

Progrès :

Nous avons terminé la rédaction du progiciel R pour le nppCART et nous avons poursuivi nos travaux empiriques en utilisant les données de l'approche participative comme échantillon non probabiliste et celles de l'Enquête sur la population active ou de la Série d'enquêtes sur les perspectives canadiennes comme échantillon probabiliste. Une des principales conclusions tirées à la suite de nos expériences est que toutes les méthodes examinées ont permis de réduire les biais, mais qu'un biais persiste. Nous avons aussi observé l'importance de former des groupes homogènes. L'algorithme nppCART a bien fonctionné avec ces données. La régression logistique avec effets principaux seulement constitue également une option raisonnable à condition que les probabilités de participation estimées du modèle logistique soient utilisées pour former des groupes homogènes.

Le projet a été présenté lors de la Morris Hansen Memorial Lecture de 2022 (Beaumont, Bosa, Brennan, Charlebois et Chu, 2022). Nous prévoyons terminer la rédaction d'un article qui sera présenté à une revue statistique à comité de lecture.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2022). Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. Présentation à la 2022 Morris Hansen Memorial Lecture, le 1er mars 2022.

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.

SOUS-PROJET : Analyse de sensibilité pour évaluer le biais plausible de non-réponse ou de participation dans les données d'échantillons

Les enquêtes à faible taux de réponse et les échantillons non probabilistes peuvent présenter des biais. Les biais peuvent être atténués par l'utilisation de modèles, comme ceux utilisés dans les ajustements pour la non-réponse, lorsque de bons renseignements auxiliaires sont disponibles. Ces modèles sont toutefois sensibles au manque de renseignements pertinents (des variables auxiliaires qui sont associées à la fois à la participation et aux variables d'intérêt). Lorsqu'il manque des renseignements pertinents, un biais peut persister après la modélisation de la non-réponse.

Le projet visait l'évaluation du biais plausible qui subsiste en raison de renseignements auxiliaires manquants. La stratégie était d'adapter une analyse de sensibilité du domaine de l'inférence causale. Une analyse de sensibilité débute par une prémisse limitant la perturbation (p. ex. l'importance des renseignements manquants) et convertit la prémisse en assurances au sujet d'une quantité d'intérêt (p. ex. le biais d'estimations fondées sur un échantillon). Pour l'inférence causale, il y a des unités de traitement et de contrôle dans une étude d'observation, et la quantité d'intérêt est l'effet causal du traitement. Une analyse de sensibilité élaborée pour une telle situation doit être adaptée à notre problème d'estimation pour des échantillons biaisés par la non-réponse.

Progrès :

Nous avons adapté l'analyse de sensibilité de Ding et VanderWeele (2016) pour évaluer les estimations d'échantillons biaisés par la non-réponse. Pour toute prémisse donnée limitant l'importance des renseignements auxiliaires manquants, la méthode fournit des limites pour les biais qui en résultent. Notre analyse de sensibilité hérite de nombreuses propriétés souhaitables de la méthode de Ding et VanderWeele : elle ne requiert que trois paramètres, elle ne pose aucune hypothèse sur la dimension ou la distribution des renseignements manquants, les limites sont nettes et la prémisse peut être interprétée et peut être fondée sur les données actuelles ou sur une expertise spécialisée. Le dernier point est essentiel, car il signifie que la prémisse est défendable, ce qui donne des limites de biais défendables.

Nous avons testé l'analyse de sensibilité à l'aide d'un échantillon non probabiliste de l'approche participative et comparé les résultats à ceux d'une enquête probabiliste contemporaine ayant le même contenu (Brennan, 2022). Nous prévoyons présenter les résultats à une revue statistique à comité de lecture.

Pour obtenir plus de renseignements, communiquez avec :
Andrew Brennan (343-548-4028, andrew.brennan@statcan.gc.ca).

Bibliographie

Brennan, A. (2022). A sensitivity analysis for assessing plausible non-response or participation bias in sample data. Rapport interne, Statistique Canada.

Ding, P., et VanderWeele, T.J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3), 368.

SOUS-PROJET : Approche bayésienne approximative pour l'intégration des données d'échantillons probabilistes et d'échantillons non probabilistes

Dans le cadre de ce projet d'intégration de données, nous examinons la situation où la variable d'intérêt et les variables auxiliaires sont observées à la fois dans un échantillon probabiliste et dans un échantillon non probabiliste. Nous cherchons à utiliser les données de l'échantillon non probabiliste pour améliorer l'efficacité des estimations pondérées par les poids d'enquête obtenues à partir de l'échantillon probabiliste. Récemment, Sakshaug, Wiśniowski, Ruiz et Blom (2019) et Wiśniowski, Sakshaug, Ruiz et Blom (2020) ont proposé une approche bayésienne pour l'intégration des données des deux échantillons aux fins de l'estimation de paramètres d'un modèle. Dans leur méthode, les données de l'échantillon non probabiliste sont utilisées pour déterminer la distribution *a priori* des paramètres du modèle, et la distribution *a posteriori* est obtenue en se fondant sur l'hypothèse selon laquelle le plan de sondage probabiliste est ignorable (ou non informatif). L'objectif du projet était d'étendre cette approche bayésienne à la prédiction de paramètres d'une population finie selon un plan de sondage non ignorable (ou informatif).

Progrès :

Nous avons proposé, conformément à Wang, Kim et Yang (2018), une procédure bayésienne approximative qui tient compte du plan de sondage probabiliste en nous appuyant sur des statistiques pondérées par des poids d'enquête appropriés. Nous avons mené des expériences par simulations et rédigé un article pour les actes du symposium de 2021 de Statistique Canada (You, Dasylva et Beaumont, 2021), qui brosse un tableau des méthodes et résume nos résultats empiriques. La principale conclusion qui découle de nos expériences est que notre approche bayésienne peut produire des gains d'efficacité non négligeables par rapport aux estimateurs pondérés par les poids d'enquête, même dans une situation où l'échantillon non probabiliste est très informatif, à condition que la variance a priori des paramètres du modèle soit soigneusement choisie. Toutefois, nous avons également observé de petites pertes d'efficacité dans un scénario où la corrélation entre la variable d'intérêt et les variables auxiliaires était faible. Le projet sera présenté à la réunion de juin du Comité consultatif des méthodes statistiques et aux Joint Statistical Meetings de 2022.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Sakshaug, J.W., Wiśniowski, A., Ruiz, D.A.P. et Blom, A.G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, 35, 653-681.

Wang, Z., Kim, J.K. et Yang, S. (2018). Approximate Bayesian inference under informative sampling. *Biometrika*, 105, 91-102.

Wiśniowski, A., Sakshaug, J.W., Ruiz, D.A.P. et Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8, 120-147.

You, Y., Dasylva, A. et Beaumont, J.-F. (2021). An approximate Bayesian approach to improving probability sample estimators using a supplementary non-probability sample. *Recueil du Symposium international de 2021 sur les questions de méthodologie*, Statistique Canada.

SOUS-PROJET : Lissage des poids pour les enquêtes non probabilistes

Les techniques d'ajustement permettant d'atténuer le biais de sélection dans les échantillons non probabilistes font souvent appel à la modélisation de la propension à participer à l'échantillon non probabiliste ainsi qu'à la pondération par l'inverse de la propension à répondre. Il est bien connu que les procédures d'estimation des poids sont efficaces si les covariables sélectionnées dans le modèle de propension sont liées à la fois à la variable d'intérêt et à l'indicateur de participation. Dans la plupart des enquêtes, il y a de nombreuses variables d'intérêt, ce qui rend les ajustements de poids difficiles à déterminer, car un poids approprié pour une variable peut ne pas convenir pour d'autres variables. Le compromis type consiste à inclure un grand nombre de covariables dans le modèle de propension, mais cela peut accroître la variabilité des estimations, particulièrement lorsque certaines covariables ont un faible lien avec les variables d'intérêt. Le lissage des poids, établi pour les enquêtes probabilistes, pourrait être utile dans de telles situations. Il permet d'éliminer la variabilité causée par les modèles de propension surajustés en remplaçant les poids de propension inverses par les poids prédits obtenus à l'aide d'un modèle de lissage. Dans le projet, nous avons étudié le lissage des poids dans le contexte des enquêtes non probabilistes, autant sur le plan théorique que sur le plan empirique, afin de comprendre son utilité pour améliorer l'efficacité des estimations.

Progrès :

Nous avons d'abord montré théoriquement que l'estimateur lissé n'est jamais moins efficace que sa version non lissée dans un modèle linéaire pour les poids de propension. Nous avons également établi les conditions favorables à des gains d'efficacité plus importants. Par exemple, de tels gains seraient réalisés lorsque les variables d'intérêt ont un lien faible avec les covariables utilisées dans le modèle de propension. Ensuite, nous avons conçu deux études de simulation, basées sur des données artificielles et des données réelles, pour évaluer les propriétés du lissage des poids et son efficacité pour réduire l'erreur quadratique moyenne des estimations. Nous avons terminé la rédaction d'un article, qui a été accepté et publié dans la revue *TEST* (Ferri-García, Beaumont, Bosa, Charlebois et Chu, 2021).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Ferri-García, R., Beaumont, J.-F., Bosa, K., Charlebois, J. et Chu, K. (2021). Weight smoothing for nonprobability surveys. *TEST*.

SOUS-PROJET : Intégration des données statistiques selon une approche par prédiction

Nous avons examiné la façon dont une grande base de données non probabilistes peut être utilisée, au moyen de techniques d'intégration des données, pour améliorer les estimations faites à partir d'un petit échantillon probabiliste. Dans la situation où la variable d'intérêt est observée dans les deux sources de données, Kim et Tam (2021) ont proposé deux estimateurs convergents par rapport au plan de sondage qui peuvent être justifiés par la théorie des enquêtes à double base de sondage. Dans la première partie du projet, nous avons déterminé les conditions garantissant que ces estimateurs sont plus efficaces que l'estimateur Horvitz-Thompson lorsque l'échantillon probabiliste est sélectionné à l'aide de l'échantillonnage de poisson ou de l'échantillonnage aléatoire simple sans remise. Dans la deuxième partie du projet, nous avons étudié une catégorie de prédicteurs, proposée par Särndal et Wright (1984), qui traite le cas où la base de données non probabiliste contient des variables auxiliaires, mais aucune variable d'intérêt. L'enquête probabiliste permet de recueillir les variables d'intérêt, et nous supposons que la base de données non probabiliste peut être couplée à l'échantillon probabiliste. Un tel cas présente un intérêt pour une enquête sur le trafic postal menée par La Poste en France.

Progrès :

Nous avons mené une étude par simulations pour comparer les propriétés par rapport au plan de différents prédicteurs dans la catégorie de prédicteurs proposée par Särndal et Wright (1984). Ces prédicteurs comprennent un prédicteur fondé sur un modèle, un estimateur assisté par un modèle et un estimateur cosmétique. Dans nos configurations de simulation, l'estimateur cosmétique a obtenu un rendement légèrement supérieur à l'estimateur assisté par un modèle. Comme prévu, le prédicteur fondé sur un modèle n'a pas donné de bons résultats lorsque le modèle sous-jacent était mal défini.

Le projet repose sur une collaboration avec La Poste en France ainsi qu'avec l'École d'économie de Toulouse et l'Université de Besançon. Les résultats préliminaires du projet ont été présentés lors d'une séance sur invitation au Colloque francophone sur les sondages à l'automne 2021. Un article a été rédigé et présenté à une revue à comité de lecture (Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine et Puech, 2022a).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Kim, J.-K., et Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89,382–401.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. et Puech, P. (2022a). QR prediction for statistical data integration. Manuscrit inédit en cours d'examen.

Särndal, C.-E., et Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146–156.

1.2 Couplage d'enregistrements

Le couplage d'enregistrements joue un rôle important dans la production de statistiques officielles. Il peut toutefois donner lieu à des erreurs, car il repose souvent sur des quasi-identificateurs non uniques qui sont enregistrés avec des variations et des erreurs typographiques. Le projet porte sur la production et l'utilisation de données couplées, y compris sur l'estimation exacte des erreurs de couplage.

SOUS-PROJET : Estimation des erreurs de couplage

Les erreurs de couplage comprennent les faux négatifs et les faux positifs; le faux négatif ne permet pas de coupler les enregistrements d'une même unité, et le faux positif couple des enregistrements de différentes unités. Ces erreurs sont mesurées par le rappel et la précision. L'estimation des mesures est essentielle pour pouvoir coupler efficacement les ensembles de données, rendre compte de la qualité des données couplées et apporter des ajustements en tenant compte des erreurs lors de l'analyse des données. Le projet permet d'examiner la convergence des estimateurs qui modélisent le nombre de couplages faits à partir d'un enregistrement donné (Blakely et Salmond, 2002; Dasylva et Goussanou, 2020; Dasylva et Goussanou, 2021). Il est essentiel d'établir cette convergence si les erreurs doivent être estimées sans accès aux données réelles sur le terrain. Cela est toutefois difficile parce que les estimateurs sont fondés sur des modèles où les observations sous-jacentes ont une structure de dépendance complexe, c'est-à-dire une structure qui n'est pas facilement décrite par les structures de dépendance habituelles, comme les grappes indépendantes (comme dans les modèles linéaires mixtes généralisés), les séries chronologiques ou les U-statistiques.

Progrès :

Il a été démontré que la précision et le rappel satisfont à une loi des grands nombres dans des conditions générales, pour une vaste catégorie de stratégies de couplage, y compris certaines stratégies dans le cadre desquelles la décision de coupler deux enregistrements peut toucher d'autres enregistrements (Dasylva et Goussanou, 2022). Les estimateurs proposés par Blakely et Salmond (2002) et par Dasylva et Goussanou (2020) se sont également révélés convergents, c'est-à-dire qu'ils convergent vers la précision et le rappel limitatifs (Dasylva et Goussanou, 2022) lorsque la population devient arbitrairement importante.

Pour obtenir plus de renseignements, communiquez avec :
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

Bibliographie

Blakely, T., et Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *Journal of Epidemiology*, 31, 1246-1252.

Dasylva, A., et Goussanou, A. (2020). Estimating linkage errors under regularity conditions. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687-692.

Dasylva, A., et Goussanou, A. (2021). Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements. *Techniques d'enquête*, 47, 2, 325-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2021002/article/00002-fra.pdf.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*.

SOUS-PROJET : Capture-recapture avec erreurs de couplage

Dans le contexte de son paradigme « données administratives d'abord », Statistique Canada donne la priorité à l'utilisation de sources autres que les enquêtes pour produire des statistiques officielles. Ce paradigme repose de façon capitale sur des sources autres que les enquêtes pouvant fournir une couverture quasi parfaite de certaines populations cibles, y compris des fichiers administratifs ou des sources de mégadonnées. Cette couverture doit toutefois être mesurée, en appliquant par exemple la méthode de capture-recapture, selon laquelle les données sont comparées à d'autres sources présentant une bonne couverture des mêmes populations, y compris un recensement. Cependant, il s'agit d'un exercice difficile lorsque des erreurs de couplage sont présentes (Ding et Fienberg, 1994; Di Consiglio et Tuoto, 2015), ce qui survient inévitablement lorsque le couplage repose sur des quasi-identificateurs, comme cela est généralement le cas.

Progrès :

Une nouvelle méthodologie de capture-recapture avec erreurs de couplage a été élaborée (Dasylva, Goussanou et Nambeu, 2021). Elle est fondée sur le modèle d'erreurs de couplage proposé par Dasylva et Goussanou (2020), et comporte une extension pour tenir compte de la sous-couverture. Comparativement aux solutions précédentes proposées par Ding et Fienberg (1994) et Di Consiglio et Tuoto (2015), la nouvelle méthodologie ne tient pas pour acquis qu'il n'y a pas de faux positifs. Elle a été appliquée avec succès dans le cadre d'une expérience avec des données publiques de recensement.

Pour obtenir plus de renseignements, communiquez avec :
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

Bibliographie

Dasylva, A., et Goussanou, A. (2020). Estimating linkage errors under regularity conditions. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687-692.

Dasylva, A., Goussanou, A. et Nambeu, C.-O. (2021). Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors. Dans le *Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

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.

SOUS-PROJET : Intersection d'ensembles privés avec erreurs de couplage

Un organisme statistique peut évaluer la couverture d'une source de données d'une tierce partie en la comparant à une source de référence au moyen du couplage d'enregistrements et de l'estimation par capture-recapture. Cette comparaison représente toutefois un défi lorsque l'organisme statistique ne peut pas accéder directement à la source de la tierce partie parce que la confiance est limitée entre les deux parties. Lorsqu'il n'y a pas d'erreur de couplage, il est possible d'utiliser le protocole décrit par Christen (2012, chapitre 8) pour estimer la taille de l'intersection entre l'ensemble de données de la tierce partie et l'ensemble de données de référence, d'une manière qui permet de préserver la confidentialité. La solution est particulièrement simple, car elle ne requiert que des comparaisons exactes. La présence d'erreurs de couplage complique toutefois les choses.

Progrès :

Une méthodologie simple a été décrite pour l'estimation par capture-recapture lorsque l'intersection est déterminée au moyen du protocole d'intersection d'ensembles privés décrit par Christen (2012, chapitre 8), que les comparaisons sont exactes et que des erreurs de couplage sont présentes (Dasylva et Zanussi, 2021a, 2021b). La solution fait appel au modèle décrit par Dasylva et Goussanou (2020), qui permet à l'organisme d'apporter des ajustements en tenant compte des erreurs de couplage lorsqu'il estime la taille de l'intersection et de la couverture de la source de la tierce partie.

Pour obtenir plus de renseignements, communiquez avec :
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

Bibliographie

Christen, P. (2012). *Data Matching*. Springer, Berlin.

Dasylva, A., et Goussanou, A. (2020). Estimating linkage errors under regularity conditions. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687-692.

Dasylva, A., et Zanussi, Z. (2021a). Measuring the coverage of a data source with a private set intersection. *Rapport interne, Statistique Canada,* Ottawa.

Dasylva, A., et Zanussi, Z. (2021b). A Private Set Intersection Use Case. Présentation à l'UNECE Input Privacy-Preserving webinar, Nov. 2021.

1.3 Estimation sur petits domaines

SOUS-PROJET : Inférence hiérarchique bayésienne pour l'estimation sur petits domaines à l'aide de différentes distributions *a priori* pour les composantes de variance

La modélisation hiérarchique bayésienne est très populaire dans l'estimation sur petits domaines, et la spécification de distributions *a priori* est très importante dans l'approche de la modélisation hiérarchique

bayésienne. Dans le cadre du projet, nous étudions l'incidence des distributions *a priori* sur l'estimation sur petits domaines selon les modèles hiérarchiques bayésiens de You et Chapman (2006), de Sugasawa, Tamae et Kubokawa (2017) et de You (2021). En particulier, nous étudierons l'utilisation d'une distribution *a priori* non hiérarchique et de distributions *a priori* gamma inverses pour les composantes de variance au moyen d'une étude par simulations et d'une analyse sur données réelles.

Progrès :

Nous avons étudié deux spécifications de distributions *a priori* pour les composantes de variance dans les modèles hiérarchiques bayésiens de You et Chapman (2006) et de You (2021). Nous avons mené une étude par simulations et appliqué les modèles aux données de l'EPA. Nos résultats révèlent que la distribution *a priori* gamma inverse dans les modèles hiérarchiques bayésiens fonctionne très bien et peut donner de meilleurs résultats que la distribution *a priori* plate. Un document de recherche (You, 2022) a été rédigé.

Pour obtenir plus de renseignements, communiquez avec :
Yong You (613-863-9263, yong.you@statcan.gc.ca).

Bibliographie

Sugasawa, S., Tamae, H. et Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

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. (2022). An empirical study of hierarchical Bayes small area estimators using different priors on model variances. Document de recherche interne, Statistique Canada, Ottawa.

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.

SOUS-PROJET : Méthodes de lissage de la variance due à l'échantillonnage pour l'estimation sur petits domaines

Le lissage de la variance due à l'échantillonnage est un sujet important dans l'estimation sur petits domaines. Dans le cadre du projet, nous étudierons différentes méthodes de lissage de la variance due à l'échantillonnage, y compris l'utilisation des effets de plan et des fonctions de variance généralisée pour l'estimation de proportions sur petits domaines. Les méthodes de lissage proposées peuvent être utilisées comme approche type et permettre de simplifier la procédure de lissage pour l'estimation sur petits domaines fondée sur un modèle.

Progrès :

Nous avons étendu le travail de You et Hidiroglou (2012) et proposé différentes méthodes de lissage de la variance due à l'échantillonnage pour l'estimation des proportions sur petits domaines. Nous avons

appliqué les méthodes de lissage proposées à différentes données d'enquête, y compris celles de l'Enquête sur la santé dans les collectivités canadiennes et de l'Enquête sur la participation et les limitations d'activités. Nous avons également appliqué les méthodes de lissage proposées aux données de l'Enquête sur la population active pour l'estimation du taux de chômage, et nous avons comparé les résultats avec ceux de la méthode proposée par Hidiroglou, Beaumont et Yung (2019). Un document de recherche (You et Hidiroglou, 2022) a été rédigé. Une étude par simulations sera réalisée, et nous prévoyons présenter le document à une revue à comité de lecture en vue d'une publication éventuelle.

Pour obtenir plus de renseignements, communiquez avec :
Yong You (613-863-9263, yong.you@statcan.gc.ca).

Bibliographie

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2019001/article/00009-fra.pdf.

You, Y., et Hidiroglou, M. (2012). Sampling variance smoothing methods for small area proportion estimators. Document de travail de la Direction générale de la méthodologie, SRID-2012-08E, Statistique Canada, Ottawa, Canada.

You, Y., et Hidiroglou, M. (2022). Application of sampling variance smoothing methods for small area proportion estimation. Document de recherche interne, Statistique Canada, Ottawa.

SOUS-PROJET : Estimation de l'erreur quadratique moyenne fondée sur le plan de sondage dans l'estimation sur petits domaines

L'utilisation du modèle de Fay-Herriot pour produire des estimations sur petits domaines a augmenté à Statistique Canada au cours des cinq dernières années. Ces estimations sont généralement accompagnées d'estimations de l'erreur quadratique moyenne (EQM) fondée sur le modèle. Les utilisateurs sont toutefois habitués aux estimations de l'EQM fondée sur le plan de sondage. Par rapport à l'EQM fondée sur le modèle, l'EQM fondée sur le plan de sondage a l'avantage de ne pas éliminer par intégration la spécificité d'un domaine particulier, et elle peut présenter un plus grand intérêt pour les utilisateurs comme indicateur de la qualité des estimations. Les estimations fondées sur le plan de l'EQM fondée sur le plan sont réputées instables (voir par exemple Rao, Rubin-Bleuer et Estevao, 2018). Nous envisageons d'étudier l'utilisation d'une approche conditionnelle pour obtenir un estimateur plus efficace de l'EQM fondée sur le plan de sondage.

Progrès :

Dans le cadre d'études antérieures, des estimateurs de l'EQM fondée sur le plan de sondage à partir d'une approche conditionnelle ont été élaborés. Cette année, une étude de simulation a permis d'évaluer plusieurs estimateurs de l'EQM fondée sur le plan de sondage, et une théorie additionnelle permettant de mieux expliquer les propriétés empiriques a été élaborée. La principale conclusion de nos travaux de recherche est qu'il ne semble pas possible d'estimer avec exactitude l'EQM fondée sur le plan. Nous prévoyons terminer nos études et résumer nos constatations dans un rapport.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Mesure de l'incertitude associée aux estimateurs pour petits domaines basés sur un modèle. *Techniques d'enquête*, 44, 2, 163-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2018002/article/54958-fra.pdf.

2 Méthodes et applications de la science des données

SOUS-PROJET : Techniques d'apprentissage automatique permettant de traiter la non-réponse aux questions des enquêtes

La non-réponse aux questions des enquêtes est un problème courant dans la production de statistiques officielles. En particulier, si la probabilité de non-réponse est corrélée à une variable d'intérêt, elle peut introduire un biais important dans les estimations finales si le biais n'est pas corrigé. Une façon courante de corriger le biais dû à la non-réponse est de modéliser la probabilité de réponse. En estimant la probabilité de réponse pour toutes les unités de l'enquête, nous pouvons ajuster le poids de chaque unité afin de compenser les non-répondants. Nous pouvons intégrer la robustesse au processus en regroupant les unités en groupes de réponses homogènes afin d'éviter les ajustements de poids extrêmes (p. ex. Gelein, Haziza et Causeur, 2018).

À l'heure actuelle, la plupart des modèles de non-réponse aux questions des enquêtes reposent sur la régression logistique (suivie de la création de groupes homogènes) ou des arbres de décision. L'objectif du projet était d'étudier l'application de méthodes d'apprentissage automatique plus complexes pour modéliser la probabilité de réponse à une enquête.

Progrès :

Dans le monde réel, la population peut être assez hétérogène pour ce qui est des probabilités de réponse. Le mécanisme sous-jacent pourrait être très difficile à modéliser au moyen de modèles simples comme la régression logistique. L'une des principales constatations du projet est qu'il est possible d'utiliser des techniques d'intelligibilité pour disséquer des modèles d'apprentissage automatique complexes et visualiser la structure des données sous-jacentes.

Dans le cadre de notre étude, nous avons créé un ensemble de données synthétiques en générant des réponses avec différentes probabilités pour différentes sous-populations. Dans un premier temps, nous avons utilisé un modèle complexe, mais efficace, d'apprentissage automatique d'optimisation (XGBoost) pour estimer la probabilité de réponse dans l'ensemble de données synthétique (Chen et Guestrin, 2016). Nous avons ensuite eu recours à une technique utilisée pour faire des prédictions de modèles d'apprentissage automatique complexes (« boîte noire ») interprétables par les humains, la technique LIME (Ribeiro, 2016), pour récupérer la majeure partie de la structure de sous-population correspondant aux différents mécanismes de non-réponse. La technique LIME permet d'ajuster un modèle linéaire local dans le voisinage de chaque point de données en utilisant les valeurs prédites du modèle complexe. Les coefficients des modèles linéaires peuvent ensuite être regroupés pour définir les tendances de non-

réponse (Van der Maaten et Hinton, 2008; Ribeiro, Singh et Guestrin, 2016; et Wattenberg, Viégas et Johnson, 2016). Étant donné que l'ensemble de données est synthétique, la véritable structure de la sous-population est connue, et la comparaison des grappes de coefficients établies au moyen de la technique LIME avec les groupes de sous-populations réels a montré une bonne cohérence.

Cette technique pourrait représenter une nouvelle façon prometteuse de former des groupes de réponse homogènes en combinant des algorithmes d'apprentissage automatique complexes et des techniques d'intelligibilité pour les populations dont la structure de non-réponse est hétérogène.

Pour obtenir plus de renseignements, communiquez avec :
Jeffery Zhang (343-551-1318, jeffery.zhang@statcan.gc.ca).

Bibliographie

Gelein, B., Haziza, D. et Causeur, D. (2018). Propensity Weighting for Survey Nonresponse Through Machine Learning. Dans *Journées de méthodologie statistique*, INSEE, Paris, 12 au 14 juin 2018.

Chen, T., et Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv: 1603.02754v3.

Ribeiro, M.T. (2016). LIME – Local Interpretable Model-Agnostic Explanations. Extrait de https://homes.cs.washington.edu/~marcotcr/blog/lime/.

Van der Maaten, L., et Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605.

Ribeiro, M. T., Singh, S. et Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv: 1602.04938v3.

Wattenberg, M., Viégas, F. et Johnson, I. (2016). How to use t-SNE effectively. *Distill*.

SOUS-PROJET : Apprentissage automatique quantique pour la classification de textes

L'informatique quantique promet de révolutionner la science des données grâce à l'émergence de l'apprentissage automatique quantique. Nous avons entrepris un projet de recherche pour étudier la faisabilité d'appliquer l'apprentissage automatique quantique aux tâches de classification de textes à Statistique Canada; le projet a été réalisé en collaboration avec l'Université de Sherbrooke et la Banque du Canada, et il a été financé par le Conseil de recherche et développement de Statistique Canada.

Progrès :

Nous avons mis en œuvre et testé trois méthodes différentes : le classificateur quantique variationnel, la méthode quantique fondée sur le noyau et l'apprentissage par transfert hybride quantique-classique (Laprade, Blanchette, Zanussi, Chikhar et Skavysh, 2021). En ce qui concerne l'apprentissage automatique quantique, l'une des grandes limites à l'application de la classification de textes est la grande dimensionnalité du codage des données. Il est notoire que les méthodes de classification de textes comportent de grands espaces des attributs, et bien que de nombreux partisans de l'apprentissage automatique quantique mentionnent la croissance exponentielle des circuits de bits quantiques, de tels

plongements denses sont coûteux et nécessitent des circuits profonds. Par conséquent, les techniques de réduction de la dimensionnalité étaient essentielles au succès du projet.

Même si les résultats du projet de recherche sont prometteurs, d'autres études sont nécessaires pour que nous puissions appliquer de telles méthodes à des cas d'utilisation typiques pour la classification de textes au sein de l'organisme.

Pour obtenir plus de renseignements, communiquez avec :
Saeid Molladavoudi (613-290-7418, saeid.molladavoudi@statcan.gc.ca).

Bibliographie

Laprade, J.-F., Blanchette, S., Zanussi, Z., Chikhar, O. et Skavysh, V. (2021). Quantum machine learning for text classification. Présentation d'affiche au 2021 Montreal AI Symposium.

SOUS-PROJET : Amélioration du rendement de la reconnaissance optique de caractères grâce au prétraitement des images

De nombreux projets de la Division de la science des données portent sur l'extraction de l'information. L'extraction de l'information vise l'extraction automatique de renseignements structurés de documents non structurés ou semi-structurés. L'un des objectifs de la Division de la science des données est que d'autres divisions utilisent des techniques de science des données pour améliorer leurs processus et obtenir plus de renseignements à partir de leurs données. L'extraction de l'information est un moyen permettant de réduire le temps de traitement nécessaire pour extraire l'information pertinente des documents, de réduire au minimum le travail manuel et d'accroître l'efficacité. Le degré de difficulté de l'extraction des données dépend de l'état des documents. Pour les documents ou les images numérisés, la reconnaissance optique de caractères (ROC) est requise. Les moteurs de ROC sont utilisés pour convertir des images textuelles en fichiers modifiables. Pour ce qui est d'améliorer le rendement des pipelines d'extraction de l'information pour les images numérisées, le rendement de la ROC joue un rôle important.

L'objectif du projet de recherche est d'étudier des techniques de pointe pour améliorer le rendement des moteurs de ROC au moyen de techniques d'amélioration du prétraitement des images et, plus particulièrement, de la modélisation au moyen de la binarisation d'images. La binarisation d'images est le processus de conversion d'une image, dans le format en niveaux de gris, en une image en noir et blanc.

Progrès :

Une revue de littérature a été effectuée, et les travaux d'Ayantha, Nilanjan, Xiao et Allegra (2021) ont révélé des résultats prometteurs. Les techniques et les méthodes proposées dans leur document ont été utilisées pour différents ensembles de données et comparées à d'autres méthodes de pointe, et elles ont été jugées supérieures en fonction de l'exactitude et du taux d'erreur des caractères. De plus, une légère amélioration à la méthode a été proposée. Le projet de recherche est terminé. Nous espérons pouvoir utiliser les résultats du projet de recherche dans de futurs projets de la Division de la science des données pour des enquêtes comportant des données de réception numérisées. Les prochaines étapes de ces travaux sont à l'étude. Voici la liste des tâches accomplies :

- Une revue de littérature a été effectuée.
- Les résultats de Ayantha et coll. (2021) ont été reproduits de près.

- Le pipeline de formation prévu dans le document de recherche a été modifié en fonction de nouvelles observations.
- Un modèle de binarisation a été entraîné à partir d'un nouvel ensemble de données.
- Un rapport sur les travaux réalisés dans le cadre du projet a été produit.
- Un rapport contenant le code et sa documentation a été créé sur Gitlab.

Pour obtenir plus de renseignements, communiquez avec :
Oladayo Ogunnoiki (289-489-1239, oladayo.ogunnoiki@statcan.gc.ca).

Bibliographie

Ayantha, R., Nilanjan, R., Xiao, X. et Allegra, L. (2021). Unknown-Box Approximation to Improve Optical Character Recognition Performance. Dans *ICDAR,* (1), 481-496.

SOUS-PROJET : Approche participative de préservation de la confidentialité pour la détection de la cyberintimidation

En collaboration avec le Centre de l'intégration et du développement des données sociales et le Secrétariat de l'éthique des données du Centre de coopération internationale et d'innovation en méthodologie, nous avons étudié la capacité d'entraîner des modèles d'apprentissage automatique à partir des données des clients sans jamais visualiser ou recueillir les données, en utilisant une technique d'apprentissage automatique distribué appelée « apprentissage fédéré ». L'apprentissage fédéré permet à un modèle d'apprentissage automatique centralisé conservé par une autorité centrale, comme Statistique Canada, d'être entraîné à une tâche précise dans un domaine spécialisé précis sans que l'autorité centrale n'ait besoin d'accéder aux données de formation ou de les conserver. L'hôte du modèle d'apprentissage automatique centralisé reçoit uniquement les pondérations numériques des modèles d'apprentissage automatique locaux qui sont entraînés à partir des données distribuées des clients conservées sur les appareils des clients. En agrégeant les pondérations obtenues, l'autorité centrale met à jour son modèle d'apprentissage automatique central avec les changements effectués par les clients sur leurs appareils.

Ce projet de validation de la technologie vise à étudier le fonctionnement de l'apprentissage fédéré, la façon dont il peut être appliqué dans un contexte d'approche participative et la faisabilité de son application dans des contextes de production. Grâce aux connaissances acquises, Statistique Canada aura une meilleure compréhension de la technique en évolution et des possibilités offertes par l'application.

Progrès :

Après avoir effectué des essais simulés à l'aide d'un ensemble de données sur la cyberintimidation accessible au public, nous avons démontré que l'approche peut fonctionner efficacement dans le cadre de différentes approches participatives, si nous portons une attention particulière lorsque nous travaillons avec des données non étiquetées. Nous avons étudié deux cadres pour évaluer la façon dont l'apprentissage fédéré peut être utilisé pour les données sur la cyberintimidation dans un contexte d'approche participative. Le premier, un cadre d'annotateurs, révèle que l'utilisation de l'apprentissage fédéré avec un ensemble d'annotateurs de données fiables peut permettre à Statistique Canada d'entraîner les données annotées tout en conservant les données du côté de l'annotateur. Le deuxième cadre, un cadre d'apprentissage fédéré semi-supervisé, permet d'utiliser des données non étiquetées pour la formation, et ces données demeurent sur les appareils des clients. Une approche d'apprentissage

fédéré semi-supervisé de base a été appliquée au cadre avec un succès limité, mais il s'est avéré que cette approche peut être mise en œuvre.

L'apprentissage fédéré est donc un outil prometteur qui devrait continuer d'être étudié en vue de son utilisation dans les systèmes de production à mesure que le concept continue d'évoluer. À la lumière de telles constatations, nous recommandons, comme prochaines étapes, d'étudier davantage l'apprentissage fédéré en la combinant à différentes technologies d'amélioration de la confidentialité (comme la confidentialité différentielle et le chiffrement homomorphique) en vue d'accroître davantage la confidentialité des données des utilisateurs, et en appliquant l'apprentissage fédéré à une page Web fictive déployée ou à une application mobile afin d'élaborer un projet étendu de validation de la technologie sur la façon dont la technique peut être appliquée à un exercice réaliste d'approche participative. La mise à l'essai plus poussée de l'approche peut mener à des collaborations qui permettront de concevoir des modèles d'apprentissage automatique robustes comportant des données sensibles et privées distribuées entre plusieurs organisations différentes.

Pour obtenir plus de renseignements, communiquez avec :
Benjamin Santos (438-459-7721, benjamin.santos@statcan.gc.ca) ou
Julian Templeton (julian.templeton@statcan.gc.ca).

SOUS-PROJET : Revue de littérature sur les itinéraires efficaces pour les activités de dénombrement

Les activités de dénombrement du recensement et de l'Enquête sur la population active qui se déroulent en personne exigent des déplacements dans de grandes régions géographiques (unités de collecte), qui sont divisées en îlots (îlots de collecte). La recherche de l'itinéraire le plus efficace dans ces îlots est un processus à étapes multiples, et sa complexité varie en fonction du nombre d'îlots à parcourir. L'ordre des îlots idéal représente un chemin hamiltonien, connu pour être un problème informatique difficile qui est NP complet.

Nous avons effectué une revue de littérature sur les progrès réalisés au chapitre de l'apprentissage automatique et des approches de l'informatique quantique pour trouver les chemins hamiltoniens, particulièrement en ce qui concerne la mise en œuvre de l'apprentissage par renforcement profond et de la programmation dynamique des processus décisionnels de Markov pour l'apprentissage automatique ainsi que la mise en œuvre du problème de satisfaction contraint pour l'informatique quantique (Montanaro, 2018; Campbell, Khurana et Montanaro, 2019).

Progrès :

En fin de compte, la recherche n'a abouti à aucune solution particulière qui pourrait être facilement déployée et permettrait de résoudre le cas d'utilisation du recensement et de l'Enquête sur la population active. Il faudrait mener d'autres expériences pour déterminer le degré de réussite possible d'une approche d'apprentissage automatique ou quantique, et il est possible qu'un ensemble hybride d'approches donne les meilleurs résultats.

Pour obtenir plus de renseignements, communiquez avec :
Reginald Maltais (613-612-9438, reginald.maltais@statcan.gc.ca).

Bibliographie

Montanaro, A. (2018). Quantum-walk speedup of backtracking algorithms. *Theory of Computing*, 14, 1-24.

Campbell, E., Khurana, A. et Montanaro, A. (2019). Applying quantum algorithms to constraint satisfaction problems. *Quantum*, 3, 167.

SOUS-PROJET : Revue de littérature sur l'intersection d'ensembles privés

Les organismes statistiques nationaux, comme Statistique Canada, procèdent souvent au couplage d'enregistrements entre des ensembles de données afin d'améliorer la valeur analytique des données disponibles. Toutefois, dans de nombreux cas, un ensemble de données ou les deux ensembles sont considérés comme sensibles, ce qui se traduit par des frais généraux juridiques et administratifs supplémentaires importants associés au couplage. Le couplage d'enregistrements préservant la confidentialité pourrait être une option pour que le couplage d'enregistrements puisse être effectué sans ces frais généraux.

Dans le cadre des travaux, nous avons effectué une revue de littérature sur le couplage d'enregistrements préservant la confidentialité dans le contexte de Statistique Canada. Diverses techniques d'appariement exact, comme les méthodes par hachage, le chiffrement totalement homomorphique, le transfert inconscient, les fonctions pseudo-aléatoires inconscientes et le calcul multipartite sécurisé, ont été étudiées dans le contexte du couplage d'enregistrements préservant la confidentialité entre deux parties.

Progrès :

Dans le cadre des travaux, nous avons décrit un certain nombre de méthodes relatives au couplage d'enregistrements préservant la confidentialité à Statistique Canada. Il s'agit d'un domaine en expansion rapide qui pourrait avoir des répercussions futures sur les travaux portant sur les couplages et l'intégration de données (Zanussi et Dugdale, 2022). D'autres études sont nécessaires pour que ces protocoles puissent être appliqués aux cas d'utilisation typiques au sein de l'organisme.

Pour obtenir plus de renseignements, communiquez avec :
Saeid Molladavoudi (613-290-7418, saeid.molladavoudi@statcan.gc.ca).

Bibliographie

Zanussi, Z., et Dugdale, C. (2022). *Practical Privacy-Aware Data Linkage and Statistical Aggregation based on Privacy Enhancing Techniques*. Rapport interne soumis pour publication, Statistique Canada, Ottawa.

SOUS-PROJET : Revue de littérature sur l'apprentissage automatique automatisé

Le principal objectif du projet était d'effectuer une revue de littérature et d'en apprendre davantage sur la façon dont la collectivité de l'apprentissage automatique définit et comprend l'apprentissage automatique automatisé et sur les moyens qui devraient être pris pour intégrer un tel paradigme aux pratiques d'apprentissage automatique de Statistique Canada.

Progrès :

Dans le cadre de la revue de littérature, nous avons examiné les différentes perspectives et approches en matière d'apprentissage automatique automatisé dans le but de répondre aux questions suivantes : Quelles composantes du processus d'apprentissage automatique sont habituellement automatisées, et quels sont les outils et les services populaires courants disponibles pour l'apprentissage automatique automatisé au moyen de licences ou d'accès aux codes sources libres ? De plus, quelques fournisseurs de services et de progiciels en matière d'apprentissage automatique automatisé ont été interrogés et comparés. Le produit livrable définitif était un document interne. Il s'agit de la première phase de l'étude de la Division de la science des données sur l'apprentissage automatique automatisé; la prochaine phase sera un projet pratique dans le cadre duquel nous réaliserons une étude empirique pour évaluer les avantages d'une approche d'apprentissage automatique automatisé.

Pour obtenir plus de renseignements, communiquez avec :
Loïc Muhirwa (343-998-7756, loic.muhirwa@statcan.gc.ca).

SOUS-PROJET : Introduction à l'intelligence artificielle explicable — examen technique de méthodes locales indépendantes d'un modèle

Les principes d'explicabilité abordés dans le cadre « Utilisation responsable de l'apprentissage automatique à Statistique Canada » (Bosa, 2021) orientent l'élaboration de processus d'apprentissage automatique responsables. Il s'agit d'une source de motivation essentielle pour cet examen technique. L'apprentissage automatique responsable fait également partie de la capacité stratégique « opérationnalisation » de la Stratégie de la science des données, qui met l'accent sur l'application des modèles en production.

Les travaux donnent un aperçu des principes généraux de l'intelligence artificielle explicable. Ils donnent une description taxonomique complète des méthodes d'intelligence artificielle explicable ainsi que des définitions et des références pour les méthodes les plus populaires. Ils montrent également brièvement la façon dont fonctionnent certaines méthodes d'intelligence artificielle explicable. Nous examinons ensuite de plus près la plupart des méthodes locales populaires indépendantes d'un modèle. Nous mettons l'accent sur quatre méthodes, pour lesquelles nous présentons la théorie, les avantages et les inconvénients de chaque méthode. Ensuite, au moyen de plusieurs ensembles de données, les applications des méthodes montrent la façon dont elles pourraient être utilisées pour différents types de problèmes relatifs à la science des données. Nous concluons par un bref exposé sur la façon d'utiliser l'intelligence artificielle explicable à diverses étapes, des expériences à la production, et sur la façon de l'utiliser dans le monde réel.

Progrès :

Dans le cadre du projet, nous mettons principalement l'accent sur les méthodes indépendantes d'un modèle, qui deviennent de plus en plus attrayantes dans le domaine de l'apprentissage automatique tout au long du processus de production. Les travaux incitent la Division de la science des données à accorder encore plus d'attention à l'application de ces méthodes d'intelligence artificielle explicable dans le développement de systèmes de production.

Comme dans tout domaine de recherche brûlant, les méthodes ne cessent d'évoluer, bien qu'il soit important de souligner que diverses méthodes comportent des limites, chacune présentant ses propres

avantages et désavantages. De plus, nous tenons à souligner que l'accent mis sur les techniques d'explicabilité locale n'est qu'une première introduction à ce sujet; notre intention est de poursuivre la recherche sur ces techniques dans le cadre d'un travail à plus long terme. Nous avons hâte de mettre ces connaissances à profit dans l'élaboration de futures lignes directrices pratiques sur la sélection de l'approche d'intelligence artificielle explicable la plus appropriée pour les projets de la Division.

Pour obtenir plus de renseignements, communiquez avec :
Soufiane Fadel (343-573-7912, soufiane.fadel@statcan.gc.ca).

Bibliographie

Bosa, K. (2021). Utilisation responsable de l'apprentissage automatique à Statistique Canada. Extrait de https://www.statcan.gc.ca/fr/data-science/network/machine-learning.

3 Études sur la non-réponse aux enquêtes

SOUS-PROJET : Analyse comportementale de la non-réponse aux programmes statistiques

À l'instar d'autres agences statistiques nationales à travers le monde, Statistique Canada a vu la participation à ses programmes statistiques diminuer de façon significative au cours des dernières années. Les raisons expliquant ce phénomène sont nombreuses et il n'existe pas de profil unique des personnes refusant de participer aux programmes statistiques. L'objectif de ce projet consiste à mener des groupes de discussion et des entrevues individuelles ciblés auprès de Canadiens et de Canadiennes afin de mieux comprendre la décision de ne pas participer aux enquêtes de Statistique Canada.

Progrès :

Nous avons développé un canevas d'entrevue articulé autour de la théorie de l'action raisonnée (Azjen et Fishbein, 1980), afin de bien cerner les facteurs menant à la décision de ne pas participer à une enquête statistique. Nous avons utilisé ce canevas pour mener huit groupes de discussion réunissant entre six et huit participants chacun, ainsi que sept entrevues individuelles. Le tout a permis de mieux comprendre les croyances, attitudes et normes ayant une influence sur la participation aux programmes statistiques, de faire des recommandations et de dégager des pistes de solutions, qui seront explorées au cours des prochaines années.

Le tout a également permis de peaufiner les connaissances et pratiques en matière de groupe de discussion menés virtuellement. Tremblay, Dean et Martineau (2022) ont publié un article présentant un sommaire des défis rencontrés et des nouvelles opportunités offertes par les groupes de discussion menés virtuellement.

Pour obtenir plus de renseignements, communiquez avec :
Patrice Martineau (613-219-5899, patrice.martineau@statcan.gc.ca).

Bibliographie

Ajzen, I., et Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice Hall.

Tremblay, L., Dean, S. et Martineau, P. (2022). Conducting Online Focus Groups: Challenges and Opportunities. SAGE Research Methods: Doing Research Online.

SOUS-PROJET : Suivi des non-répondants aux enquêtes auprès des entreprises

Au cours des deux dernières décennies, les taux de réponse aux enquêtes ont diminué régulièrement. Dans un tel contexte, il devient de plus en plus important pour les organismes de statistique d'élaborer et d'utiliser des méthodes qui réduisent les effets néfastes de la non-réponse sur l'exactitude des estimations d'enquêtes. Le suivi des non-répondants peut être un remède efficace contre le biais de non-réponse, même s'il exige beaucoup de temps et de ressources. L'objectif du projet de recherche est de faire la lumière sur certaines questions pratiques concernant le suivi de la non-réponse. Par exemple, en supposant un budget fixe pour le suivi des non-répondants, quelle est la meilleure façon de choisir les unités devant faire l'objet d'un suivi ? Combien d'effort faudrait-il consacrer au suivi répété des non-répondants pour qu'une réponse soit reçue ? Faudrait-il faire le suivi de tous les non-répondants ou seulement d'un échantillon d'entre eux ? Si le suivi est fait auprès d'un échantillon, de quelle manière celui-ci devrait-il être sélectionné ?

Progrès :

Une étude par simulations a déjà été réalisée pour permettre de répondre à ces questions. Nous avons comparé les biais relatifs Monte Carlo et les racines des erreurs quadratiques moyennes relatives selon divers plans de sondage de suivi, tailles d'échantillon et scénarios de non-réponse. Nous avons également déterminé la taille minimale de l'échantillon de suivi nécessaire pour que le budget soit dépensé, en moyenne, et nous avons montré qu'elle maximise le taux de réponse attendu. Une des principales conclusions de notre simulation est que cette taille de l'échantillon minimale semble aussi minimiser approximativement le biais et l'erreur quadratique moyenne des estimations.

Pendant l'année en cours, nous avons révisé un document, qui a ensuite été accepté et publié dans le numéro de juin 2022 de *Techniques d'enquête* (Neusy, Beaumont, Yung, Hidiroglou et Haziza, 2022).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Neusy, E., Beaumont, J.-F., Yung, W., Hidoroglou, 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.

SOUS-PROJET : Initiative de suivi de la non-réponse

L'initiative de suivi de la non-réponse recueillera des données sur les non-répondants des principaux programmes de la statistique sociale de Statistique Canada. Les buts de cette initiative sont de faciliter

l'analyse afin de comparer les répondants aux non-répondants et de mesurer l'ampleur de l'erreur de non-réponse. Dans le contexte de taux de réponse faibles et généralement en baisse, ainsi que du plan d'action sur les données désagrégées, des mesures pour détecter l'impact potentiel sur des sous-populations plus petites et spécifiques doivent être mises en œuvre.

Les trois étapes consistent en une preuve de concept, deux pilotes et enfin la mise en place du programme permanent comme tel. À chaque étape, l'approche scientifique est utilisée en identifiant la portée, en formulant et en testant des hypothèses, pour finalement documenter et communiquer les résultats pour alimenter les prochaines itérations de développement du programme. Les plans ont été présentés à la conférence annuelle 2022 de la Société statistique du Canada (Wright, Brisebois et Martineau, 2022). Le programme est conçu pour contribuer au résultat ultime de la production de renseignements statistiques de haute qualité pour les Canadiens.

Progrès :

Pour l'enquête de preuve de concept, l'équipe de gestion de l'enquête a rédigé le questionnaire et a reçu des commentaires pour en améliorer son contenu. Le plan d'échantillonnage a été conçu et le calendrier de production pour préparer l'enquête et recueillir les données a été élaboré. Pour faciliter la participation, divers modes de collecte seront offerts, incluant des entrevues téléphoniques assistées par ordinateur, des questionnaires papier à retourner par la poste et des questionnaires électroniques sur le Web.

L'enquête de preuve de concept sera en collecte au cours de l'année fiscale 2022-2023. Également au cours de l'année fiscale 2022-2023, la planification, la préparation et la mise en œuvre de l'enquête pilote auront lieu.

Pour obtenir plus de renseignements, communiquez avec :
François Brisebois (613-222-8310, francois.brisebois@statcan.gc.ca).

Bibliographie

Wright, P., Brisebois, F. et Martineau, P. (2022). Improving response by studying citizen participation in social surveys. Communication présentée à la conférence annuelle de la Société statistique du Canada, juin 2022.

4 Problèmes d'estimation dans les enquêtes

SOUS-PROJET : Estimation de la variance bootstrap pour un échantillonnage à plusieurs degrés avec application à la non-réponse

Le bootstrap sert fréquemment à l'estimation de la variance dans les enquêtes avec un plan de sondage stratifié à plusieurs degrés. Il est souvent mis en œuvre par la production d'un ensemble de poids bootstrap qui est mis à la disposition des utilisateurs et qui tient compte de la complexité du plan de sondage. La méthode de Rao, Wu et Yue (1992) sert dans bien des cas à produire les poids bootstrap requis. Elle est valide pour un échantillonnage stratifié avec remise au premier degré ou un échantillonnage de taille fixe sans remise si les fractions de sondage au premier degré sont négligeables.

Certaines enquêtes reposent sur des plans qui ne satisfont pas à ces conditions. Le but du projet était de proposer une méthode bootstrap qui tient compte de cette limite des poids bootstrap de Rao, Wu et Yue (1992).

Progrès :

Au cours des dernières années, nous avons élaboré une méthode bootstrap simple et unifiée qui s'applique à tout plan de sondage stratifié à plusieurs degrés dans la mesure où des poids bootstrap valides peuvent être produits pour chaque degré distinct d'échantillonnage. Notre méthode s'applique aussi aux plans de sondage à deux phases à condition que l'échantillonnage de poisson soit utilisé à la deuxième phase. Nous utilisons ce plan pour modéliser la non-réponse aux enquêtes et dériver des poids bootstrap qui tiennent compte de la pondération pour la non-réponse.

Pendant l'année en cours, nous avons mené trois études de simulation limitées pour évaluer les propriétés de notre méthode bootstrap, et nous avons terminé la rédaction d'un article qui a été accepté et publié dans un numéro spécial de *Stats* sur les méthodes de rééchantillonnage (Beaumont et Émond, 2022).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Beaumont, J. F., et Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5, 339–357.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 2, 225-234. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1992002/article/14486-fra.pdf.

SOUS-PROJET : Estimation bootstrap du biais conditionnel pour mesurer l'influence dans les enquêtes complexes

Dans les enquêtes qui visent à recueillir des données sur des variables asymétriques, il est souvent souhaitable d'évaluer l'influence des unités de l'échantillon sur l'erreur d'échantillonnage des estimateurs pondérés des paramètres de population finie. Le biais conditionnel est une mesure attrayante de l'influence qui tient compte du plan de sondage et de la méthode d'estimation. Il se définit comme l'espérance par rapport au plan de sondage de l'erreur d'échantillonnage conditionnellement à ce qu'une unité donnée soit sélectionnée dans l'échantillon. L'estimation de ce biais conditionnel est relativement simple pour les plans de sondage et les estimateurs simples. Pour les plans ou les estimateurs complexes, en revanche, il peut être fastidieux de dégager une expression explicite du biais conditionnel. Dans ces enquêtes complexes, l'estimation de la variance s'obtient fréquemment par des méthodes de répliques comme le bootstrap. Les méthodes bootstrap d'estimation de la variance s'appliquent normalement par la production d'un ensemble de poids bootstrap qui est mis à la disposition des utilisateurs avec les données d'enquête. Dans le projet, notre objectif était de montrer la façon d'utiliser ces poids bootstrap pour obtenir un estimateur du biais conditionnel. Cet estimateur peut alors servir à construire des estimateurs robustes des paramètres de population finie, qui sont moins négativement affectés par les unités influentes que les estimateurs pondérés ordinaires.

Progrès :

La théorie a été élaborée dans des travaux de recherche antérieurs, et notre estimateur bootstrap a été évalué dans une étude par simulations. Cette année, nous avons révisé un article et ajouté une illustration à l'aide de données tirées de l'Enquête canadienne sur les dépenses des ménages. Notre article a été accepté et publié en ligne dans le *Journal of Survey Statistics and Methodology* (Beaumont, Bocci et St-Louis, 2021).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., Bocci, C. et St-Louis, M. (2021). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. *Journal of Survey Statistics and Methodology*.

SOUS-PROJET : Élaboration d'un prototype de système pour réaliser une estimation robuste

Dans de nombreuses enquêtes économiques et dans quelques enquêtes sociales, des variables aux distributions asymétriques sont recueillies, ce qui peut donner lieu à la présence de valeurs aberrantes et d'unités influentes dans l'échantillon. Les méthodes classiques d'estimation peuvent produire des estimateurs hautement inefficaces dans de tels scénarios. L'objectif de l'estimation robuste est de réduire l'effet des unités influentes de l'échantillon sur les estimations. Le biais conditionnel est utilisé comme mesure de l'influence des unités de l'échantillon. Les estimations classiques sont réduites par une fonction du biais conditionnel des unités de l'échantillon. La notion de biais conditionnel a d'abord été avancée par Moreno-Rebollo, Munoz-Reyez et Munoz-Pichardo (1999); elle a ensuite été utilisée par Beaumont, Haziza et Ruiz-Gazen (2013) pour concevoir un estimateur robuste. Ce travail est pertinent pour un grand nombre d'enquêtes économiques et sociales de Statistique Canada.

Progrès :

Un prototype SAS efficace incluant les spécifications d'estimation robuste formulées dans Beaumont (2017) a été mis en œuvre et évalué. Il s'agit d'un groupe de neuf macros pour les diverses fonctions liées à la production d'estimations de totaux, de ratios et de moyennes d'un domaine. Des renseignements auxiliaires peuvent être utilisés dans de nombreuses fonctions pour accroître l'efficacité des estimations.

Le prototype comprend une macro pour le calcul d'estimations classiques et d'estimations robustes pour les domaines. Les estimations robustes au niveau des domaines n'ont pas la propriété d'additivité des estimations classiques. Il y a donc une autre macro qui crée des estimations cohérentes au niveau des domaines à partir des estimations robustes en modifiant légèrement leurs valeurs. Une autre macro produit ensuite des poids par recalage pour les unités de l'échantillon qui reproduisent les estimations cohérentes au niveau des domaines et les totaux connus de variables auxiliaires.

Une méthode bootstrap pour l'estimation de la variance a été incluse pour produire des estimations de la variance pour les estimations robustes au niveau des domaines.

Un guide de l'utilisateur dans lequel figurent des exemples exposant les neuf fonctions du prototype est disponible (voir Estevao, 2022). Chaque macro du prototype effectue une validation complète des entrées

spécifiées; elle affiche des renseignements sur les erreurs et les problèmes potentiels liés aux ensembles de données d'entrée et aux options sélectionnées.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca) ou
Victor Estevao (613-863-9038, victor.estevao@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, J.M. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Beaumont, J.-F. (2017). *Robust Estimation Prototype, Methodology Specifications*. Rapport interne, Statistique Canada.

Estevao, V.M. (2022). *Robust Estimation – Parameter Description and User Guide, Methodology Specifications*. Document interne, Statistique Canada.

Moreno-Rebollo, J.L., Munoz-Reyez, A.M. et Munoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.

SOUS-PROJET : Degrés de liberté liés à l'estimation du questionnaire long du recensement

Récemment, Statistique Canada a décidé de publier des intervalles de confiance afin d'exprimer la qualité des estimations. En autant que la couverture annoncée est respectée, la longueur d'un intervalle de confiance témoigne de la qualité de l'estimation. 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 (règle du pouce). Il est connu que, dans le cas de petits domaines, cette règle approximative n'estime pas adéquatement le nombre de degrés de liberté réel. Ceci engendre des intervalles de confiance qui ne donnent pas toujours la couverture attendue de 95 %.

Dans ce projet, on cherche à savoir dans quelle mesure la couverture des intervalles de confiance peut être améliorée par l'utilisation d'un nombre de degrés de liberté approprié. Celui-ci est obtenu en utilisant l'approximation de Satterthwaite. Le projet se situe dans le contexte de l'estimation du questionnaire long du recensement de la population. Dans ce cas, la variance est estimée selon la méthode « *Partial Balanced Replicated Replications epsilon* » (PBRR-epsilon) telle que décrite par Devin et Verret (2016). De plus, le projet cherche à obtenir une meilleure compréhension des facteurs qui influencent le nombre de degrés de liberté dans le but d'améliorer la règle approximative.

Progrès :

Une formule des degrés de liberté a été obtenue dans un contexte d'estimation d'un total sur un domaine du questionnaire long du recensement de la population. On a considéré un plan aléatoire stratifié avec une fraction de sondage d'un sur quatre, dont les unités primaires d'échantillonnage sont les ménages et lorsque la variance est estimée par la méthode PBRR-epsilon. La formule des degrés de liberté nous apprend que ceux-ci sont influencés par le coefficient d'aplatissement (*Kurtosis*) et la variance des totaux des ménages sur le domaine. Une étude par simulations a été réalisée pour l'estimation du total de variables continues et du total de variables dichotomiques (effectifs). Le but étant de comparer la

couverture résultante des intervalles de confiance calculés avec la règle approximative et avec la formule explicite. Deux types d'intervalles de confiance, soient de Student dans le cas continu et Wilson modifié dans le cas discret, ont été considérés. Les résultats suggèrent qu'en utilisant le nombre de degrés de liberté adéquat, la couverture est rehaussée et on atteint souvent le seuil nominal dans le cas problématique de petits domaines. L'écriture d'un article en vue d'une soumission à un journal scientifique a déjà été entreprise et se conclura dans l'année financière 2022-2023.

Pour obtenir plus de renseignements, communiquez 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, Virginie.

SOUS-PROJET : Une revue des approches d'inférence fréquentistes et bayésiennes pour des données d'enquête

Dans les enquêtes par sondage, les données sont obtenues sur un sous-ensemble d'une population finie, habituellement par des procédures d'échantillonnage probabiliste. Ces données permettent de calculer des estimations ponctuelles des paramètres de la population finie ainsi que les estimations de la variance et les intervalles de confiance associés. Les méthodes permettant de faire des inférences et d'évaluer les propriétés des procédures d'échantillonnage et d'estimation ont fait l'objet de discussions et de débats dans la seconde moitié du 20^{e} siècle. Dans le cadre de ce projet, nous cherchons à produire une revue critique de trois approches inférentielles dans un contexte de population finie : l'approche fondée sur le plan de sondage, l'approche fréquentiste fondée sur un modèle et l'approche bayésienne.

Progrès :

Nous avons rédigé un article qui a été accepté dans un numéro spécial célébrant le 50^{e} anniversaire de la *Revue canadienne de la statistique* (Beaumont et Haziza, 2022).

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., et Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and Bayesian approaches. Accepté pour publication dans *The Canadian Journal of Statistics.*

SOUS-PROJET : Sondage indirect double de type plusieurs à un et sondage indirect simple

Dans les enquêtes sociales et économiques, il peut être difficile de joindre directement des unités de la population cible, et le sondage indirect est souvent préconisé comme solution au problème. Dans le sondage indirect, l'échantillon est tiré de la base de sondage qui est liée à la population cible, et l'estimation des paramètres de la population cible est habituellement réalisée au moyen de la méthode généralisée de partage des poids. Cette méthode donne un poids, pour chaque unité de la population cible, qui dépend des poids d'échantillonnage dans la base de sondage et des poids de lien entre la population de la base de sondage et la population cible. Dans ce projet, nous mettons l'accent sur la

situation dans laquelle les unités de la population de la base de sondage sont liées à une et une seule unité de la population cible (cas « plusieurs à un »). Une telle situation se produit au service postal français, où l'échantillon est tiré d'une population d'adresses, mais les unités de la population cible correspondent à des tournées de facteurs. Nous cherchons à comprendre l'incidence des poids de lien sur l'efficacité des estimateurs obtenus par la méthode généralisée du partage des poids.

Progrès :

Nous avons dérivé des expressions de la variance et des résultats d'optimalité pour une grande classe de plans de sondage. De plus, nous avons constaté que le cas « plusieurs à un » pourrait nécessiter l'observation d'un grand nombre de liens. Nous atténuons le problème en ajoutant une population intermédiaire et un sondage indirect double. Nous avons élaboré une théorie et mené des études empiriques pour évaluer la perte de précision résultant de l'utilisation d'un sondage indirect double. Ces constatations nous aident à expliquer la perte de précision des estimateurs par la double méthode généralisée du partage des poids observée au service postal français. Nous avons rédigé un article qui présente nos résultats théoriques et empiriques (Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine et Puech, 2022b), et l'article a été accepté dans les *Annals of Applied Statistics*.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

Bibliographie

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. et Puech, P. (2022b). Many-to-one indirect sampling with application to the French postal traffic estimation. Accepté pour publication dans *The Annals of Applied Statistics*.

5 Confidentialité et protection des renseignements personnels

Les travaux de recherche sur la confidentialité à Statistique Canada continuent à mettre l'accent sur l'élaboration de nouvelles méthodes et idées qui offrent d'autres formes d'accès tout en continuant à garantir que les renseignements personnels des particuliers et des entreprises ne sont divulgués d'aucune façon. Des progrès ont été réalisés dans le cadre de trois projets décrits ci-après (voir la section 5.1 et la section 5.2). Le groupe du Centre de la confidentialité et de l'accès de Statistique Canada continue également à 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 6.9).

5.1 Méthodes axées sur la perturbation

SOUS-PROJET : Ajustement tabulaire aléatoire

L'ajustement tabulaire aléatoire est une méthode de contrôle de la divulgation qui consiste à ajouter du bruit aléatoire aux estimations plutôt que de supprimer celles-ci. Le but premier est d'éviter la suppression pour des variables continues observées dans les enquêtes économiques.

Progrès :

Un certain nombre de programmes de Statistique Canada continuent à adopter lentement la méthodologie de l'ajustement tabulaire aléatoire. Dans l'Enquête annuelle sur la recherche et le développement dans l'industrie canadienne et l'Enquête sur l'innovation et les stratégies d'entreprise, on continue à utiliser l'ajustement tabulaire aléatoire pour les produits tabulaires. Le Recensement de l'agriculture, l'Enquête annuelle sur les dépenses de protection de l'environnement et le programme PorcTRACÉ ont tous évalué l'utilisation de l'ajustement tabulaire aléatoire et prévoient l'utiliser pour leurs publications à l'avenir.

Le prototype SAS pour l'ajustement tabulaire aléatoire avec ajustements aléatoires corrélés a été mis à jour, les algorithmes d'optimisation et la gestion des données ayant été améliorés. Le prototype a également été modifié pour inclure des caractéristiques comme les poids et des garanties de sécurité appropriées pour les personnes ayant des contributions multiples. Des progrès ont également été réalisés en ce qui concerne la documentation du programme et un document de travail exposant la méthodologie utilisée. Nous poursuivons les travaux pour terminer cette phase du projet.

Pour obtenir plus de renseignements, communiquez avec :
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

5.2 Modernisation de l'accès

SOUS-PROJET : Accès du Centre de données de recherche aux fichiers des entreprises

Statistique Canada recherche des solutions d'accès et en élabore pour permettre aux chercheurs d'avoir accès à des microdonnées réelles. Il s'agit d'un scénario utile pour les chercheurs, mais qui présente un certain risque d'accès s'il n'est pas mis en œuvre de manière stratégique. Le programme du Centre de données de recherche offre un cadre sécuritaire pour les chercheurs autorisés où les résultats sont contrôlés avec soin. Des méthodes ont été élaborées pour permettre l'accès aux données sur les entreprises par l'entremise du programme du Centre de données de recherche, et une stratégie générale de contrôle des résultats a été élaborée. Le processus était auparavant géré par le programme du Centre canadien d'élaboration de données et de recherche économique.

Progrès :

Une stratégie de contrôle a été élaborée pour tenir compte des différences produites par l'accès aux données sociales qui sont présentes avec les données sur les entreprises. La stratégie comprend un cadre, des lignes directrices et des règles de contrôle. Le cadre commence par les administrateurs de données et le secteur spécialisé, et il a comme objet de fournir le plus de renseignements possible à la Division de l'accès aux données et aux analystes du Centre de données de recherche responsables du contrôle des résultats. Un questionnaire détaillé a été élaboré et il permet de saisir les subtilités et les risques de chaque fichier de données fourni au Centre de données de recherche. Les lignes directrices ont été élaborées de sorte que les analystes du Centre de données de recherche et les chercheurs qui accèdent aux fichiers opérationnels disposent des outils et des renseignements nécessaires pour contrôler adéquatement les résultats conformément aux règles de contrôle. Les règles de contrôle permettent d'évaluer les risques de chaque type de résultats demandés par les chercheurs et de veiller à ce que les risques de divulgation soient adéquatement pris en compte.

Pour obtenir plus de renseignements, communiquez avec :
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

SOUS-PROJET : Données synthétiques

La Directive sur le gouvernement ouvert du gouvernement du Canada vise à garantir que les Canadiens ont accès à la plus grande quantité possible de données et de renseignements gouvernementaux. Les données synthétiques représentent une solution pour les données ouvertes. Une version synthétique d'une base de données permettrait de traiter les problèmes de confidentialité des données personnelles tout en conservant la plus grande valeur analytique possible. Les méthodes utilisées par Statistique Canada pour créer des données synthétiques ont été documentées par Kenza Sallier, qui a ainsi remporté le prix des jeunes statisticiens de 2020 (Sallier, 2020).

Progrès :

L'un des défis liés aux données synthétiques est la terminologie et la nomenclature utilisées pour décrire les divers types de données synthétiques ainsi que les diverses méthodes et les divers outils disponibles. Un guide a été élaboré pour la dernière fois en 2002; il est en cours de révision et la nouvelle version traitera de bon nombre des méthodes modernes axées sur les fichiers synthétiques qui préservent le contenu analytique plutôt que de créer de faux fichiers simples. Le guide peut contribuer à certains des travaux en cours de la Commission économique des Nations Unies pour l'Europe et du Groupe de haut niveau pour la modernisation des statistiques officielles ou il peut s'appuyer sur ces travaux. Statistique Canada a également contribué au groupe en lui faisant part de ses expériences (Sallier, 2021).

Enfin, des efforts ont été consacrés à l'étude d'autres outils et méthodes qui pourraient compléter ou remplacer ceux utilisés jusqu'à présent. Une étude comparant les ensembles logiciels R Synthpop avec SimPop a été réalisée (Zhao, 2021).

À mesure que Statistique Canada acquérait une plus grande expertise dans la production de fichiers de données synthétiques de haute valeur analytique, nous avons eu à relever de nouveaux défis en 2021-2022, y compris en ce qui concerne la synthèse de données qui préserve les structures hiérarchiques sous la forme de familles où les enregistrements sont couplés et partagent des traits communs qui doivent être préservés. De tels défis se posent également lorsque d'autres données structurées comme les données d'entreprise sont synthétisées. Un exposé, qui a été présenté au Symposium de 2021 de Statistique Canada, donnait un exemple concret de l'application d'une telle stratégie aux données sur le revenu des familles (Gauvin, 2021). Le nouvel objectif pour 2022-2023 est de synthétiser les données historiques sur le revenu et les pensions de ces familles de 1966 à 2015. Le défi est de maintenir la corrélation des données longitudinalement.

Pour obtenir plus de renseignements, communiquez avec :
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

Bibliographie

Gauvin, H. (2021). Generating structured microdata: The example of synthesizing hierarchical data. Dans *le Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

Sallier, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. *Statistical Journal of the IAOS*, 36, 1059-1066.

Sallier, K. (2021). Statistics Canada's experience creating public synthetic datasets using the FCS and the Synthpop package. Présentation au UNECE/HLG-MOS working group on synthetic data.

Zhao, Z. (2021). Exploring available tools to generate synthetic data with high analytical value: R packages synthpop and simPop. Présentation interne, Statistique Canada, Ottawa.

6 Soutien (centres de ressources)

6.1 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 de traitement des séries chronologiques;
- développement et soutien pour la désaisonnalisation et l'estimation des tendances-cycles;
- modélisation et prévision de séries chronologiques, particulièrement dans le contexte de l'estimation en temps réel.

SOUS-PROJET : Consultation et formation sur les séries chronologiques

Le Centre de recherche et d'analyse en séries chronologiques est chargé d'élaborer et de dispenser la formation sur les méthodes en séries chronologiques, y compris la désaisonnalisation, le rapprochement et la modélisation en séries chronologiques, à l'intention des participants de Statistique Canada et aussi d'autres organismes, principalement par l'entremise de l'Institut de formation de Statistique Canada. De plus, le Centre offre des conseils et des consultations sur les projets de séries chronologiques en général pour les programmes de l'ensemble de Statistique Canada.

Progrès :

Des cours sur l'ajustement saisonnier (H-0431 et H-0434), l'étalonnage (H-0436) et le ratissage (H-0437) ont été donnés au cours de l'année par l'entremise de l'Institut de formation de Statistique Canada (Statistique Canada, 2022); ces cours ont été mis à jour et comprennent des démonstrations et des ateliers permettant de montrer l'aspect pratique des techniques. Des cours sur la modélisation et la prévision des séries chronologiques ainsi que sur les techniques d'ajustement saisonnier et de rapprochement ont été donnés à des participants d'organismes externes sous la forme de présentations à distance. 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) et les agents financiers, 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.

Le Centre a également offert des consultations au Système de comptabilité nationale dans un certain nombre de domaines. Le Centre a produit des séries chronologiques rapprochées pour les importations et les exportations de services par pays afin d'évaluer l'ajout de cette étape au programme des comptes et ainsi d'améliorer la cohérence des données. Dans le cadre du projet Environnement mondial pour l'écosystème de la statistique économique, le Centre mène également des consultations sur l'application des étapes pertinentes de rapprochement et la recherche et l'élaboration d'outils appropriés. De plus, des représentants du Centre assistent régulièrement au forum des analystes du *Quotidien* pour maintenir une présence dans la collectivité des analystes et offrir un soutien en matière de séries chronologiques à divers programmes, y compris pour le produit intérieur brut mensuel.

Le Centre mène régulièrement des consultations sur l'extrapolation rétrospective afin de préserver ou de rétablir la comparabilité au fil du temps, et il met la dernière main à une directive 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, et nous avons récemment présenté le document à la haute direction pour obtenir leurs commentaires et leur approbation, et nous collaborons avec des responsables de la diffusion pour mettre la dernière main au document en vue de sa diffusion publique. Le document (Matthews et Saint-Pierre, 2022) devrait être publié au cours du premier trimestre de l'exercice 2022-2023.

Pour obtenir plus de renseignements, communiquez avec :
Steve Matthews (613-854-3174, steve.matthews@statcan.gc.ca).

Bibliographie

Matthews, S., et Saint-Pierre, É. (2022). Guidelines on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics (v1.1). Document interne disponible sur demande, Statistique Canada.

Statistique Canada (2022). Ateliers, formation et conférences.
https://www.statcan.gc.ca/fr/afc/formation.

SOUS-PROJET : Soutien et amélioration du système et des outils de traitement des séries chronologiques

Le système de traitement des séries chronologiques est une application SAS personnalisable qui permet d'appliquer des techniques de séries chronologiques, y compris des techniques de désaisonnalisation et de rapprochement, largement utilisées dans la production d'estimations désaisonnalisées pour les programmes essentiels à la mission de Statistique Canada. Le système est dans un état assez mature et stable, mais il doit être mis à jour de façon continue; il faut élargir la fonctionnalité et répondre aux nouveaux besoins des programmes de l'organisme. À plus long terme, nous envisageons une nouvelle version du système qui permettrait d'intégrer les outils et les nouvelles techniques disponibles à partir de logiciels libres.

Progrès :

La diffusion de la version V3.08 du système de traitement des séries chronologiques a permis d'améliorer considérablement la fonctionnalité des modules d'étalonnage et d'équilibrage (Ferland, 2022). Le travail a été effectué en soutien à l'extrapolation rétrospective et à l'étalonnage pour un projet particulier, mais la nouvelle fonctionnalité est largement applicable et a donc été intégrée aux modules du système de

traitement. Plus précisément, les changements permettent l'étalonnage proportionnel des variables des stocks et certaines mises à niveau des validations et des paramètres disponibles dans le module d'équilibrage. Une application peut inclure les liens comptables généralement présents dans les données financières lorsque les contraintes ne sont pas simplement additives, ainsi que l'étalonnage lorsque les variables ne sont pas agrégées temporairement pour correspondre aux totaux annuels (p. ex. lors du rapprochement des inventaires trimestriels par rapport à une valeur annuelle).

Des études ont permis d'évaluer les outils disponibles pour l'application de l'étalonnage, du ratissage et de la désaisonnalisation au moyen d'outils libres, y compris ceux disponibles en R, Python et les programmes en Java mis au point et publiés par la Commission européenne. Un prototype de fonction R a été programmé avec une fonctionnalité équivalente à PROC benchmarking dans G-SERIES, dont l'utilisation interne est en cours d'évaluation. Statistique Canada a continué à participer au Seasonal Adjustment Center of Excellence d'Eurostat en tant qu'organisme partenaire afin de pouvoir participer aux discussions et à l'élaboration d'outils connexes.

Pour obtenir plus de renseignements, communiquez avec :
Steve Matthews (613-854-3174, steve.matthews@statcan.gc.ca).

Bibliographie

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

SOUS-PROJET : Développement et soutien pour la désaisonnalisation et l'estimation de tendances-cycles

L'objectif du projet est d'effectuer des analyses et des évaluations de nouvelles méthodes et techniques de désaisonnalisation et d'estimation de tendances-cycles et de mener des consultations et de centraliser l'expertise au chapitre de l'application de la désaisonnalisation.

Progrès :

Le Centre de recherche et d'analyse en séries chronologiques a effectué une importante assurance de la qualité pour la désaisonnalisation, car les effets de la pandémie de COVID-19 se sont poursuivis. Des échanges ont eu lieu avec des représentants d'autres bureaux nationaux de statistique, par courriel ainsi que dans le cadre de conférences et de réunions virtuelles. Ces échanges ont eu lieu avec des membres d'Eurostat, de l'Office for National Statistics, du United States Census Bureau, du United States Bureau of Labor Statistics, de Statistics Norway, de l'INSEE (Institut national de la statistique et des études économiques), de Statistics Israel, du Bureau de la statistique de l'Australie, de Statistics New Zealand et d'autres organismes. Les échanges ont été extrêmement utiles et ont permis de comparer et de mettre en opposition les approches à court et à long terme, qui étaient presque universellement conformes à l'approche suggérée par Eurostat (2020). Pour que l'information soit largement accessible, le Centre a communiqué les conclusions et les recommandations découlant des consultations aux personnes-ressources pertinentes au sein de l'organisme au moyen de mises à jour périodiques.

Nous avons élaboré une stratégie pour réduire la validation accrue appliquée pendant le choc initial de la pandémie et ainsi réduire le soutien de façon continue (stratégie de retour graduel à l'approche d'avant la pandémie — ajustement des valeurs critiques pour la détection des valeurs aberrantes, détection automatique des valeurs aberrantes au cours des derniers mois, etc.). Cette stratégie globale a été

documentée et partagée avec d'autres spécialistes de la désaisonnalisation (Matthews, 2021a). Le Centre a également participé à une table ronde sur la désaisonnalisation lors des Joint Statistical Meetings de 2021 de l'American Statistical Association (Matthews, 2021b). À mesure que chaque indicateur économique se stabilise, une stratégie est élaborée pour que nous revenions graduellement aux pratiques d'assurance de la qualité qui étaient en place avant la pandémie sans causer d'effets défavorables aux estimations désaisonnalisées en temps réel et rétrospectivement. Nous avons étudié les modèles d'espace d'états afin de produire des indications précoces de ruptures structurelles, et le travail se poursuivra.

Le Centre de recherche et d'analyse en séries chronologiques a également tenu de vastes consultations sur la désaisonnalisation au sein de l'organisme pour des programmes qui ne reçoivent pas l'appui officiel de l'équipe. Compte tenu des défis liés à la désaisonnalisation qui découlent de la pandémie et de ses chocs économiques, une réunion trimestrielle est maintenant tenue pour que nous puissions offrir des conseils et des consultations sur des problèmes pratiques à plusieurs des programmes du Système de comptabilité nationale. La réunion est coordonnée avec la diffusion trimestrielle faite par ces programmes, et le Centre offre des conseils ponctuels et fait des suivis pour faciliter l'analyse et la production des données désaisonnalisées. Des consultations ont été offertes aux groupes qui produisent des statistiques désaisonnalisées sur le commerce des services et sur le nombre d'inscriptions et de sorties dans le Registre des entreprises.

Des progrès continus ont été réalisés pour que le tableau de bord sur la désaisonnalisation soit accessible aux analystes pour un nombre croissant de programmes, quatre autres enquêtes essentielles à la mission pouvant maintenant avoir accès à l'outil pour comprendre et expliquer les résultats désaisonnalisés, et ce nombre étant appelé à augmenter au cours du prochain exercice. Le tableau de bord a de plus été présenté lors du Symposium international de 2021 sur les questions de méthodologie de Statistique Canada (Verret, 2021).

De plus, les méthodes d'estimation des tendances-cycles ont été évaluées dans le contexte de la pandémie de COVID-19. En particulier, compte tenu de la gravité des chocs économiques, la tendance-cycle actuellement publiée pour certains programmes à Statistique Canada (selon le filtre linéaire en cascade proposé dans Dagum et Luati, 2009) peut présenter une impression trop en douceur de l'économie, de sorte qu'un certain nombre d'autres mesures ont été déterminées, y compris la rupture de la série à un point donné et l'introduction d'effets aberrants pour modéliser les chocs plus directement. Les méthodes ont été appliquées aux résultats de plusieurs programmes, et nous les présenterons à l'interne et à l'externe afin de déterminer si les ajustements à la méthodologie devraient être adoptés.

Nous avons de plus commencé à tenir des réunions d'équipe trimestrielles avec l'équipe des séries chronologiques de Statistique Canada et l'équipe correspondante du United States Census Bureau. Ces réunions permettent d'échanger sur les priorités et les nouveautés en matière de recherche et de collaborer aux enjeux actuels concernant l'analyse en matière de séries chronologiques.

Pour obtenir plus de renseignements, communiquez avec :
Steve Matthews (613-854-3174, steve.matthews@statcan.gc.ca).

Bibliographie

Dagum, E.B., et Luati, A. (2009). A cascade linear filter to reduce revisions and false turning points for real time trend-cycle estimation. *Econometric Reviews*, 28, 40-59.

Eurostat (2020). Guidance on time series treatment in the context of the COVID-19 crisis. https://ec.europa.eu/eurostat/documents/10186/10693286/Time_series_treatment_guidance.pdf.

Matthews, S. (2021a). Seasonal Adjustment during the COVID-19 pandemic: Statistics Canada's Approach. *Proceedings of the European Establishment Statistics Workshop*.

Matthews, S. (2021b). *Time Series and Seasonal Adjustment Estimation during the COVID-19 Pandemic*. Round-table session at the 2021 Joint Statistical Meetings of the American Statistical Association.

Verret, F. (2021). *Statistics Canada's Seasonal Adjustment Dashboard*. Présenté au Symposium international de 2021 sur les questions de méthodologie, Statistique Canada.

SOUS-PROJET : Modélisation et prévision de séries chronologiques, en particulier dans le contexte de l'estimation en temps réel

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.

Progrès :

L'évaluation de modèles statistiques dans un tel contexte s'est poursuivie, et les modèles d'apprentissage automatique ont été comparés avec d'autres approches de séries chronologiques, les modèles ARIMA-X et les modèles d'espace d'états étant les candidats les plus attrayants. Les modèles d'espace d'états comprennent l'application de modèles de facteurs dynamiques, qui combinent la réduction de la dimensionnalité, les données de fréquence mixte et d'autres questions pratiques, et ils sont de plus en plus utilisés dans des applications de prévisions immédiates à l'échelle internationale. Une présentation a été faite au Comité consultatif des méthodes statistiques de Statistique Canada; l'objectif était de communiquer les progrès et d'étudier l'aspect de l'explicabilité ainsi que les plans futurs pour le projet (Matthews et Le Moullec, 2021). L'évaluation de modèles de facteurs dynamiques a été réalisée en partenariat avec Rafal Kulik de l'Université d'Ottawa ainsi qu'avec un étudiant des cycles supérieurs, Ismael Zie Diamoutene, qui collabore également à l'analyse. L'évaluation de la méthode concerne le réglage des hyperparamètres et l'estimation de la variance, ainsi que les futurs domaines de recherche (Diamoutene, 2022). L'enquête a permis de conclure que les modèles de facteurs dynamiques pourraient être grandement automatisés et efficaces avec des variables auxiliaires fortement corrélées; certains modèles manuels de présélection semblent toutefois contribuer à l'élimination des variables ayant des corrélations plus faibles ou potentiellement fausses qui n'améliorent pas la précision (Patak, 2022). Une présentation sommaire des travaux sur les prévisions immédiates a été donnée à la sixième International Conference on Establishment Surveys (voir Matthews, 2021c).

Deux études de cas ont été réalisées en collaboration avec la Division de la science des données, et elles ont permis d'évaluer les méthodes prédictives pour le produit intérieur brut mensuel. Ces études ont permis de comparer des modèles plus traditionnels, comme la régression avec erreurs ARMA, avec des modèles prédictifs d'apprentissage automatique, et elles ont permis de révéler un rendement semblable et de mettre en évidence non seulement les différences entre les méthodes, mais aussi de nombreuses similitudes (Ritter et Patak, 2021). Une présentation sur les travaux a été donnée lors du Symposium international de 2021 sur les questions de méthodologie de Statistique Canada.

Nous avons préparé un document provisoire visant à présenter un cadre pour les indicateurs avancés, à faire la promotion de la normalisation de la terminologie pour la production d'indicateurs avancés (y

compris les prévisions immédiates fondées sur des modèles), à décrire les critères appuyant la décision de publier de nouveaux indicateurs avancés, et à offrir des conseils sur les méthodes appropriées de prévision immédiate ainsi que sur les avantages et les désavantages (Matthews, 2022). De plus, nous avons rédigé un article sommaire pour un numéro de 2021 sur la convergence afin de faire connaître le travail effectué récemment sur les prévisions immédiates (Picard, 2021).

Pour obtenir plus de renseignements, communiquez avec :
Steve Matthews (613-854-3174, steve.matthews@statcan.gc.ca).

Bibliographie

Diamoutene, I. (2022). *Technical Notes on Dynamic Factor Models*. Thèse de doctorat, University of Ottawa (à publier).

Matthews, S. (2021c). *Toward Near Real-Time Economic Indicators Using Time Series Models: Statistics Canada's Progress*. Présenté à la Sixth International Conference on Establishment Statistics.

Matthews, S. (2022). *Framework for Development and Production of Advance Indicators at Statistics Canada*. Document interne disponible sur demande, Statistique Canada.

Matthews, S., et Le Moullec, J. (2021). *Development of More Timely Indicators of Gross Domestic Product*. Présenté à la 73e réunion du Comité consultatif sur les méthodes statistiques, Statistique Canada.

Patak, Z. (2022). Nowcasting the Canadian GDP. Document interne disponible sur demande, Statistique Canada.

Picard, F. (2021). Le Nowcasting. *Convergence*, 26, 15-17, Septembre 2021, Association des statisticiennes et statisticiens du Québec.

Ritter, C., et Patak, Z. (2021). *On the Path to More Timely Economic Indicators: A Comparison of Traditional and New Machine Learning Nowcasting Methods*. Présenté au Symposium international de 2021 sur les questions de méthodologie, Statistique Canada.

6.2 Systèmes généralisés de statistiques économiques

L'équipe des systèmes généralisés pour les statistiques économiques est responsable du soutien et du développement de quatre systèmes généralisés : G-SAM, le système généralisé d'échantillonnage; BANFF, le système généralisé de contrôle et d'imputation; G-EST, le système généralisé d'estimation; G-SERIES, le système généralisé pour les techniques de séries chronologiques.

Progrès :

Un volume habituel de cas de soutien a été traité par l'équipe de projet, principalement pour G-EST, BANFF et G-SAM. La plupart des cas ont été résolus par des suggestions sur la façon d'appliquer les systèmes en termes pratiques, mais plusieurs cas ont nécessité une intervention plus poussée.

Une nouvelle version de G-EST visant à corriger une erreur a été diffusée cette année. Une erreur de programmation a été découverte dans la version 2.03.002, et elle a donné lieu à des erreurs dans l'estimation de la variance dans certaines conditions (pour les estimateurs calés lorsque la matrice de plan d'expérience est singulière). L'erreur a été communiquée aux utilisateurs, et aucun cas où l'erreur a causé des problèmes concernant des valeurs publiées n'a été trouvé. La version G-EST 2.03.003 a été diffusée pour corriger l'erreur (Statistique Canada, 2021).

La mise à l'essai s'est poursuivie et a permis de mettre au point la version de G-EST pour qu'elle tienne compte des grands ensembles de données complexes dans l'application de SEVANI (partie de G-EST) pour estimer la variance due à l'imputation. Cette nouvelle version permettra à plusieurs programmes de Statistique Canada de calculer des estimations améliorées de la variance, ce qui est important pour permettre l'utilisation de l'ajustement tabulaire aléatoire afin d'améliorer le contrôle de la divulgation. La mise à l'essai des prototypes est presque terminée, et une nouvelle version de G-EST sera diffusée au cours des prochains mois.

Beaucoup de progrès ont également été réalisés dans le cadre d'une importante initiative pour les systèmes généralisés, un exercice de planification à long terme tenant compte de l'évolution à venir des systèmes. Chaque système fait l'objet d'un examen permettant d'étudier la possibilité d'utiliser des outils libres, y compris de nouvelles méthodes de pointe, et de revoir l'approche de gestion des développements futurs (Gray et Matthews, 2021). L'examen est terminé pour les systèmes BANFF et G-SERIES, et la recommandation est de moderniser ces systèmes en les développant indépendamment de SAS (Gray et Arsenault, 2021, et Matthews et Mathieu, 2021). Ces projets ont été proposés à titre d'investissements pour l'année à venir, et des travaux sont en cours pour la modernisation de BANFF. L'évaluation des options pour G-SAM et G-EST sera terminée dans la prochaine année, et les recommandations qui en découleront seront présentées au comité directeur des systèmes généralisés statistiques.

L'élaboration d'ImpACT, un système permettant de visualiser l'imputation, s'est poursuivie, et le système a été utilisé pour l'évaluation des stratégies d'imputation de deux projets de Statistique Canada. Les travaux ont été présentés au Symposium de 2021 sur les questions de méthodologie de Statistique Canada (Gray, 2021). Il est prévu d'étendre l'outil afin d'y intégrer d'autres visualisations et de l'appliquer à d'autres programmes. De plus, des recherches ont été effectuées sur l'ajout possible de l'ajustement proportionnel entier à deux dimensions au système BANFF (Baillargeon, 2022a).

Pour le système G-SAM, nous avons produit des documents additionnels pour mieux décrire le contexte, la fonctionnalité et l'application du logiciel. Des études ont également permis d'étudier la généralisation des procédures de stratification et de répartition au moyen d'une approche d'optimisation (Baillargeon, 2022b).

Les membres de l'équipe ont donné de la formation dans le cadre de cours officiels de l'Institut de formation de Statistique Canada, de séminaires à l'intention des statisticiens récemment recrutés et d'autres présentations ponctuelles faites à des analystes et à d'autres organismes. L'équipe a également contribué à l'organisation de l'atelier sur la vérification des données de la Commission économique des Nations Unies pour l'Europe prévu pour l'automne 2022.

Pour obtenir plus de renseignements, communiquez avec :
Steve Matthews (613-854-3174, steve.matthews@statcan.gc.ca).

Bibliographie

Baillargeon, J. (2022a). Bidirectional Pro-rating. Document interne disponible sur demande, Statistique Canada.

Baillargeon, J. (2022b). Optimization for sample designs: Progress report on research activities. Document interne disponible sur demande, Statistique Canada.

Gray, D. (2021). Improving decision-making in imputation design through data visualization. Présenté au *Symposium international de 2021 sur les questions de méthodologie*, Statistique Canada.

Gray, D., et Arsenault, S. (2021). BANFF – Generalized Systems evolution plan: PHASE III. Presentation to the Statistical Generalized Systems Steering Committee, Rapport interne, Statistique Canada.

Gray, D., et Matthews, S. (2021). Rethinking the role of Generalized Systems for economic statistics. Présenté à la 73e réunion du Comité consultatif sur les méthodes statistiques, Rapport interne, Statistique Canada.

Matthews, S., et Mathieu, M.-C. (2021). G-SERIES evolution plan: Phase 1 Summary. Présenté à la 73e réunion du Comité consultatif sur les méthodes statistiques, Rapport interne, Statistique Canada.

Statistique Canada (2021). G-EST 2.03.003 User Guide. Document interne disponible sur demande, Statistique Canada.

6.3 Centre de ressources en couplage d'enregistrements

Les objectifs du Centre de ressources en couplage d'enregistrements (CRCE) consistent à offrir des services de consultations aux utilisateurs internes et externes des méthodes de couplage d'enregistrements, ce qui comprend la formulation de recommandations au sujet de logiciels et de méthodes à utiliser, des travaux concertés sur les applications de couplage d'enregistrements, l'évaluation de méthodes de couplage d'enregistrements et le développement de méthodes améliorées. Nous facilitons aussi la diffusion de l'information sur les méthodes, les logiciels et les applications de couplage d'enregistrements aux personnes intéressées à l'intérieur et à l'extérieur de Statistique Canada.

Progrès :

Nous avons continué à soutenir l'équipe de développement de G-Coup et à participer aux réunions du groupe de travail sur le couplage d'enregistrements. Les membres de l'équipe se sont rencontrés toutes les deux semaines et le CRCE a fait le suivi des procès-verbaux mentionnant des sources possibles, passées ou présentes, de corrections, de bogues ou d'améliorations pour G-Coup. Le CRCE a aussi offert un soutien aux utilisateurs internes et externes de G-Coup qui ont demandé de l'aide, ont formulé des commentaires ou ont présenté des suggestions à G-Coup_info au moyen de billets et de requêtes.

Durant l'année la majorité du travail méthodologique s'est porté sur la maintenance, le développement et le support auprès des utilisateurs de la nouvelle version 3.5 de G-Coup qui inclut l'addition d'un couplage basé sur les profils et des indicateurs de qualité. Deux outils fondés sur la révision manuelle pour estimer les taux de faux positifs et faux négatifs ont été intégrés dans la version 3.3 de G-coup. De plus,

on a entrepris un travail exploratoire de couplage sur l'infonuagique et le test du logiciel de couplage splink, codé en python (développé par l' « Office for National Statistics » au Royaume-Uni), en utilisant la plateforme Databricks.

Le CRCE a aussi travaillé sur une variété d'autres projets de couplage probabiliste et a animé des réunions avec des collègues internationaux. Ces couplages nous ont aidés à analyser la performance du logiciel et les solutions à apporter. Le travail sur ces données a permis d'élaborer des approches plus systématiques de définir et d'ajuster les couplages d'enregistrements sur des serveurs et sur la plateforme SAS Grid. Le RLRC a également orienté des recherches sur l'évaluation de la qualité du couplage probabiliste qui a amené à des échanges internationaux avec l'agence Française INSEE et à la présentation de nos travaux à la conférence des Journées Mondiales de Statistique (JMS) de 2022 à Paris.

Pour obtenir plus de renseignements, communiquez avec :
Abdelnasser Saïdi (613-863-7863, abdelnasser.saidi@statcan.gc.ca).

6.4 Centre de ressources en analyse de données

Le Centre de ressources en analyse de données a pour objectif premier de donner 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 Centre de ressources en analyse de données, 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 ou des centres de données de recherche.

Progrès :

Consultations

Des services de consultation ont été offerts à la demande de clients internes et externes. Les questions variaient en complexité et portaient sur l'utilisation de poids bootstrap d'enquête, la construction d'intervalles de confiance et les tests d'hypothèses avec des données d'enquête, l'estimation de l'importance des effets avec des données d'enquête, l'analyse avec des données couplées et l'ajustement de modèles de régression logistique. Le Centre de ressources en analyse de données a également aidé les clients à mettre en œuvre des méthodes dans les logiciels SUDAAN, SAS, STATA et R.

Services et matériel de formation

Hao et Provençal (2021) ont conçu et présenté un atelier sur la saisie des politiques et l'analyse factorielle à l'intention du groupe de l'Équité salariale du Conseil du Trésor.

Le Centre de ressources en analyse de données a également fait une présentation sur l'utilisation des statistiques descriptives dans le cadre d'un atelier sur le recensement et l'analyse de données organisé par le Projet régional d'avancement de la statistique dans les Caraïbes pour l'Institut de la statistique du Belize.

Le Centre de ressources en analyse de données a fait une présentation sur l'analyse de données avec des données d'enquête complexes et sur l'utilisation de statistiques descriptives lors de l'atelier sur l'interprétation des données organisé par Statistique Canada.

Le Centre de ressources en analyse de données a présenté la séance sur la régression linéaire avec des données d'enquête complexes dans le cadre du cours sur la modélisation statistique à Statistique Canada. Le Centre de ressources en analyse de données a également fait plusieurs présentations sur l'analyse des données tirées d'une enquête complexe dans la série de séminaires pour les recrues.

Collaboration

Le Centre de ressources en analyse de données a collaboré à l'élaboration de stratégies de mesure pour trois projets : i) le Projet de mesure du rendement en santé mentale en milieu de travail; ii) l'Indice de renouvellement de la fonction publique au-delà de 2020; iii) la Stratégie d'accessibilité pour la fonction publique du Canada. Pour ces projets, les données recueillies dans le cadre du Sondage auprès des fonctionnaires fédéraux de 2019 et de 2020 ont été utilisées pour mesurer des variables latentes comme les facteurs de risque psychologiques, les comportements, etc. Pour les projets i) et ii), le Centre de ressources en analyse de données a calculé les scores factoriels pour différents niveaux d'agrégation, et il a produit la documentation sur la méthodologie. Pour le projet iii), le Centre de ressources en analyse de données a dérivé un modèle. Les scores factoriels établis pour le projet i) ont servi à la création du tableau de bord sur la santé mentale en milieu de travail dans la fonction publique fédérale lancé le 17 mai 2022 : Stratégie pour la fonction publique fédérale sur la santé mentale en milieu de travail. Les modèles de mesure ont été élaborés à l'aide de l'analyse factorielle et de la modélisation des équations structurelles, comme l'ont expliqué le Centre de ressources en analyse de données (2020), Blais, Mach, Michaud et Simard (2020) et Blais, Michaud, Simard, Mach et Houle (2021).

Pour obtenir plus de renseignements, communiquez avec :
Harold Mantel (613-863-9135, harold.mantel@statcan.gc.ca) ou
Fritz Pierre (613-720-4318, fritz.pierre@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 interne au Workplace Mental Health Performance Measurement Steering Committee, le 7 octobre 2020.

Blais, A.-R., Michaud, I., Simard J.-S., Mach, L. et Houle, S. (2021). Mesurer les facteurs psychosociaux en milieu de travail. *Rapports sur la santé*, 32, Statistique Canada.

Centres de données de recherche (2020). *Measuring Culture of Accessibility, A Brief Introduction to Factor Analysis*. Présentation interne à l'Accessibility Culture Goal Working Group, le 1er mai 2020.

Hao, Y., et Provençal, J.-S. (2021). Policy Capturing: an overview for Treasury Board of Canada Secretariat. Atelier au Pay Equity Group, Treasury Board of Canada Secretariat, 26 octobre et 2 novembre 2021.

6.5 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 ou des nouvelles utilisations de données, 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 joue également un rôle de renforcement des capacités.

Progrès :

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. Un document sur les fondements des examens éthiques dans le contexte statistique, en cours d'élaboration, présente les six principes directeurs. Le document devrait être disponible en 2022. Enfin, les travaux se sont poursuivis en vue de l'élaboration d'une échelle de sensibilité, un outil qui aide à surveiller la sensibilité des nouvelles acquisitions de données ou des nouveaux projets.

Pour obtenir plus de renseignements, communiquez avec :
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

6.6 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.

SOUS-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 également présenté deux webinaires ouverts au public sur l'importance de la qualité des données, dans le cadre de la série de webinaires organisée par le Centre de services de données de Statistique Canada. Des documents sur la qualité des données et les bonnes pratiques de gestion de la qualité ont été présentés à l'Initiative de formation en littératie des données de Statistique Canada. 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. Un cadre de qualité des données est en cours d'élaboration et sera disponible au printemps 2022. À l'échelle internationale, la participation au Groupe d'experts des Nations Unies sur les cadres nationaux d'assurance de la qualité s'est poursuivie en vue de la mise en œuvre du *National Quality Assurance Framework Manual for Official Statistics* des Nations Unies (Nations Unies, 2019).

Pour obtenir plus de renseignements, communiquez avec :
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

Bibliographie

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

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

Afin de fournir aux utilisateurs des indicateurs de qualité pour les programmes qui combinent des sources de données administratives, le Secrétariat de la qualité a entrepris de mettre au point un indicateur composite qui combine les indicateurs de qualité liés aux différentes étapes du traitement des données (couplage d'enregistrements, imputation, géocodage, etc.) en un seul indicateur. L'objectif est de donner une vue globale de la qualité d'une estimation compte tenu de plusieurs facteurs qui peuvent introduire des erreurs (Gagnon, Qian, Yeung, Lebrasseur et Beaulieu, 2022). Un premier programme, le Programme de la statistique du logement canadien, a publié ces indicateurs avec les estimations en septembre 2021. Le projet a été présenté au European Establishment Statistics Workshop en septembre 2021 (Beaulieu, Lebrasseur et Gagnon, 2021) ainsi qu'au Comité consultatif des méthodes statistiques de Statistique Canada (Beaulieu et Gagnon, 2021) et au Symposium international de 2021 sur les questions de méthodologie de Statistique Canada (Gagnon, Beaulieu, Lebrasseur, Qian et Yeung, 2021).

Pour obtenir plus de renseignements, communiquez avec :
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

Bibliographie

Beaulieu, M., Lebrasseur, D. et Gagnon, R. (2021). Measuring and communicating quality for programs using administrative data sources exclusively. *Proceedings of the 2021 European Establishment Statistics Workshop.*

Beaulieu, M., et Gagnon, R. (2021). *Measuring and Communicating Quality for Statistical Programs Based on Administrative Data: The Canadian Housing Statistics Program's Composite Quality Indicator*. *Communication présentée au Comité consultatif sur les méthodes statistiques, novembre 2021, Statistique Canada.*

Gagnon, R., Beaulieu, M., Lebrasseur, D., Qian, W. et Yeung, A. (2021). Creation of a composite quality indicator for estimates based on administrative data using clustering. *Recueil du Symposium international de 2021 sur les questions de méthodologie.*

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, Publié le 6 janvier 2022, https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-fra.htm.

6.7 Centre de ressources en assurance de la qualité

Le Centre de ressources en assurance de la qualité a comme objectif de mener des activités de recherche et de développement sur les méthodes statistiques d'assurance et de contrôle de la qualité, afin d'améliorer la qualité après contrôle des opérations de collecte et de traitement des données d'enquête au sein de Statistique Canada. Il offre notamment des services méthodologiques pour G-code, qui est utilisé à Statistique Canada pour la création de bases de données de codage pour le traitement des données. La recherche sur l'assurance et le contrôle de la qualité est souvent de nature générique et porte sur des questions d'efficacité et d'automatisation qui sont fréquemment appliquées à de nombreuses étapes des opérations d'enquête.

Progrès :

L'équipe de soutien méthodologique a aidé l'équipe de développement de G-Code et elle a fait le suivi des commentaires des utilisateurs pour chercher à trouver des idées d'améliorations possibles pour G-Code. Le Centre de ressources en assurance de la qualité a également offert du soutien aux utilisateurs internes et externes de G-Code lorsqu'ils avaient besoin d'aide, ou avaient des commentaires ou suggestions au sujet de G-Code.

Au cours de l'année, les travaux ont porté sur la mise en œuvre d'une nouvelle version de G-Code (version 3.3), qui comprend l'ajout de capacités d'apprentissage automatique (XgBoost, FastText et Pytorch). Plus précisément, l'équipe du Centre de ressources en assurance de la qualité a participé à une validation du concept de codage et de classification en vue d'examiner l'intégration de l'algorithme FastText à G-Code. Les nouveaux algorithmes ont été largement utilisés pour le codage de diverses classifications pour l'Enquête sur la population active, l'Enquête sur les postes vacants et les salaires, l'Enquête sur la santé dans les collectivités canadiennes, l'Enquête sur les permis de bâtir, le Système d'information sur les étudiants postsecondaires, le Registre statistique des entreprises et le Recensement de la population. De plus, ces nouvelles fonctionnalités ont été présentées à des organismes externes (Bureau de la statistique de l'Australie, Statistics New Zealand et Institut de la statistique du Belize) et à l'interne dans le cadre de cours et de démonstrations axées sur des projets. Dernièrement, l'équipe du Centre de ressources en assurance de la qualité a contribué à l'intégration de Pytorch à G-Code. PyTorch est une bibliothèque d'apprentissage automatique libre basée sur la bibliothèque Torch, qui est utilisée pour des applications comme la vision informatique et le traitement du langage naturel, principalement mis au point par le laboratoire de recherche sur l'intelligence artificielle de Facebook.

Le Centre de ressources en assurance de la qualité a également travaillé à la mise en œuvre de divers contrôles de la qualité des processus de codage pour l'Enquête sur la population active, l'Enquête sur la santé dans les collectivités canadiennes, le Système d'information sur les étudiants postsecondaires, l'Enquête sur les permis de bâtir, l'Enquête sur les postes vacants et les salaires, le Registre statistique des entreprises et le nouveau Registre statistique des immeubles.

De plus, un article portant sur l'intégration d'un modèle FastText et d'un processus de contrôle de la qualité aux activités de codage de l'Enquête sur la population active a été présenté au Symposium de 2021 de Statistique Canada (Evans et Oyarzun, 2022).

Pour obtenir plus de renseignements, communiquez avec :
Javier Oyarzun (613-302-8454, javier.oyarzun@statcan.gc.ca).

Bibliographie

Evans, J., et Oyarzun, J. (2022). Need for speed: Using fastText (Machine Learning) to code the Labour Force Survey. Dans le *Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

6.8 Centre de ressources en conception de questionnaires

Le Centre de ressources en conception de questionnaires est le centre d'expertise de Statistique Canada en matière de conception et d'évaluation de questionnaires. Il offre des services de consultation et de soutien et mène des projets et des recherches concernant l'élaboration, la mise à l'essai et l'évaluation de questionnaires d'enquête. Il joue un rôle très important dans la gestion de la qualité et il répond aux exigences des programmes de l'ensemble de Statistique Canada en consultant les clients, les répondants et les utilisateurs de données et en procédant à l'essai préliminaire de 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 la Banque du Canada, l'Agence de la santé publique du Canada, Patrimoine canadien, Ressources naturelles Canada, Services publics et Approvisionnement Canada, le Conseil de la santé du Nouveau-Brunswick et d'autres.

Le Centre de ressources en conception de questionnaires a continué à expérimenter la recherche sur les méthodes mixtes. De plus, certaines recherches et expérimentations ont commencé par des méthodes qualitatives asynchrones.

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

Pour obtenir plus de renseignements, communiquez avec :
Paul Kelly (613-371-1489, paul.kelly@statcan.gc.ca).

6.9 Confidentialité

Une partie du rôle et de la responsabilité de Statistique Canada demeure la sensibilisation et le soutien aux stratégies en matière de confidentialité. Parmi les activités menées tout au long de l'année, mentionnons la tenue d'un atelier sur l'anonymisation pour le Projet régional d'avancement de la statistique dans les Caraïbes, des discussions avec l'Agence du revenu du Canada au sujet de la confidentialité et l'élaboration de solutions en matière d'accès pour Emploi et Développement social Canada. D'autres activités de recherche et développement en matière de confidentialité sont décrites à la section 5.

Pour obtenir plus de renseignements, communiquez avec :
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

6.10 Communautés de pratique en science des données

SOUS-PROJET : Communauté de pratique de l'apprentissage automatique

La Communauté de pratique de l'apprentissage automatique de Statistique Canada a pour but de faciliter la collaboration et le transfert des connaissances ainsi que d'améliorer nos opérations d'apprentissage automatique à Statistique Canada. Grâce à diverses activités liées à l'apprentissage automatique réunissant de 50 à 80 personnes, comme des présentations, des groupes de lecture, des groupes de visionnement et des échanges d'information sur un site créé et mis à jour par les membres, la Communauté de pratique, par sa présence active, continue à collaborer au développement des capacités en apprentissage automatique des employés de Statistique Canada.

Progrès :

Malgré le travail à distance et le remaniement de la composition du comité, la Communauté de pratique a été active cette année et a pris plusieurs nouvelles orientations. Un nouveau président a été nommé cette année après que le président précédent ait quitté son poste. Plusieurs nouveaux membres du comité ont été recrutés. La Communauté de pratique a commencé à produire deux nouveaux bulletins bimensuels, publiés en alternance chaque mois. L'un des bulletins, qui met l'accent sur les récents progrès de l'apprentissage automatique, est également inclus dans le bulletin des réseaux de science des données. Un centre « porte ouverte » sur l'apprentissage automatique a été configuré dans Microsoft Teams. Tous les vendredis matins, sur un canal dédié de Teams, n'importe qui de Statistique Canada peut venir poser des questions et obtenir des conseils sur tout ce qui a trait à l'apprentissage automatique. Cette année, la Communauté de pratique organise un nouveau défi (trimestriel) de programmation en apprentissage automatique. Des défis de différents niveaux sont proposés, et les équipes ayant présenté des propositions impressionnantes sont invitées à présenter leur approche et leurs travaux. La Communauté de pratique met à jour et tient la page Confluence pour la Communauté de pratique de l'apprentissage automatique, et elle ajoute actuellement des « faits mensuels », rédigés du point de vue d'une personne qui apprend ou désire apprendre au sujet de l'apprentissage automatique. Enfin, la Communauté de pratique continue à communiquer des renseignements à ses membres sur diverses activités externes pertinentes concernant la science des données; elle envoie notamment des renseignements et des invitations pour des marathons de programmation et pour le séminaire de recrutement sur l'apprentissage automatique au sein de l'organisme.

Pour obtenir plus de renseignements, communiquez avec :
Nicholas Denis (613-618-9948, nicholas.denis2@statcan.gc.ca).

SOUS-PROJET : Communauté de pratique de l'apprentissage automatique appliqué à l'analyse de textes

La Communauté de pratique de l'apprentissage automatique appliqué à l'analyse de textes est un lieu interministériel centralisé où les spécialistes ayant diverses expertises peuvent discuter des applications pratiques du traitement du langage naturel au sein du gouvernement du Canada. Divers spécialistes du gouvernement du Canada se réunissent chaque mois pour apprendre et échanger sur les applications éthiques du traitement du langage naturel et les adopter. Les réunions mensuelles rassemblent de 50 à 60 participants qui échangent sur leurs solutions et leurs problèmes. Environ la moitié des participants proviennent de ministères fédéraux à l'extérieur de Statistique Canada.

Progrès :

Tout au long de 2021, cinq spécialistes de divers domaines de Statistique Canada et huit spécialistes d'autres ministères ou organismes, comme le Bureau du surintendant des institutions financières, l'Agence des services frontaliers du Canada, Immigration, Réfugiés et Citoyenneté Canada et l'Agence du revenu du Canada, ont présenté leurs solutions de grande qualité en matière de traitement du langage naturel. Chaque présentateur a illustré sa méthodologie moderne pour traiter rapidement sa source de données, qu'il s'agisse de données d'enquête, de données administratives ou de rapports publics. Les discussions qui ont suivi la présentation ont permis de décomposer les concepts complexes afin que les participants puissent les comprendre. Plus de 140 participants provenaient de 20 ministères différents et environ 130 participants provenaient de Statistique Canada.

Pour obtenir plus de renseignements, communiquez avec :
Joanne Yoon (343-542-5625, joanne.yoon@statcan.gc.ca).

7 Autres activités

7.1 Revue *Techniques d'enquête*

Techniques d'enquête est une revue statistique gratuite à comité de lecture publiée en ligne deux fois par année par Statistique Canada depuis 1975. La revue publie des articles novateurs de recherche théorique ou appliquée, et parfois des articles de synthèse, qui offrent de nouvelles perspectives sur les méthodes statistiques présentant un intérêt pour les organismes nationaux de statistique et d'autres organismes statistiques. Les articles sont publiés gratuitement dans les deux langues officielles et en formats HTML et PDF entièrement accessibles à l'adresse suivante : https://www150.statcan.gc.ca/n1/pub/12-001-x/index-fra.htm. 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 2021 (47-1 et 47-2) ont été publiés. Le numéro de juin 2021 contient neuf articles, dont l'article Waksberg 2020 de Roger Tourangeau : « Science et gestion d'enquête ». Huit

articles ont été publiés dans le numéro de décembre 2021, qui présente l'article Waksberg 2021 de Sharon Lohr : « Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples ».

En 2021, 66 articles ont été soumis à la revue. Le nombre moyen de jours entre la soumission et l'évaluation initiale était de 38. Tous les articles soumis ont été évalués dans un délai de 126 jours, et 88 % d'entre eux l'ont été dans un délai de 90 jours. Parmi les 66 articles soumis, 47 ont été rejetés, 11 ont été acceptés et 8 n'avaient pas fait l'objet d'une décision définitive en date de juin 2022. D'avril 2021 à mars 2022, les pages de *Techniques d'enquête* ont été consultées 45 071 fois, et 20 750 copies d'articles ont été téléchargées.

Nous planifions actuellement des articles spéciaux avec discussion et des numéros spéciaux pour les prochaines diffusions de la revue. En particulier, un article avec discussion de Changbao Wu sur l'inférence pour des échantillons non probabilistes devrait paraître dans le numéro de décembre 2022. Les auteurs qui discuteront l'article sont Michael Bailey, Michael Elliott, Jae-Kwang Kim, Sharon Lohr et Xiao-Li Meng. Ce numéro contiendra également l'article Waksberg 2022 de Roderick Little. Un article spécial avec discussion est actuellement prévu pour honorer la mémoire et les contributions de Jean-Claude Deville. D'autres articles avec discussion et numéros spéciaux sont en cours de planification et pourraient paraître dans les prochains numéros de la revue.

Pour obtenir plus de renseignements, communiquez avec :
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

7.2 Transfert de connaissances — formation en statistique

Le Groupe de travail sur le développement de talent statistique, dont le mandat principal demeure la modernisation de la formation en statistique au sein de la direction et de l'organisme, a connu une autre année occupée et productive. Notre programme de cours en classe est maintenant entièrement adapté à la formation dans un cadre virtuel. Les cours sur le couplage d'enregistrements et la conception de questionnaires ont été les derniers à être adaptés et offerts virtuellement pour la première fois cette année. Le cours sur l'estimation sur petits domaines, le cours sur l'imputation en présence de non-réponse et le cours d'introduction à l'échantillonnage sont d'autres cours dignes de mention qui ont été offerts cette année. Pour ce qui est des nouvelles activités, le groupe continue 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'éthique des données a été offert pour la première fois, par un éminent professeur d'université. Le cours traitait de sujets comme la définition de l'éthique des données, les principales théories éthiques, l'éthique appliquée et les problèmes actuels dans le domaine de l'éthique des données. L'accent était mis sur ce qui concerne Statistique Canada. Le cours a été offert dans les deux langues officielles et d'autres séances sont prévues pour l'an prochain. Le cours a fait l'objet de très bons commentaires et quelques ajustements mineurs ont été suggérés pour les séances à venir.

Comme la science des données et la modélisation des données demeurent des domaines prioritaires au sein de l'organisme, les nouvelles initiatives sont restées principalement axées sur ces sujets. Deux activités de modélisation moderne ont été offertes ou sont en préparation. La première est la répétition d'une activité mise à l'essai l'an dernier et faisant maintenant partie de notre programme. Trois thèmes ont été abordés : la régression linéaire, la régression logistique et la validation croisée. Les participants

sont invités à visionner une série de vidéos, qui sont suivies d'une discussion approfondie avec le groupe et un animateur, et enfin d'une démonstration de l'application de la méthode dans un programme existant de Statistique Canada. Pour la deuxième activité, actuellement en préparation, un format similaire est utilisé pour traiter de la modélisation par apprentissage automatique pour la classification, qui englobe un certain nombre d'approches, comme XGBoost, FastText et les modèles transformeurs (auto-attentifs). De telles activités sont idéales en ce sens qu'elles peuvent être préparées rapidement, offrir de la formation sur les méthodes de pointe et favoriser la participation active des employés.

Le cours pilote sur la science des données à l'intention des gestionnaires a également été achevé cette année. Ce cours s'adressait aux gestionnaires de l'organisme et portait sur ce que les méthodes peuvent et ne peuvent pas faire plutôt que sur la théorie mathématique qui sous-tend les méthodes. Le cours, combiné à un cours plus technique sur les méthodes de la science des données, fera partie du programme régulier pour les années à venir. À mesure que l'utilisation de R s'est répandue au sein de l'organisme, un plus grand nombre de séances du nouveau cours sur R ont également été offertes.

Les faits intéressants sur la méthodologie constituent une autre activité d'apprentissage qui s'est poursuivie cette année. Il s'agit de courtes vidéos, de 30 à 40 minutes, conçues par des employés de Statistique Canada et portant sur un large éventail de thèmes. Les thèmes abordés cette année comprennent la linéarisation appliquée à l'estimateur Hájek, l'inférence conditionnelle appliquée à l'estimateur Hájek et l'estimation pour des domaines.

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

Pour obtenir plus de renseignements, communiquez avec :
Pierre Caron (613-612-6910, pierre.caron@statcan.gc.ca).

7.3 Symposium international sur les questions de méthodologie de Statistique Canada

SOUS-PROJET : Symposium international de 2021 sur les questions de méthodologie de Statistique Canada

Le Symposium international de 2021 sur les questions de méthodologie de Statistique Canada, dont le thème était « Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société », s'est déroulé de façon virtuelle pendant plusieurs vendredis consécutifs, du 15 octobre au 5 novembre 2021, et il a été couronné de succès. Le symposium gratuit a attiré au total plus de 1 000 participants et a offert plus de 50 présentations officielles au cours des quatre séances plénières, des sept séances avec des conférenciers spécialement invités et des sept séances auxquelles des intervenants ont contribué, en plus d'une séance de présentation d'affiches virtuelle « portes ouvertes ». En outre, plus de 500 personnes ont participé à l'un des trois ateliers tenus le jeudi précédant le début officiel de la conférence : « Éthique et vie privée », « Aperçu d'applications en science des données à Statistique Canada » et « Une approche en science des données pour l'estimation des statistiques officielles : exploiter le pouvoir des modèles d'apprentissage automatique ». Les actes du Symposium sont prêts à être catalogués aux fins de publication, et 30 manuscrits ont été soumis, aux fins de publication en ligne dans un format bilingue au cours de l'été 2022. Nous sommes reconnaissants de la participation

de chacun à cet exercice enrichissant de partage de connaissances, et nous avons hâte de mettre en œuvre les leçons retenues de notre tout premier symposium à grande échelle entièrement virtuel.

Pour obtenir plus de renseignements, communiquez avec :
Joseph Duggan (613-371-1410, joseph.duggan@statcan.gc.ca).

SOUS-PROJET : Symposium international de 2022 sur les questions de méthodologie de Statistique Canada

Le Symposium international de 2022 sur les questions de méthodologie de Statistique Canada, dont le thème sera « Désagrégation des données : dresser un portrait de données plus représentatif de la société », s'appuiera sur le succès du symposium de l'an dernier en offrant à nouveau une conférence gratuite sous forme virtuelle. Le symposium de 2022 sera accessible en ligne et, en incluant la première journée d'ateliers, il se déroulera du 2 au 4 novembre 2022. Le symposium devrait à nouveau offrir des séances plénières et des séances parallèles qui porteront sur divers thèmes.

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 du programme a rédigé plusieurs thèmes pour les séances en fonction du thème choisi pour cette année, et il organise les différentes activités entourant le symposium. 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.

Au cours de la première partie de la prochaine année de déclaration, le comité du programme lancera une invitation générale à présenter des communications et il désignera les organisateurs des séances avec des conférenciers spécialement invités. Une ébauche du programme devrait être prête en août. D'ici là, un conférencier de marque sera trouvé, et le discours du lauréat du prix Waksberg sera organisé. Le comité de la logistique configurera la version 2022 du site Web du symposium, et le formulaire de présentation des résumés en ligne pourra être utilisé de nouveau avant l'été. D'autres mises à jour seront apportées au site Web lorsque le comité organisateur décidera d'activer la page d'inscription et à mesure que le programme évoluera.

Vous trouverez de plus amples renseignements à l'adresse :
https://www.statcan.gc.ca/fr/conferences/symposium2022/index.

Pour obtenir plus de renseignements, communiquez avec :
Joseph Duggan (613-371-1410, joseph.duggan@statcan.gc.ca).

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

Baillargeon, J. (2022a). *Bidirectional Pro-Rating*. Document interne disponible sur demande, Statistique Canada.

Baillargeon, J. (2022b). *Optimization for Sample Designs: Progress Report on Research Activities*. Document interne disponible sur demande, Statistique Canada.

Beaulieu, M., Lebrasseur, D. et Gagnon, R. (2021). Measuring and communicating quality for programs using administrative data sources exclusively. *Proceedings of the 2021 European Establishment Statistics Workshop.*

Beaulieu, M., et Gagnon, R. (2021). *Measuring and Communicating Quality for Statistical Programs Based on Administrative Data: The Canadian Housing Statistics Program's Composite Quality Indicator*. Communication présentée au Comité consultatif sur les méthodes statistiques, novembre 2021, Statistique Canada.

Beaumont, J.-F., Bocci, C., et St-Louis, M. (2021). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. *Journal of Survey Statistics and Methodology*.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2022). Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. Présentation à la 2022 Morris Hansen Memorial Lecture, le 1er mars 2022.

Beaumont, J.-F., et Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5, 339–357.

Beaumont, J.-F., et Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and Bayesian approaches. Accepté pour publication dans *Canadian Journal of Statistics.*

Blais, A.-R., Michaud, I., Simard J.-S., Mach, L. et Houle, S. (2021). Mesurer les facteurs psychosociaux en milieu de travail. *Rapports sur la santé*, 32, Statistique Canada.

Brennan, A. (2022). A sensitivity analysis for assessing plausible non-response or participation bias in sample 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*.

Dasylva, A., Goussanou, A. et Nambeu, C.-O. (2021). Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors. Dans *Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

Dasylva, A., et Zanussi, Z. (2021a). *Measuring the Coverage of a Data Source with a Private Set Intersection*. *Rapport interne, Statistique Canada,* Ottawa.

Dasylva, A., et Zanussi, Z. (2021b). A Private Set Intersection Use Case. Presentation at the UNECE Input Privacy-Preserving webinar, Nov. 2021.

Estevao, V.M. (2022). *Robust Estimation – Parameter Description and User Guide, Methodology Specifications*. Document interne, Statistique Canada.

Evans, J., et Oyarzun, J. (2022). Need for speed: Using fastText (Machine Learning) to code the Labour Force Survey. Dans *Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

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

Ferri-García, R., Beaumont, J.-F., Bosa, K., Charlebois, J. et Chu, K. (2021). Weight smoothing for nonprobability surveys. *TEST*.

Gagnon, R., Beaulieu, M., Lebrasseur, D., Qian, W. et Yeung, A. (2021). Creation of a composite quality indicator for estimates based on administrative data using clustering. *Recueil du Symposium international de 2021 sur les questions de méthodologie.*

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, Publié le 6 janvier 2022, https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-fra.htm.

Gauvin, H. (2021). Generating structured microdata: the example of synthesizing hierarchical data. Dans le *Recueil du Symposium international de 2021 sur les questions de méthodologie,* Statistique Canada, Ottawa (à paraître).

Gray, D. (2021). Improving decision-making in imputation design through data visualization. Présenté au *Symposium international de 2021 sur les questions de méthodologie*, Statistique Canada.

Gray, D., et Matthews, S. (2021). *Rethinking the role of Generalized Systems for economic statistics*. Présenté à la 73e réunion du Comité consultatif sur les méthodes statistiques, Statistique Canada.

Hao, Y., et Provençal, J.-S. (2021). *Policy Capturing: An Overview for Treasury Board of Canada Secretariat*. Atelier au Pay Equity Group, Treasury Board of Canada Secretariat, 26 octobre et 2 novembre 2021.

Laprade, J. F., Blanchette, S., Zanussi, Z., Chikhar, O. et Skavysh, V. (2021). Quantum Machine Learning for Text Classification. Présentation d'affiche au 2021 Montreal AI Symposium.

Matthews, S. (2021a). Seasonal Adjustment during the COVID-19 pandemic: Statistics Canada's Approach. *Proceedings of the European Establishment Statistics Workshop*.

Matthews, S. (2021b). *Time Series and Seasonal Adjustment Estimation during the COVID-19 Pandemic*. Round-table session at the 2021 Joint Statistical Meetings of the American Statistical Association.

Matthews, S. (2021c). *Toward Near Real-Time Economic Indicators Using Time Series Models: Statistics Canada's Progress*. Présenté à la Sixth International Conference on Establishment Statistics.

Matthews, S. (2022). *Framework for Development and Production of Advance Indicators at Statistics Canada*. Document interne disponible sur demande, Statistique Canada.

Matthews, S., et Le Moullec, J. (2021). *Development of More Timely Indicators of Gross Domestic Product*. Présenté à la 73e réunion du Comité consultatif sur les méthodes statistiques, Statistique Canada.

Matthews, S., et Saint-Pierre, É. (2022). *Guidelines on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics (v1.1)*. Document interne disponible sur demande, Statistique Canada.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. et Puech, P. (2022a). *QR Prediction for Statistical Data Integration*. Manuscrit inédit en cours d'examen.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. et Puech, P. (2022b). Many-to-one indirect sampling with application to the French postal traffic estimation. Accepté pour publication dans *The Annals of Applied Statistics*.

Neusy, E., Beaumont, J.-F., Yung, W., Hidoroglou, 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.

Picard, F. (2021). Le Nowcasting. *Convergence*, 26, 15-17, Septembre 2021, Association des statisticiennes et statisticiens du Québec.

Sallier, K. (2021). *Statistics Canada's Experience Creating Public Synthetic Datasets Using the FCS and the Synthpop Package*. Présentation au UNECE/HLG-MOS working group on synthetic data.

Statistique Canada (2021). *G-EST 2.03.003 User Guide*. Document interne disponible sur demande, Statistique Canada.

Tremblay, L., Dean, S. et Martineau, P. (2022). Conducting Online Focus Groups: Challenges and Opportunities. SAGE Research Methods: Doing Research Online.

Wright, P., Brisebois, F. et Martineau, P. (2022). *Improving Response by Studying Citizen Participation in Social Surveys*. Communication présentée à la Statistical Society of Canada Annual Meeting, juin 2022.

You, Y. (2022). *An Empirical Study of Hierarchical Bayes Small Area Estimators Using Different Priors on Model Variances*. Document de recherche interne, Statistique Canada, Ottawa.

You, Y., Dasylva, A. et Beaumont, J.-F. (2021). An approximate Bayesian approach to improving probability sample estimators using a supplementary non-probability sample. Recueil du Symposium international de 2021 sur les questions de méthodologie, Statistique Canada.

You, Y., et Hidiroglou, M. (2022). *Application of Sampling Variance Smoothing Methods for Small Area Proportion Estimation*. Document de recherche interne, Statistique Canada, Ottawa.

Zanussi, Z., et Dugdale, C. (2022). *Practical Privacy-Aware Data Linkage and Statistical Aggregation based on Privacy Enhancing Techniques*. Rapport interne soumis pour publication, Statistique Canada, Ottawa.

Zhao, Z. (2021). *Exploring Available Tools to Generate Synthetic Data with High Analytical Value: R Packages Synthpop and SimPop*. Présentation interne, Statistique Canada, Ottawa.