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

Programme de recherche et développement en méthodologie : réalisations, 2019-2020

Date de diffusion : le 29 septembre 2020

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 STATCAN.infostats-infostats.STATCAN@canada.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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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.

Le présent rapport fait la synthèse des réalisations en 2019-2020 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 d'enquête de l'organisme; ce sont des activités qui, autrement, ne s'exerceraient pas dans le cadre des services de méthodologie offerts à ces programmes d'enquête. 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 :

Susie Fortier
(613-220-1948; susie.fortier@canada.ca)

Programme de recherche et développement en méthodologie

Rapport annuel 2019-2020

Table des matières

5. Recherche divisionnaire et autres activités

6. Documents de recherche parrainés par le Programme de recherche et développement en méthodologie

1. Modélisation, intégration des données et science des données

1.1 Estimation sur petits domaines

Les estimations classiques fondées sur le plan de sondage des paramètres de population, ce qu'on appelle les estimations directes, sont généralement fiables, à condition que la taille de l'échantillon dans les domaines d'intérêt ne soit pas trop petite. Les estimations indirectes, qui empruntent de l'information à d'autres domaines ou à d'autres périodes, permettent souvent de réaliser des gains d'efficacité importants dans le cas des petits domaines, le prix à payer étant l'introduction d'hypothèses de modélisation. Ces dernières années, on a observé à Statistique Canada un regain d'intérêt pour l'étude des méthodes d'estimation indirecte par modèle sur petits domaines et un système a été mis au point. Le système et les méthodes en question sont décrits dans Hidiroglou, Beaumont et Yung (2019). Le but ultime est d'utiliser ces méthodes pour la production de statistiques officielles, lorsque cela est jugé approprié. Les principaux objectifs de ce projet sont les suivants :

i) élaborer de nouvelles méthodes d'estimation sur petits domaines qui tiennent compte des problèmes observés dans les enquêtes réelles;
ii) étudier les propriétés des méthodes existantes selon différents scénarios afin de mieux comprendre comment et quand employer ces dernières;
iii) établir une méthodologie d'estimation sur petits domaines convenant à certaines enquêtes prometteuses;
iv) mettre au point et à l'essai des prototypes appliquant des méthodes nouvelles ou existantes susceptibles d'être utiles dans le cadre des programmes statistiques.

Des progrès ont été réalisés dans les sous-projets suivants. En voici une description.

SOUS-PROJET : Modélisation des meilleurs prédicteurs linéaires sans biais empirique (MPLSBE) et modélisation hiérarchique bayésienne (HB) pour l'estimation sur petits domaines dans le cadre de l'EPA avec lissage de la variance d'échantillonnage par opposition à la modélisation

Ce projet vise l'étude et l'évaluation du modèle de Fay-Herriot (FH) au moyen de différentes méthodes de modélisation et de lissage de la variance d'échantillonnage pour l'estimation des caractéristiques de la population active. Nous examinons en particulier les modèles de You et Chapman (2006) et de You (2016) pour l'estimation du taux de chômage et comparons les estimations du modèle aux estimations du recensement. Nous prévoyons établir un modèle approprié pour les petits domaines par modélisation ou lissage de la variance

d'échantillonnage aux fins de l'estimation du taux de chômage dans l'Enquête sur la population active (EPA).

Progrès :

Nous avons étudié les approches du meilleur prédicteur linéaire sans biais empirique (MPLSBE) et du traitement hiérarchique bayésien (HB) pour estimer le taux de chômage à l'aide du modèle de Fay-Herriot avec modélisation ou lissage de la variance d'échantillonnage. Les estimations par modèle obtenues ont fait l'objet d'une comparaison avec les estimations du recensement pour les régions métropolitaines et les agglomérations de recensement (RMR/AR) et différentes tailles d'échantillon. Les résultats démontrent que la méthode de lissage de la fonction de variance généralisée (FVG) de You et Hidiroglou (2012) est la plus efficace dans l'estimation sur petits domaines avec les données de l'EPA. La modélisation hiérarchique bayésienne par les modèles loglinéaires de You (2016) se révèle aussi efficace par rapport aux estimations directes d'enquête pour ce qui est de la réduction du coefficient de variation (CV) et du biais. Employer une variance d'échantillonnage non lissée avec le MPLSBE de Wang et Fuller (2003) donne un piètre résultat, plus particulièrement lorsque la taille de l'échantillon est petite. À nos yeux, le lissage de la variance d'échantillonnage est nécessaire et constitue une étape importante dans l'estimation du taux de chômage par le modèle de Fay-Herriot si l'approche MPLSBE est retenue. Dans le cas d'une modélisation hiérarchique bayésienne, nous recommandons de modéliser la variance de l'échantillonnage à l'aide d'un modèle loglinéaire FVG. Les modèles et les résultats sont résumés dans un rapport de recherche (You, 2020a).

SOUS-PROJET : Estimation par modèle des totaux de l'EPA

Dans ce projet, nous étudions le modèle linéaire de Fay-Herriot et les modèles non linéaires non appariés pour l'estimation des totaux de l'EPA, dont le total de l'emploi. Nous prévoyons évaluer les estimations étalonnées ou non pour les totaux et les rapprocher des estimations du recensement. Cette étude fournira des modèles pour l'estimation des totaux et permettra d'évaluer l'utilisation de modèles loglinéaires non appariés et d'une procédure possible d'étalonnage dans l'estimation de totaux par modèle.

Progrès :

Le modèle log linéaire non apparié de Fay-Herriot (You et Rao, 2002), qui est fondé sur les méthodes hiérarchiques bayésiennes, est appliqué aux données des totaux de l'EPA en vue de l'estimation des totaux de l'emploi, de la population active et du chômage au niveau des RMR/AR. Dans l'estimation des totaux, le modèle linéaire de Fay-Herriot ne donne pas un bon résultat en cas de modélisation directe. Une transformation des taux ou des moyennes est

nécessaire si le modèle de Fay-Herriot est employé. Dans l'estimation du total de l'emploi, un modèle loglinéaire non apparié par traitement bayésien hiérarchique fonctionne très bien et se révèle efficace en réduisant à la fois le biais et le coefficient de variation. Les distributions antérieures gamma inversent et par modèle loglinéaire de You (2016) produisent toutes deux des résultats semblables dans l'estimation des totaux. Le modèle loglinéaire non apparié fonctionne un peu mieux que le traitement par transformation, car le coefficient de variation obtenu est légèrement moindre. L'étalonnage ne diminue pas nécessairement le biais, puisque le modèle est adéquat pour les données. Des programmes et des rapports sommaires ont été écrits et produits (You, 2020b).

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

Ces dernières années, on a utilisé davantage le modèle de Fay-Herriot à Statistique Canada pour produire des estimations sur petits domaines. Ces estimations s'accompagnent habituellement d'estimations de l'erreur quadratique moyenne (EQM) par modèle. Toutefois, les utilisateurs sont habitués aux estimations de l'EQM par plan. L'avantage de l'EQM par plan sur l'EQM par modèle réside dans le fait de ne pas écarter par intégration la spécificité d'un domaine donné. Cette estimation par plan pourrait être plus intéressante pour les utilisateurs comme indicateur de la qualité des estimations. On sait que les estimations par plan de l'EQM sont instables (voir, par exemple, Rao, Rubin-Bleuer et Estevao, 2018). Nous prévoyons étudier le recours à une approche conditionnelle pour obtenir un estimateur plus efficace de l'EQM par plan.

Progrès :

Nous avons élaboré un estimateur de l'EQM par plan dans un traitement conditionnel. La théorie en question a d'abord été conçue pour le cas du meilleur prédicteur où on suppose que tous les paramètres du modèle sont connus. On s'attend à ce que cet estimateur soit plus efficace que l'estimateur sans biais de plan de l'EQM par plan. Le cas du meilleur prédicteur linéaire sans biais est actuellement à l'étude. Pour le meilleur prédicteur linéaire sans biais empirique (MPLSBE) où tous les paramètres du modèle font l'objet d'une estimation, nous avons conçu une procédure bootstrap d'estimation de l'EQM par plan. Notre traitement conditionnel sera évalué dans une étude en simulation. Un rapport interne provisoire a été produit (Beaumont, Lesage et Rao, 2020).

SOUS-PROJET : Évaluation de la robustesse du modèle de Fay-Herriot pour les petits domaines

Il est bien connu que les enquêtes probabilistes permettent d'obtenir des estimations fiables pour des domaines de la population pour lesquels la taille d'échantillon est suffisamment grande. La demande d'estimations pour des domaines de plus en plus fins s'est accrue dans

les dernières années. Pour répondre à cette demande, sans augmenter drastiquement la taille d'échantillon globale et les coûts de collecte, on peut avoir recours à des techniques d'estimation sur petits domaines. Ces techniques reposent sur la détermination d'un modèle qui relie les données d'enquête à des données auxiliaires. Le modèle de Fay-Herriot est le plus fréquemment utilisé en pratique. Si le modèle est correctement spécifié, l'approche est valide et peut mener à des augmentations de précision importantes. Que se passe-t-il dans le scénario réaliste où le modèle ne tient pas la route parfaitement ? L'objectif de ce projet est d'évaluer et quantifier, théoriquement et/ou par simulations, les conséquences d'une mauvaise spécification du modèle sur le biais et la variance des estimations sur petits domaines, particulièrement pour les plus petits domaines. Différents types de mauvaise spécification du modèle sont évaluées.

Progrès :

Nous avons entrepris une étude en simulation à l'aide de divers scénarios pour vérifier les effets de la non-linéarité de la moyenne du modèle. Une grande conclusion est que les estimations sur petits domaines demeurent plus efficaces que les estimations directes en cas d'erreur de spécification du modèle, plus particulièrement dans le cas des domaines les plus petits. Un rapport de Buresi (2019) renseigne plus en détail sur ces résultats.

SOUS-PROJET : Diagnostics locaux pour le modèle de Fay-Herriot

Les outils de validation de modèle sous forme par exemple de graphiques des résidus servent souvent à évaluer la vraisemblance du modèle de Fay-Herriot. Les estimations d'erreur quadratique moyenne (EQM) par modèle permettent alors d'évaluer les gains d'efficacité que présentent les estimateurs sur petits domaines par rapport aux estimateurs directs. Toutes ces techniques sont utiles s'il s'agit de jauger le rendement global des estimations sur petits domaines. Il reste que les utilisateurs s'intéressent souvent à leur seul domaine bien précis et qu'un indicateur de qualité de leur estimation propre présente plus d'intérêt pour eux. C'est le but que réalise partiellement l'EQM par modèle, mais en écartant par intégration l'effet aléatoire local (erreur de modèle de couplage) qui intéresse les utilisateurs d'un domaine particulier. L'EQM par plan serait plus utile à ces utilisateurs, mais on sait que les estimations sans biais de plan de l'EQM par plan sont très instables (voir, par exemple, Rao, Rubin-Bleuer et Estevao, 2018). Le but du présent projet est de concevoir et d'étudier de nouveaux diagnostics d'effet local dans l'évaluation des estimations sur petits domaines.

Progrès :

Dans des recherches antérieures, nous avons conçu deux nouveaux diagnostics d'effet local. Ils font tous deux intervenir le résidu centré de chaque domaine, ainsi que le rapport entre la

variance de modèle de couplage et la variance de modèle d'échantillonnage. Dans la dernière année, nous avons produit une étude qui a été soumise à Techniques d'enquête (Lesage, Beaumont et Bocci, 2020).

SOUS-PROJET : Estimation sur petits domaines des indicateurs de santé dans les quartiers en Ontario

Dans ce projet, nous avons examiné la capacité des techniques de traitement de petits domaines à estimer 19 indicateurs de santé (proportions) dans 147 quartiers ontariens à l'aide des données annuelles de l'Enquête sur la santé dans les collectivités canadiennes (ESCC). Cette enquête n'est pas conçue pour produire des estimations fiables de toutes les proportions en question à ce niveau. Nous avons modélisé les estimations d'enquête comme fonction des variables du recensement pour produire des estimations sur petits domaines par le cadre méthodologique de Fay-Herriot.

Progrès :

Nous avons produit des estimations sur petits domaines des 19 proportions visées. Nous avons produit des documents internes avec un exposé de stratégie générale et un résumé des résultats. Parfois, l'erreur quadratique moyenne était bien moindre pour les estimations sur petits domaines que pour les estimations directes de l'ESCC.

SOUS-PROJET : Estimation sur petits domaines des caractéristiques du travail selon les zones de travail autonomes

Il est possible d'établir des estimations directes des taux de chômage et des totaux de l'emploi à partir de l'Enquête sur la population active (EPA) à divers niveaux d'agrégation pour une période quelconque. Deux demandes ont été présentées en vue d'étudier la faisabilité de la production d'estimations sur petits domaines au niveau fin des zones de travail autonomes pour lesquelles les estimations directes de l'EPA ne sont normalement pas fiables. Nous avons utilisé le modèle de Fay-Herriot au niveau des domaines pour produire des estimations annuelles et mensuelles des taux de chômage et des totaux de l'emploi par une modélisation des estimations de l'EPA avec les estimations démographiques et les données de l'assurance-emploi comme information auxiliaire.

Progrès :

Nous avons produit pour les années 2011 à 2016 des estimations sur petits domaines des taux de chômage et des totaux de l'emploi en valeur annuelle selon les zones de travail autonomes et nous les avons soumises aux analystes.

Nous avons produit des estimations semblables des taux de chômage du 4e au 11e mois de 2019 pour chaque zone de travail autonome. Deux documents de Bocci et Beaumont (2019a, 2019b) décrivent le cadre méthodologique ayant servi à produire ces estimations sur petits domaines; ils présentent des diagnostics, des mesures de la qualité et des comparaisons. Les travaux se poursuivent en ce qui concerne l'estimation sur petits domaines du total de l'emploi.

SOUS-PROJET : Cours traitant de l'estimation sur petits domaines

Ce projet visait à créer un cours traitant des principes fondamentaux de l'estimation et du diagnostic sur petits domaines, ainsi que d'un système logiciel conçu par Statistique Canada pour produire de telles estimations dans la pratique.

Progrès :

Le cours a été mis au point et présenté en anglais. Il a été traduit en français. Après révision de la traduction, la version française sera prête.

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

Bibliographie

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

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2016). Modélisation hiérarchique de la variance d'échantillonnage de Bayes pour l'estimation sur petits domaines basée sur des modèles au niveau des zones avec applications. Document de recherche de la Direction de la méthodologie, ICCSMD-2016-03-E.

You, Y., et Chapman, B. (2006). Estimation sur petits domaines à l'aide de modèles au niveau des zones et des variances d'échantillonnage estimées. *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.

You, Y., et Hidiroglou, M. (2012). Méthodes de lissage de la variance d'échantillonnage pour les estimateurs de proportion de petits domaines. Document de travail de la Direction de la méthodologie, SRID-2012-08E, Statistique Canada, Ottawa, Canada.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics, 30, 3-15.*

1.2 Estimation en temps réel par la modélisation de séries chronologiques

D'autres projets liés aux séries chronologiques sont décrits à la section 4.6.

SOUS-PROJET : Modélisation des séries chronologiques et progrès de l'estimation en temps réel

La modélisation des séries chronologiques (en prévisions immédiates) est considérée comme un moyen possible de produire des indicateurs économiques avancés pour l'organisme. Nous avons sondé des modèles devant permettre de produire des indicateurs avancés et examiné la possibilité de présenter des indicateurs en temps réel qui pourraient fréquemment être mis à jour en fonction de l'information disponible la plus récente.

Progrès :

Nous avons recensé les études pour relever les modèles couramment utilisés en prévisions immédiates (modèles ARMMI et modèles à composantes inobservées) et enrichir notre connaissance des méthodes et des outils disponibles pour leur application. De nouveaux travaux ont permis d'étudier d'autres modèles disponibles en R (TBATS, ARMMI robustes et lissage exponentiel pour la prévision) ainsi que les modèles d'apprentissage automatique. Une version provisoire de lignes directrices de pratique pour l'établissement de prévisions immédiates a vu le jour (Matthews, Patak, Picard et Mischler (2020)). Ce document recense les familles disponibles de modèles de séries chronologiques avec leurs avantages et leurs inconvénients dans le contexte des prévisions immédiates; il examine la question de la production de prévisions immédiates dans le cas d'un organisme statistique national.

Un certain nombre de critères sont proposés en vue de juger si un estimateur avancé devrait être produit dans le contexte des statistiques officielles. Nous avons procédé à un certain nombre d'exercices de validation de principe portant sur la faisabilité des méthodes en question dans certains projets et dans des circonstances variables. L'équipe a évalué la modélisation des séries chronologiques et les méthodes d'établissement de prévisions immédiates avec les données mensuelles sur le commerce de détail (Matthews et Patak, 2020). L'évaluation consistait notamment à utiliser des prédicteurs tirés d'indicateurs avancés sous forme de produits statistiques, de données secondaires ou de réponses partielles à des enquêtes afin de quantifier le compromis entre précision et actualité pour des modèles fondés sur des sources d'information auxiliaire. Les résultats de ces travaux ont été présentés aux analystes de différents secteurs de Statistique Canada, ce qui devait mener à des applications des méthodes dans d'autres contextes. Ainsi, la modélisation a été appliquée à d'autres indicateurs. Les techniques ont été employées sur des données mixtes de fréquence pour

produire et mettre à jour des prévisions immédiates, les résultats ont été comparés à ceux des techniques d'apprentissage automatique et l'utilisation d'indicateurs de sentiment comme variable auxiliaire a été étudiée.

L'équipe a continué à étendre l'application d'outils de prévision de haute performance en SAS à des fins d'utilisation générale en se proposant aussi d'établir une plateforme de traitement stable et d'améliorer la compréhension des méthodes. Il y a eu étude d'un certain nombre de caractéristiques du processus de sélection de modèle, ce qui a permis d'obtenir une certaine rétroaction au SAS, c'est-à-dire de signaler de légères incohérences et de clarifier les questions techniques au sujet de l'incidence des événements sur les degrés de liberté. Le logiciel continue à appuyer la modélisation des non-répondants critiques, à cerner les ruptures dans les séries et à permettre d'autres applications de modélisation.

Pour une constatation efficace et intuitive des tendances saisonnières dans les modèles de prévision, nous avons progressé dans la voie menant à un test de saisonnalité avec une périodicité souple (non paramétrique) pour dégager les tendances saisonnières dans les données quotidiennes, hebdomadaires et mensuelles. Nous avons envisagé une double approche de ré échantillonnage et de comparaison multiple. Nous avons réalisé et documenté une étude d'un test non paramétrique de saisonnalité fondé sur un rééchantillonnage pour produire une fonction de distribution empirique (Lapointe et Mischler, 2019).

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

1.3 Science des données – Apprentissage automatique

SOUS-PROJET : Modélisation des opioïdes

Il existe actuellement au Canada une véritable crise des surdoses et des décès aux opioïdes. Il faut comprendre les circonstances plus générales que vivent les gens avec les opioïdes pour mieux éclairer l'adoption de politiques visant à traiter les facteurs en amont. Pour répondre à ce besoin d'information, Statistique Canada, en collaboration avec plusieurs organismes britanno-colombiens (notamment le British Columbia Opioid Steering Committee) a créé une source complète de données couplées réunissant une information administrative pour mieux comprendre les déterminants socioéconomiques des victimes d'événements indésirables liés aux opioïdes (décès et surdoses). Le but est de trouver des techniques d'intelligence artificielle et d'apprentissage automatique pour appuyer les analyses de prédiction et de trajectoire à l'aide de mégadonnées structurées portant sur ces événements indésirables à titre d'étude de cas.

Progrès :

Nous avons réalisé une analyse d'hétérogénéité de la cohorte étudiée. Il s'agissait d'une analyse en grappes faisant appel à 15 variables discontinues et à 2 variables continues. Comme les variables mêlent le discontinu et le continu, la méthode à *k* prototypes de Huang (1998) a servi à l'analyse en grappes. Pour juger du nombre de grappes à prévoir, nous avons exécuté neuf cycles indépendants de mise en grappes à *k* prototypes (pour 2 à 10 grappes); nous avons effectué des diagnostics pour jauger l'optimalité du nombre de grappes ainsi que l'expertise en la matière. Pour évaluer la stabilité des grappes (en ce qui concerne les initialisations aléatoires), nous avons exécuté 10 autres cycles indépendants de cette mise en grappes à *k* prototypes. Pour chaque paire de grappes, nous avons calculé deux mesures de similarité. Nous avons évalué la stabilité au moyen du graphique de densité et de l'histogramme résultants des deux scores de similarité. Nous avons vérifié si l'analyse de stabilité des grappes par l'un ou l'autre de ces scores livrerait la même conclusion qu'avec l'autre score en produisant le diagramme de dispersion des paires de grappes des deux et en constatant que les deux scores épousent très étroitement une transformation non linéaire strictement croissante l'un de l'autre. Les grappes ainsi obtenues présentent des profils distincts des événements de la vie qui corroborent fortement ce que décrivent les études de santé publique. Un manuscrit est en préparation sur les résultats de l'analyse en grappes. Une étude aura lieu par la suite en vue d'estimer les risques conditionnels des événements liés aux opioïdes par grappe, le but étant d'étudier la faisabilité de la prédiction d'événements imminents aux opioïdes par les antécédents récents d'événements dans la vie de chaque personne.

Bibliographie

Huang, Z. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 1998*; volume 2.

SOUS-PROJET : Prédiction des rendements culturaux

Statistique Canada diffuse des estimations annuelles de rendement des cultures (production par unité de superficie de terres en récolte) à la fin de chaque année de référence. Il publie en outre des prédictions de rendement des cultures sur l'année plusieurs fois dans la seconde moitié de l'année de référence. Pour les prédictions de juillet 2019, il a appliqué une méthode par modèle, principalement une sélection de variables par le système LASSO, suivie d'une régression linéaire robuste, pour éliminer progressivement la question sur les rendements culturaux dans l'enquête de juillet dans une province, soit le Manitoba. Dans ce projet, on examine si certaines techniques d'apprentissage automatique peuvent livrer de meilleures prédictions que la méthode actuellement employée dans les prédictions de rendement au Manitoba en juillet 2019 à l'aide de mesures longitudinales de janvier à juillet du degré de végétation au niveau des parcelles, de données sur les conditions météorologiques locales, de précisions sur les types de cultures des parcelles, etc.

Progrès :

Pour évaluer et dégager un meilleur modèle de prédiction des rendements que le modèle actuel, nous devons tenir compte de la contrainte opérationnelle suivante : le modèle d'apprentissage à déployer chaque année ne peut avoir été formé que par les données d'années strictement antérieures. Nous avons donc élaboré un protocole de validation prospective où, pour une année, les données qui s'y rattachent servent de données de validation et où les données des années d'une période fixe d'apprentissage strictement antérieure à l'année en question servent de données d'apprentissage. Nous avons répété pour plusieurs années de validation. Des données remontant à l'an 2000 étaient disponibles. Nous avons employé les données de 2000 à 2017 pour l'apprentissage et la validation et les données de 2018 et 2019 comme test. Nous avons appliqué plusieurs techniques d'apprentissage automatique comme XGBoost, Support Vector Machine (SVM), Random Forest et glmnet. Nous avons fait une recherche par quadrillage pour bien régler les hyperparamètres. Nous avons élaboré deux mesures pour comparer les résultats à ceux de la méthode actuelle, à savoir l'erreur relative pondérée par la production moyenne et l'écart-type de l'erreur relative en pondération par la production moyenne. Grâce à ces mesures, nous avons pu démontrer que les modèles d'apprentissage automatique les plus performants innovent en réalité sur le

modèle en place. Nous sommes en train de mettre cette nouvelle approche en exploitation pour l'année de référence 2021 avec un essai parallèle en 2020.

SOUS-PROJET : Approche d'apprentissage automatique pour la classification statistique

Nous désirons améliorer l'efficacité et l'exactitude de la classification statistique par laquelle le texte fourni par un répondant est catégorisé (ou codé) de manière à faciliter la totalisation et l'inférence. Ainsi, les réponses textuelles aux questions sur le titre du poste et les principales fonctions sont régulièrement converties en une classe professionnelle (code numérique à quatre chiffres) de la Classification nationale des professions. Les algorithmes d'apprentissage automatique promettent un rendement supérieur à celui du codage manuel classique où quelqu'un examine le texte fourni et choisit un code, ce qui assurera une exactitude comparable, mais plus rapidement et à moindre coût.

Progrès :

Nous avons ajouté au logiciel G-code des fonctions automatiques de codage fondées sur l'apprentissage automatique et plus particulièrement sur fastText. Nous avons procédé à divers essais et évaluations avec des textes d'entrée tirés d'enquêtes et de recensements antérieurs pour divers systèmes de classification (industries, professions, principaux domaines d'études, etc.).

SOUS-PROJET : Projets d'apprentissage automatique pour le Groupe de haut niveau sur la modernisation des statistiques officielles (GHN-MSO)

Le projet d'apprentissage automatique du Groupe de haut niveau pour la modernisation des statistiques officielles (GHN-MSO) a débuté au printemps 2019 et se compose de trois lots de travaux: 1) Preuve de concepts; 2) Mesure de la qualité et 3) Intégration de l'apprentissage automatique (ML) dans les organisations. Statistique Canada a participé aux trois modules de travail à des degrés divers. Notre participation sera discutée dans ce rapport.

Le but de la preuve de concepts est de partager les connaissances, l'expérience et le code (le cas échéant) entre les groupes impliqués dans le projet. La preuve de concepts couvre de multiples utilisations différentes du ML telles que la classification, l'imputation, le traitement d'image et l'analyse des sentiments. L'objectif du module de travail sur la mesure de la qualité est de développer un cadre de qualité qui permettra aux statisticiens officiels de comparer les algorithmes statistiques, y compris ceux d'apprentissage automatique. Le cadre sera basé sur des concepts existants mais sera adapté aux algorithmes statistiques. Enfin, l'objectif du module de travail Intégration est d'identifier les meilleures pratiques et de guider les

organisations statistiques nationales novices en apprentissage automatique sur la meilleure façon de l'intégrer dans leur environnement.

Progrès :

Dans le cadre de l'ensemble de preuve de concept, les employés de Statistique Canada ont participé à une preuve de concept de codage et de classification en vue de l'intégration de l'algorithme FastText dans notre outil de codage généralisé (G-CODE). Le nouvel algorithme a été utilisé pour coder l'industrie et la profession pour deux enquêtes sur la santé à Statistique Canada et les résultats, les expériences et le code ont été partagés avec le groupe d'apprentissage automatique HLG-MOS. Les employés de la Direction générale de la méthodologie ont également fourni une assistance ou des conseils sur deux preuves de concept entreprises par d'autres pays. La preuve de concept de l'Office des statistiques nationales (ONS) a étudié l'utilisation d'arbres de décision pour élaborer des règles de vérification pour l'une de leurs enquêtes sur les dépenses des ménages. Une étude similaire a été entreprise dans le cadre du Programme de données fiscales agricoles de Statistique Canada et le méthodologiste impliqué a fourni des conseils à l'ONS. Une preuve de concept de la Belgique a étudié l'utilisation de l'apprentissage automatique pour prédire les bilans énergétiques afin de produire des données plus opportunes. La section des séries chronologiques de la Branche de la méthodologie a travaillé avec les chercheurs en Belgique pour voir comment les méthodes d'apprentissage automatique se comparent aux méthodes de séries chronologiques. Les résultats ont été obtenus et sont partagés avec les chercheurs en Belgique.

Un employé de la succursale dirige le lot de travail sur les mesures de la qualité qui produira un cadre de qualité pour les algorithmes statistiques. Le cadre comportera les cinq dimensions suivantes: interprétabilité, précision, rapidité, rentabilité et reproductibilité. Une version finale du cadre devrait être disponible à l'automne 2020.

Pour le package d'intégration, un petit sprint en face à face a été organisé à Statistique Canada en octobre 2019. Les participants du sprint étaient les trois responsables du lot de travail et le gestionnaire de projet. Au cours du sprint, les participants ont eu l'occasion de rencontrer le statisticien en chef et le statisticien en chef adjoint, Stéphane Dufour, qui copréside le conseil exécutif du HLG-MOS. Le principal résultat du sprint était une feuille de route pour le lot de travail 3. Le lot de travail 2 des employés de la Direction générale était l'organisateur principal du sprint et y a également participé.

Pour obtenir plus de renseignements, communiquez avec :
Yanick Beaucage (613-854-2397, yanick.beaucage@canada.ca).

1.4 Intégration des données – Couplage d'enregistrements

Le couplage d'enregistrements joue un grand rôle dans la production des statistiques officielles. Il s'expose cependant aux erreurs, parce qu'il fait souvent appel à des quasi-identificateurs non uniques qui sont enregistrés avec des variantes et des erreurs typographiques. Dans le présent projet, nous examinons la production et l'utilisation de données couplées, et notamment l'estimation fidèle des erreurs de couplage. D'autres activités dans le domaine du couplage d'enregistrements sont décrites à la section 4.1.

SOUS-PROJET : Estimation des erreurs de couplage dans un contexte de protection de la vie privée

On a conçu diverses méthodes pour coupler des données chiffrées, dont les méthodes par filtrage de Bloom (Kroll, Niedermeyer, Schnell et Steinmetzer, 2014). Un grand défi à cet égard a cependant été l'estimation fidèle des erreurs de couplage et le problème lié de l'établissement des paramètres de couplage. Les solutions adoptées par le passé faisaient appel à des données d'apprentissage qui étaient souvent indisponibles.

Progrès :

Le projet s'est attaqué à ce problème en faisant la démonstration d'une estimation fidèle des erreurs de couplage dans un contexte de protection de la vie privée, et ce, sans données d'apprentissage. On y a appliqué un nouveau modèle d'erreur en fonction du nombre de voisins d'un enregistrement donné (Dasylva, Goussanou, Ajavon et Abousaleh, 2019). On a créé une population finie avec des noms, des dates de naissance, des adresses, etc. On a créé deux registres complets sans duplication et procédé au couplage à l'aide de filtres de Bloom (Kroll et coll., 2014) en prévoyant des critères de blocage fondés sur les tout premiers bits et des comparaisons par l'indice de Dice. Les paramètres de couplage étaient le nombre de bits servant au blocage et le seuil de l'indice de Dice. Les paramètres ont été réglés à diverses valeurs et, pour chaque paramétrage, on a estimé les taux d'erreur et les a comparés aux taux réels en fonction de l'état d'appariement effectif. Dans chaque réglage, les taux estimatifs étaient proches des taux réels. Ces résultats démontrent la possibilité de coupler des données chiffrées en toute précision sans données d'apprentissage. Les détails sont présentés par He (2019).

SOUS-PROJET : Analyse de survie avec données couplées

Les erreurs de couplage sont source de biais en analyse, notamment dans l'analyse de survie. La correction de ces erreurs présente de nombreux défis, dont une estimation fidèle des erreurs et la justesse de la méthode de correction compte tenu des taux estimés.

Progrès :

Dans ce projet, nous avons examiné l'étude de mortalité des cohortes par couplage de la Base canadienne de données sur les décès et de l'Enquête sur la santé dans les collectivités canadiennes. Nous y avons mis à l'essai deux nouvelles méthodes d'analyse de survie avec des données couplées, dont une nouvelle méthode d'estimation d'erreur (Dasylva et coll., 2019) et une nouvelle méthode de correction (Dasylva, 2018). Nous avons dégagé un certain nombre de questions à examiner dans de futurs travaux. Une première difficulté est la longue durée d'exécution de la procédure numérique servant à calculer les paramètres corrigés de survie. Une autre est de tenir compte du plan de sondage de l'Enquête sur la santé dans les collectivités canadiennes au moment d'estimer les erreurs de couplage. Les détails sont présentés par Miller (2019).

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

Bibliographie

Dasylva, A. (2018). Pairwise Estimating Equations for the Analysis of Linked Data, thèse de doctorat, Carleton University, 2018.

Kroll, M., Niedermeyer, F., Schnell, R. et Steinmetzer, S. (2014). Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 6, 59-79.

1.5 Méthodes mixtes et approche qualitative

SOUS-PROJET : Engagement international – Évaluation par les méthodes mixtes

Statistique Canada exerce des activités internationales pour diverses raisons. Pour un organisme statistique de premier plan, apporter un leadership et une expertise technique à la communauté internationale est un effort de tous les instants. De tels efforts sont surveillés de près par le programme d'engagement international.

Jusqu'à présent, la surveillance exercée par ce programme a privilégié les éléments de mesure quantitative. Il y a une déclaration trimestrielle de caractéristiques comme le nombre de participants, la nature des rôles, le coût des activités, les frais de déplacement et la participation par secteur. Toutefois, on désire de plus en plus comprendre davantage la valeur de cet engagement international au-delà des statistiques quantitatives du programme. La présente étude recourt à la recherche par les méthodes mixtes pour nous éclairer à ce sujet. C'est une recherche qui tire parti des atouts possibles des méthodes tant qualitatives que quantitatives pour analyser des sujets dans leur complexité.

Progrès :

L'évaluation du programme d'engagement international est terminée et ses résultats en décrivent les forces et les faiblesses. Le projet réalisé a consisté en interviews cognitives auprès des participants à ce programme et en collecte de données quantitatives, ces deux volets étant intégrés dans l'optique retenue de recherche par les méthodes mixtes. Une analyse rétrospective a permis de jauger la faisabilité d'une application généralisée des méthodes mixtes à Statistique Canada.

SOUS-PROJET : Étude par les méthodes mixtes de la mesure des facteurs psychosociaux de santé mentale en milieu de travail

Le groupe du Projet de mesure du rendement en matière de santé mentale en milieu de travail, codirigé par Services publics et Approvisionnement Canada et Statistique Canada, en consultation avec les professionnels des ressources humaines et les psychologues organisationnels, a entrepris un exercice collectif de schématisation des concepts dans une analyse des indicateurs liés à 13 facteurs psychosociaux dans les organismes fédéraux, et ce, à l'aide de diverses sources d'information, dont le Sondage auprès des fonctionnaires fédéraux (SAFF). Cette étude visait principalement à présenter une méthode rigoureuse d'évaluation de la mise en correspondance proposée des concepts entre les éléments du SAFF de 2017 et les 13 facteurs psychosociaux.

Progrès :

Nous avons adopté une approche d'enquête sociale par les méthodes mixtes (Greene, 2007) avec triangulation et analyse de complémentarité. D'abord, nous avons mené un exercice de mise en correspondance des concepts (sur le mode qualitatif) afin de voir dans quelle mesure les éléments du SAFF s'harmonisaient avec les 13 facteurs psychosociaux selon la perception d'un groupe choisi de fonctionnaires. Cet exercice de schématisation conceptuelle était différent de l'exercice précédent du groupe de projet, en ce sens qu'il était demandé aux participants de faire l'exercice indépendamment plutôt que collectivement. Nous avons calculé des cotes d'accord entre les évaluateurs. En second lieu, nous avons procédé à des analyses de facteurs exploratoires et confirmatoires (sur le mode quantitatif) pour examiner si les éléments du SAFF déjà mis en correspondance mesuraient convenablement les 13 facteurs psychosociaux. Nous avons pris divers critères en considération pour évaluer l'ajustement du modèle. Nous avons mis en triangulation et en rapprochement de complémentarité (par les méthodes mixtes) les résultats de l'exercice de schématisation conceptuelle et d'analyse factorielle pour évaluer la mise en correspondance des concepts proposés par le groupe de projet. La conclusion a été qu'il faudrait sans doute développer les éléments d'information et envisager d'autres sources de données pour mieux mesurer les facteurs psychosociaux.

Un rapport détaillé (Arim, Bougie, Michaud, Tabuchi, Yung et Kohen, 2019) a été préparé pour le groupe du Projet de mesure du rendement en santé mentale en milieu de travail.

Pour obtenir plus de renseignements, communiquez avec :
Laurie Reedman (613-894-2779, laurie.reedman@canada.ca).

Bibliographie

Greene, J.C. (2007). Mixed Methods in Social Inquiry. San Francisco: Jossey-Bass.

2. Confidentialité

La recherche sur la confidentialité à Statistique Canada continue de se concentrer sur le développement de nouvelles méthodes et idées qui offrent des formes alternatives d'accès tout en continuant de s'assurer que les renseignements personnels des particuliers et des entreprises ne sont divulgués d'aucune façon. Le groupe du Centre pour la confidentialité et l'accès de Statistique Canada continue également d'offrir des services de consultation aux partenaires internes et externes afin d'aider à développer la capacité de détection et de traitement des risques de divulgation.

2.1 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 dans le cas des variables continues observées dans le cadre des enquêtes économiques.

Progrès :

À la fin de 2018-2019, Statistique Canada a fait un grand pas dans la mise à disposition de cette stratégie de rechange en diffusant les données de son Enquête sur l'innovation et les stratégies d'entreprise (https://www150.statcan.gc.ca/n1/daily-quotidien/190326/dq190326b-fra.htm). C'était la première occasion qui s'offrait d'appliquer la méthode de l'ajustement tabulaire aléatoire (ATA).

En 2019-2020, les chercheurs ont tiré parti de ce succès en voulant appliquer la méthode à un plus large éventail de produits de Statistique Canada, l'organisme s'étant donné un objectif de suppression zéro pour les données de ses enquêtes économiques. Le principal défi étudié était d'ajouter une fonction de bruit corrélé. Il y avait à cela plusieurs avantages : maintien des structures de corrélation entre les variables apparentées, atténuation de l'incidence du bruit ajouté sur les cellules agrégées grâce à l'introduction d'un bruit négatif dans les cellules connexes, préservation des tendances dans les enquêtes répétées. Un document est en cours qui décrira le nouveau cadre méthodologique. Un nouvel outil SAS est en chantier en vue de l'application de cette méthode. Nous sommes enfin en train de concevoir un matériel didactique pour former une base d'utilisateurs.

2.2 Données synthétiques

Fidèle à son souci d'offrir de nouvelles possibilités d'accès, Statistique Canada investit dans la recherche de méthodes de création de données synthétiques. Ces données peuvent prendre diverses formes et présenter différentes caractéristiques de qualité, mais visent toujours à offrir une possibilité d'accès aux microdonnées sans autre risque de divulgation, d'où une diffusion grand public.

Progrès :

Le but en 2019-2020 était de se donner des principes fondamentaux clairs quant à la terminologie de la synthèse de données en mettant l'accent sur la création de données synthétiques d'une grande valeur analytique. Le premier défi était de décrire les données synthétiques et de les distinguer des fichiers « fictifs » revêtant un même caractère synthétique. La mise à jour d'un document d'orientation est en cours.

Les enquêtes-échantillons présentent des défis particuliers, qu'il s'agisse du traitement des poids d'échantillonnage ou de la justesse de l'analyse. Il a été question des difficultés de la création de données synthétiques d'une haute valeur analytique à la réunion 68 du Comité consultatif des méthodes statistiques de Statistique Canada en avril 2019.

Le 2 novembre 2019, Le Partenariat canadien contre le cancer s'est réuni à l'occasion du marathon de programmation de Hack4Cancer un jour avant la Conférence canadienne sur la recherche sur le cancer à Ottawa (Ontario). Statistique Canada a élaboré les données synthétiques de cette séance; c'était la seconde fois que des données de cette nature étaient diffusées publiquement par l'organisme. Des préoccupations particulières, en matière de contraintes d'utilité par exemple, ont été soulevées et commanderont désormais l'attention (https://cc-arcc.ca/hack4cancer_hackathon/).

Le succès du travail qui se fait à Statistique Canada dans ce domaine de recherche a été souligné par Kenza Sallier, lauréate du prix des jeunes statisticiens de l'International Association for Official Statistics pour son article « Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis » (Sallier, 2020).

2.3 Soutien et consultation des utilisateurs

Statistique Canada continue à appuyer ses partenaires internes et externes dans les questions de confidentialité et d'accès.

Progrès :

Le programme des centres de données de recherche permet à des chercheurs de confiance d'accéder aux microdonnées détaillées dans des conditions de sécurité. Le programme de recherche appuie à cette fin les utilisateurs et élabore des règles de contrôle pour les produits analytiques et le soutien de projets particuliers en matière de confidentialité.

Le système généralisé G-Confid est la solution généralisée qu'a adoptée Statistique Canada pour évaluer et traiter les risques de confidentialité dans le cas des variables continues qui se présentent normalement dans le contexte des enquêtes économiques. Ce programme de recherche répond aux demandes de mise en œuvre de ce système par les utilisateurs.

Statistique Canada garantit l'anonymat de tout fichier de microdonnées avec le concours de son Comité de diffusion des microdonnées. Ce comité reçoit à son tour le soutien de la Direction de la méthodologie et de son équipe d'experts en confidentialité, lesquels veillent à ce que les bonnes stratégies de constatation et d'atténuation des risques soient en place pour toute diffusion publique de données. En 2019-2020, une vingtaine de fichiers de données publiques ont été diffusés dans le cadre de ce processus.

Pendant cette période, diverses consultations ont eu lieu avec les partenaires nationaux et internationaux. Les pays étrangers consultés sont notamment le Danemark, le Royaume-Uni, l'Australie et les Antilles. Statistique Canada joue également un rôle actif auprès de la Commission économique des Nations Unies pour l'Europe et du Groupe de haut niveau sur la modernisation des statistiques officielles. Un document soulignant les défis en matière d'accès et de confidentialité à Statistique Canada a été présenté à la session de travail d'octobre 2019 sur la confidentialité des données statistiques.

Au nombre des organismes nationaux consultés, on compte le Conseil canadien des registres du cancer, l'Agence de la santé publique du Canada, Santé Canada, l'International Association of Privacy Professionals (IAPP) Canada, la Société canadienne d'hypothèques et de logement, CANON, BC Stats, le Conseil des arts du Canada et le Secrétariat des programmes interorganismes à l'intention des établissements, pour ne citer que ceux-là. Il y a eu présentation des stratégies d'amélioration de l'accès à l'occasion des Joint Statistical Meetings de 2019.

Pour obtenir plus de renseignements, communiquez avec :
Steven Thomas (613-882-0825, steven.thomas@canada.ca).

3. Théorie et cadre

3.1 Théorie et cadre – Intégration des données

Intégration des données : Combiner les données d'échantillons probabilistes et non probabilistes

L'application de méthodes de collecte ou d'acquisition de données ne reposant pas sur un plan de sondage probabiliste a récemment augmenté. Notons, par exemple, un recours croissant à l'« externalisation ouverte » depuis le début de la pandémie COVID-19. Un échantillon participatif peut se définir comme tout échantillon de bénévoles généralement constitué par Internet. Ces échantillons sont parfois d'assez grande taille, mais peuvent donner des estimations entachées d'importants biais de sélection. Le moyen d'obtenir des estimations significatives et des inférences valides à partir de ces grands échantillons non probabilistes est une question importante qui nécessite encore des recherches et des expériences. Les méthodes par lesquelles on tente de répondre à la question se trouvent souvent à combiner les données d'échantillons non probabilistes et probabilistes. C'est ce qu'on appelle l'intégration des données statistiques.

Le présent projet vise principalement trois objectifs :

- évaluer la possibilité d'utiliser les méthodes actuelles d'intégration des données pour obtenir des estimations avec moins de biais introduits par les échantillons non probabilistes;
- concevoir de nouvelles méthodes ou en adapter pour résoudre les problèmes qui se posent dans la pratique;
- concevoir et expérimenter des prototypes qui incarnent les méthodes les plus prometteuses.

SOUS-PROJET : Utilisation d'arbres de classification pour pondérer un échantillon non probabiliste

Une méthode d'intégration des données qui serait propre à réduire le biais consiste à modéliser la probabilité de participation à un échantillon non probabiliste et à pondérer chaque participant par l'inverse de sa probabilité estimée de sélection. Chen, Li et Wu (2019) ont employé une fonction logistique pour modéliser la probabilité de participation, des variables explicatives étant posées. Un cas particulier est le modèle des groupes homogènes où la probabilité de participation est supposée uniforme dans chaque groupe. Nous nous appuyons sur ce travail pour étendre l'idée aux arbres de classification (Breiman, Friedman, Stone et

Olshen, 1984). Dans cette perspective, un arbre de classification est là pour créer un ensemble de groupes exhaustifs et mutuellement exclusifs qui sont individuellement homogènes sur le plan des probabilités de participation, puis pour appliquer la méthode de Chen, Li et Wu (2019) à l'estimation des probabilités en question. On peut voir dans les arbres de classification une méthode non paramétrique pour trouver des variables explicatives d'intérêt avec leurs interactions.

Progrès :

Dans la dernière année, nous avons conçu un programme en R qui se sert d'un algorithme pour construire un arbre entier. Nous avons évalué cet algorithme par une étude en simulation. Les premiers résultats indiquent qu'il est prometteur s'il s'agit de diminuer le biais des estimations obtenues par voie d'échantillon non probabiliste. Les résultats ont été présentés à la réunion de 2019 de la Société statistique du Canada (Chu et Beaumont, 2019).

Nos résultats disaient certes que ces arbres étaient de nature à réduire le biais de sélection, mais ils montraient en même temps que les estimations ainsi obtenues étaient quelque peu inefficaces. Cela peut s'expliquer par un excès d'ajustement, c'est-à-dire par la création de trop de groupes. L'élagage est habituellement recommandé en cas de surajustement. Cela se fait d'ordinaire en deux étapes. D'abord, on établit une suite de sous-arbres en ordre décroissant à partir de l'arbre entier. Ensuite, on choisit le meilleur de ces sous-arbres, normalement par validation croisée. Cette validation ne paraît pas facile à appliquer dans le contexte de l'intégration des données. Nous nous sommes efforcés de trouver un autre mode de sélection de sous-arbres par le critère d'information d'Akaike. Cette application a été décrite dans un document présenté au Comité consultatif des méthodes statistiques en juin 2020 (Beaumont et Chu, 2020).

SOUS-PROJET : Mode bayésien d'estimation d'enquête avec des échantillons probabilistes et non probabilistes

Une méthode bayésienne d'intégration des données a récemment été présentée dans le Journal of Official Statistics (Sakshaug, Wisniowski, Ruiz et Blom, 2019). L'article traite de l'estimation des paramètres de modèle lorsque la variable dépendante *y* et le vecteur de variables explicatives sont observés à la fois dans un échantillon probabiliste et un échantillon non probabiliste. L'échantillon non probabiliste est un moyen d'obtenir une distribution antérieure dans l'hypothèse d'un plan d'échantillonnage aléatoire simple. Le but dans ce projet est d'étendre la méthode à l'estimation de paramètres de population finie dans des plans de sondage complexes et de voir s'il est possible d'obtenir par modèle des estimations plus efficaces que les estimations types à pondération d'enquête.

Progrès :

Nous avons lu et étudié l'étude de Sakshaug et coll. (2019). Un rapport sommaire a été produit (You, 2019). Nous prévoyons élargir le travail de Sakshaug et coll. (2019) pour concevoir une approche d'inférence bayésienne fondée sur un modèle de superpopulation pour estimer les moyennes et les totaux de population avec un plan d'échantillonnage aléatoire simple et des plans de sondage plus complexes, ainsi que pour emprunter de l'information à des données d'échantillon non probabiliste pour obtenir une distribution antérieure dans le cadre bayésien.

SOUS-PROJET : Intégration de données statistiques dans un mode de prédiction

Nous étudions le problème où est disponible un échantillon non probabiliste contenant un vecteur de variables auxiliaires pour chaque unité échantillonnée. Nous posons que cet échantillon non probabiliste rend compte d'une partie significative de la population. Un échantillon probabiliste pourrait aussi être disponible où il y aurait aussi la variable d'intérêt *y* pour chaque unité de l'échantillon. L'indicateur de participation à l'échantillon non probabiliste se présente ici dans l'échantillon probabiliste. Ce scénario vaut pour une enquête sur le trafic postal menée par La Poste en France. Alain Dessertaine a proposé un prédicteur dans ce scénario. Le but du projet est d'étudier les propriétés de ce prédicteur.

Progrès :

Nous avons conçu des estimateurs de variance, dont un estimateur bootstrap, pour évaluer la qualité du prédicteur proposé. Les détails figurent dans un rapport provisoire produit à l'interne. Nous entendons maintenir la collaboration avec La Poste, l'École d'économie de Toulouse et l'Université de Besançon et transformer cette ébauche en un article commun pouvant être publié dans un journal contrôlé par les pairs.

SOUS-PROJET : Examen des méthodes d'intégration des données

Le but de ce projet est de dresser un vaste bilan documentaire traitant des méthodes d'intégration des données statistiques. Nous avons passé en revue des méthodes par plan de sondage comme les méthodes à bases multiples et la méthode de calage par le plan d'échantillonnage, tout comme des méthodes par modèle (appariement statistique, calage dépendant du modèle, pondération par l'inverse du score de propension et estimation sur petits domaines).

Progrès :

Un document a déjà été rédigé et présenté à Techniques d'enquête. Nous avons reçu les commentaires des examinateurs cette année et révisé le document en conséquence. Celui-ci

sera publié dans le numéro de juin 2020 de Techniques d'enquête (Beaumont, 2020a). La partie des méthodes par modèle a également été présentée à une conférence sur la méthodologie d'enquête qui a eu lieu en juin 2019 à Florence, en Italie.

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

Bibliographie

Breiman, L., Friedman, J.H., Stone, C.J. et Olshen, R.A. (1984). Classification and Regression Trees. CRC Press.

Chen, Y., Li, P. et Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (publié en ligne).

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

3.2 Théorie et cadre – Qualité

Le Secrétariat de la qualité dont les activités de support sont décrites dans la section 5.4 est également impliqué dans des activités de développement.

SOUS-PROJET : Mesure et communication de la qualité pour les programmes statistiques utilisant des données intégrées

Statistique Canada, comme bien d'autres agences statistiques nationales, fait face à un changement de paradigme dans la production de ses statistiques officielles. Ce changement se traduit par une transition entre un modèle traditionnel de statistiques produites à partir de méthodes développées pour des plans d'enquête à échantillon probabiliste et de collecte directe auprès des répondants à un nouveau modèle où les données sont extraites de différentes sources secondaires, telles que des données administratives ou des données satellites. Ces dernières sont ensuite combinées ou intégrées afin de produire les statistiques recherchées. Dans ce nouveau contexte, être capable de bien mesurer et communiquer la qualité des données devient un défi important, considérant que les méthodes actuelles et les termes utilisés sont étroitement liés à la théorie de l'échantillonnage.

Progrès :

Un état des lieux et un plan de travail ont été développés sur la base de l'article de Rancourt (2018) et ont fait l'objet de présentations au Comité d'examen scientifique et à la conférence International Total Survey Error Workshop (Beaulieu, Chepita, Fortier, 2019). Un groupe de travail a été formé pour faire une revue des indicateurs de qualité existants et identifier les lacunes pour les différentes étapes de la production de statistiques officielles. Le travail de développement d'une étiquette sur le format des étiquettes nutritionnelles a été amorcé et discuté avec certains programmes.

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

Bibliographie

Rancourt, E. (2018). L'administration d'abord comme paradigme statistique pour les statistiques officielles canadiennes : Signification, défis et possibilités. Présenté au Symposium international de 2018 sur les questions de méthodologie.

3.3 Cadre – Apprentissage automatique responsable

Statistique Canada utilise des techniques d'apprentissage automatique pour résoudre des problèmes de données à grande échelle. Certains des projets de recherche dans ce domaine sont décrits dans la section 5.6. Parallèlement, les instituts nationaux de statistique sont confrontés à une pression sans précédent pour démontrer aux citoyens, aux entreprises et aux utilisateurs qu'ils sont des institutions dignes de confiance et transparentes. Par conséquent, une approche responsable consiste à développer un cadre adapté à l'utilisation des techniques d'apprentissage automatique, y compris des lignes directrices pour la construction et la mise en œuvre de processus éthiquement et méthodologiquement solides, et des mesures pour évaluer et communiquer la qualité des processus d'apprentissage automatique et de leurs produits.

Progrès :

L'élaboration du cadre et des lignes directrices a été effectuée par une équipe multidisciplinaire et les versions provisoires ont été approuvées par diverses communautés internes (science des données, méthodologie, haute direction (comité de la qualité, des méthodes et des normes)) et des partenaires externes (comme Postes Canada et le Comité consultatif sur les méthodes statistiques). Le cadre est aligné sur la Directive sur la prise de décision automatisée et son outil d'analyse d'impact algorithmique, élaborés par le Secrétariat du Conseil du Trésor (2020). Il s'articule autour de 4 thèmes: Respect des personnes, Respect des données, Application rigoureuse et Méthodes rigoureuses et un certain nombre d'attributs pour chaque thème. Les lignes directrices n'ont pas encore été officiellement adoptées, mais sont actuellement testées par les gestionnaires des applications de la science des données. Une liste de contrôle accompagnant les directives est également en cours de test et est utilisée pour aider à évaluer les processus d'apprentissage automatique responsables. L'adoption finale des lignes directrices et de la liste de contrôle, l'élaboration d'un tableau de bord pour le suivi et la mise en œuvre d'un processus d'examen formel sont encore à faire.

Pour obtenir plus de renseignements, communiquez avec :
Yanick Beaucage (613-854-2397, yanick.beaucage@canada.ca).

Bibliographie

Secrétariat du Conseil du Trésor (2020). Directive sur la prise de décision automatisée, site Web du gouvernement du Canada https://www.tbs-sct.gc.ca/pol/doc-fra.aspx?id=32592.

3.4 Cadre – Nécessité et proportionnalité

À l'ère de l'information, les sources de données et les besoins d'information ne cessent d'augmenter et d'évoluer. Pour répondre à ces demandes croissantes et au besoin tout aussi important de protéger les données des répondants, Statistique Canada travaille en consultation avec des experts en statistiques et en protection de la vie privée du monde entier pour élaborer un nouveau cadre méthodologique fondé sur les principes de nécessité et de proportionnalité.

Progrès :

En octobre 2019, Statistique Canada a adopté un nouveau cadre de nécessité et de proportionnalité afin de maximiser conjointement la production de renseignements et la protection de la vie privée lors de l'élaboration d'approches de collecte de données. Ce cadre fournit à la fois une justification et un guide pour la conception de stratégies de collecte de données sensibles à l'aide d'enquêtes, de sources administratives obtenues du secteur public ou privé ou de toute autre méthode. L'approche est le résultat de consultations menées auprès d'experts et de praticiens nationaux et internationaux dans le domaine des statistiques officielles, des associations professionnelles de statistiques, des experts en protection de la vie privée, des experts en éthique et le Bureau du commissaire à la protection de la vie privée. Le cadre de nécessité et de proportionnalité est une adaptation de l'approche scientifique (Rancourt, 2019) au contexte à la fois de la méthodologie statistique et de la protection de la vie privée. Il repose sur une description solide des raisons pour lesquelles une source de données donnée est nécessaire et une évaluation éthique approfondie. Pour soutenir ces travaux, un secrétariat fonctionnant sous la direction d'un chef de l'éthique et de l'intégrité scientifique a été mis en place. Les travaux du secrétariat appuient les activités d'un comité interne d'éthique des données. Statistique Canada a intégré le Cadre de nécessité et de proportionnalité dans son processus d'acquisition de données de sorte que l'ingestion de toute nouvelle source de données doit suivre le cadre. Pour plus d'informations, visitez Le Centre de confiance de Statistique Canada (https://www.statcan.gc.ca/fra/confiance).

Pour obtenir plus de renseignements, communiquez avec :
Eric Rancourt (613-298-9403, eric.rancourt@canada.ca).

4. Soutien (Centre de ressources)

4.1 Centre de ressources en couplage d'enregistrements

Le Centre de ressources en couplage d'enregistrements (CRCE) a pour but de fournir des services de consultation aux utilisateurs tant internes qu'externes des méthodes de couplage, ainsi que de recommander des logiciels, des méthodes et des travaux en collaboration sur les applications de couplage à des fins d'évaluation de méthodes nouvelles ou améliorées dans ce domaine. Nous évaluons les progiciels et paquets de couplage d'enregistrements et, au besoin, nous élaborons des prototypes de logiciels incorporant des méthodes indisponibles dans les progiciels et paquets existants, tout en aidant à diffuser de l'information sur les méthodes, les logiciels et les applications de couplage pour les intéressés à l'intérieur et à l'extérieur de Statistique Canada.

Progrès :

L'équipe de support a aidé l'équipe de développement et suivi les entrées des utilisateurs pour aider à identifier les idées d'améliorations potentielles. Le CRCE a également fourni un soutien aux utilisateurs internes et externes de G-Link lorsque de l'aide/des commentaires/des suggestions concernant G-Link étaient recherchés sur G-Link info via des tickets JIRA.

Au cours de l'année, une grande partie du travail en méthodologie s'est articulée autour de l'élaboration d'une nouvelle version de G-Link (version 3.5) avec couplage par profil (comme pour le couplage déterministe), repérage et traitement des enregistrements orphelins et pseudoclés intégrées. Se sont ajouté des travaux visant à intégrer un outil d'examen de bureau (évaluation de la qualité) et à corriger et améliorer certains estimateurs de seuil.

Le CRCE a travaillé à divers autres projets de couplage d'enregistrements au cours de l'année et a notamment tenu deux nouvelles réunions du Forum sur le couplage d'enregistrements. Nos couplages nous ont aidés à documenter le rendement et les problèmes de gestion et de développement. Ils ont été l'occasion de mettre à l'essai sur le terrain de nouvelles caractéristiques de la version 3.5 et de concevoir des méthodes plus systématiques et théoriquement plus cohérentes de définition et d'adaptation des couplages dans les serveurs et le réseau SAS. Le CRCE a mis à jour le tutoriel de la version 3.5 de G-Link.

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

4.2 Systèmes généralisés

Recherche-développement – Systèmes généralisés

L'équipe des systèmes généralisés (GenSys) est chargée de la recherche, du développement et du soutien des systèmes suivants :

- G-Est : Système d'estimation généralisé;
- G-Sam : Système d'échantillonnage généralisé;
- BANFF : Système de vérification et d'imputation généralisé.

En dehors du soutien et de la formation aux systèmes généralisés, l'équipe fait de la recherche-développement sur la visualisation des données, l'estimation de variance et d'autres méthodes liées aux processus d'enquête.

SOUS-PROJET : Soutien et développement continus des systèmes généralisés

L'équipe des systèmes généralisés facilite l'utilisation des systèmes aux fins des enquêtes nouvelles ou existantes et des programmes statistiques en cours de restructuration.

Progrès :

L'équipe de soutien des systèmes généralisés a appuyé en permanence les utilisateurs réels ou éventuels (tant à Statistique Canada que dans d'autres organismes), mis à jour et présenté le contenu de la formation à diverses tribunes et rencontré les délégués internationaux pour discuter de l'état actuel et de l'évolution future des systèmes en question. L'équipe du G-Est a commencé à travailler à une nouvelle version du SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) pour résoudre les problèmes que font peser sur le système les grands ensembles de données aux stratégies complexes d'imputation. Ce développement se fait en consultation avec les clients et les intervenants. Le travail en est maintenant à l'étape du prototypage et des essais sont en cours auprès des utilisateurs. Dans la planification stratégique des systèmes généralisés, il faut passer régulièrement les possibilités en revue. Cet examen ne porte pas sur les problèmes techniques exigeant un développement d'ordre méthodologique, mais plutôt sur les problèmes des plateformes de traitement et l'utilisation efficace d'outils existants comme les langages de source ouverte. Dans le cadre de cet examen, nous avons entrepris une analyse des outils disponibles dans d'autres logiciels pour en reconnaître les fonctions et les comparer à celles de nos systèmes généralisés. De plus, nous avons conçu un modèle pour d'éventuels perfectionnements des systèmes généralisés dans un cadre d'analyse de rentabilisation.

SOUS-PROJET : Développement des systèmes généralisés – Étude de méthodes nouvelles et d'autres techniques de validation

Progrès :

L'équipe des systèmes généralisés a poursuivi le développement d'un outil d'évaluation fondé sur des simulations de Monte Carlo de la non-réponse et a étudié des outils de visualisation pour compléter les estimations empiriques actuellement produites. Ce travail s'est fait conjointement avec Keren Li, qui fait des études supérieures à l'Université d'Ottawa et qui a rédigé un rapport de stage de travail (Li, 2019) traitant en détail des outils de visualisation et de la façon de les produire avec l'outil Impact. Ce travail a également été présenté aux Joint Statistical Meetings de 2019 de l'American Statistical Association (Gray, 2019). L'outil a en outre permis d'évaluer la stratégie d'imputation d'une enquête courante dans une validation de principe. Des recherches préliminaires ont enfin porté sur l'élargissement des méthodes d'imputation par le plus proche voisin (système de BANFF); cet exercice comportait un bilan documentaire et des tests de mesure des distances en imputation par le plus proche voisin.

SOUS-PROJET : Développement des systèmes généralisés – Cadre de vérification des données statistiques et soutien des initiatives de la Commission économique des Nations Unies pour l'Europe

Progrès :

L'équipe a continué à faire partie du groupe de travail qui, à la Commission économique des Nations Unies pour l'Europe, a pour tâche d'élaborer un cadre de vérification. Elle a notamment aidé à mettre la dernière main au rapport officiel et à la documentation du Modèle générique de vérification des données statistiques (GSDEM) diffusé en juin 2019. Le groupe de travail a donné sa forme définitive au rapport pendant cette période avec des apports importants de l'équipe (CEE-ONU, 2019). Un membre de l'équipe de BANFF a siégé au comité d'organisation d'un atelier consacré à la vérification des données statistiques et prévu pour 2020 dans le cadre de la Commission économique pour l'Europe. Jusqu'ici, ce travail a consisté à organiser la thématique, à établir les communications et à examiner les résumés analytiques.

SOUS-PROJET : Développement des systèmes généralisés et mise à jour des versions

Progrès :

Une nouvelle version de G-Sam a été élaborée, essayée et diffusée. Cette version 1.03.001 a été rendue publique en mars 2020 (Statistique Canada, 2020). Elle comprenait un certain nombre de correctifs non essentiels et une mise à jour du modèle d'attribution visant à

éliminer les redondances et à simplifier les apports des utilisateurs. Plus précisément, un changement a permis d'introduire des valeurs auxiliaires de zéro dans le problème d'optimisation de la répartition, ce qui a exigé l'intégration de nouveaux outils de diagnostic. Une nouvelle version de G-EST a été élaborée, essayée et diffusée en réaction à une erreur qui s'était glissée dans la création de poids d'itération bootstrap dans des cas bien précis d'estimation de variance. La version 2.03.001 de G-EST a été diffusée en novembre 2019 (Statistique Canada, 2019a). Elle venait rectifier une erreur qui influait sur la création d'une pondération par itération bootstrap. L'équipe du G-Est a travaillé rapidement pour cerner et corriger l'erreur et communiquer avec les clients et les intervenants tout au long du processus.

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

4.3 Centre de ressources en conception de questionnaire

Le Centre de ressources en conception de questionnaires (CRCQ) de la Direction de la méthodologie est le centre d'expertise de Statistique Canada en matière de conception et d'évaluation de questionnaires. Il fournit des services de consultation et de soutien et mène des projets et des recherches sur l'élaboration, l'essai et l'évaluation des questionnaires d'enquête. Il joue un très grand rôle en gestion de la qualité et répond aux besoins des programmes dans tout Statistique Canada en consultant les clients, les enquêtés et les utilisateurs des données et en se chargeant des essais préalables de questionnaires d'enquête.

Le gros de son travail se fait en recouvrement des frais, mais la section est fréquemment sollicitée à titre spécial en vue de fournir des évaluations d'expert et des services-conseils dans une grande diversité d'enquêtes. Elle donne aussi des cours sur la conception des questionnaires.

Progrès :

Le CRCQ a contribué à une initiative de l'organisme consistant à examiner les perceptions qu'ont les citoyens canadiens de l'utilisation secondaire de données administratives comme source d'information statistique. Des forums de discussion et des consultations ont eu lieu.

Le CRCQ a en outre commencé à élaborer un plan de collecte de renseignements pouvant alimenter une échelle de sensibilité liée aux sujets et aux sources de données d'enquête.

Le groupe a enfin pris part à diverses initiatives de consultation des entreprises.

Pour obtenir plus de renseignements, communiquez avec :
Paul Kelly (613-371-1489, paul.kelly2@canada.ca).

4.4 Secrétariat de la qualité

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

SOUS-PROJET : Mise à jour des Lignes directrices concernant la qualité

Cette initiative, qui a pour but d'actualiser les Lignes directrices concernant la qualité, vise trois objectifs : a) fournir un document de référence pertinent à tous les autres producteurs de données du système statistique canadien; b) s'adapter à la nouvelle réalité des données administratives en traitant des principaux processus de production statistique; c) concourir au respect des méthodes actuelles d'assurance de la qualité.

Progrès :

Certaines révisions et ajouts ont été apportés à l'ébauche rédigée au cours de l'année 2018-19. En particulier, des principes directeurs et pratiques exemplaires qui touchent les données alternatives ainsi qu'une plus grande emphase sur les pratiques éthiques, la protection de la vie privée et la proportionnalité ont été ajoutés. La nouvelle édition des Lignes directrices concernant la qualité a été publiée en décembre 2019 (Statistique Canada, 2019b).

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

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

Progrès :

Le Secrétariat de la qualité a pris des mesures de renforcement de capacités à l'intention de nombreux partenaires au cours de la période visée. À l'interne, des ateliers de formation ont été offerts par l'entremise de différents cours offerts au personnel ainsi que par des formations plus ciblées pour des équipes travaillant sur un programme particulier. Au niveau des partenaires nationaux, des présentations formelles sur les pratiques de gestion de la qualité

ont été faites à six organisations, en plus de tenir de nombreuses discussions dans le cadre du « Data Governance Standardization Collaborative » ainsi que groupe de travail sur la qualité des données. Ce dernier groupe, co-présidé par Statistique Canada, vise à définir un cadre de la qualité des données applicable à tous les organismes du gouvernement du Canada, dans le cadre de la mise en œuvre de la stratégie des données. Une participation à un panel sur la qualité des données (Beaulieu, 2020) a également mené à des consultations avec d'autres partenaires. La validation de la qualité d'un processus statistique effectué par un autre organisme fédéral a également été complétée. Au niveau international, des présentations et discussions ont eu lieu dans le cadre de visites formelles de délégation d'autres pays. De la consultation a également été offerte à un pays rédigeant son propre cadre d'assurance de la qualité et l'implication comme Groupe d'experts des Nations Unies sur les cadres nationaux d'assurance de la qualité s'est poursuivie en vue de compléter la rédaction du « United Nations National Quality Assurance Framework Manual for Official Statistics » (United Nations, 2019).

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

Bibliographie

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

4.5 Centre de ressources en analyse de données

Le Centre de ressources en analyse de données (CRAD) 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. Ses services – qui portent surtout sur les données d'enquête et de recensement ou les données administratives – sont mis à la disposition des employés de l'organisme ou d'autres, ainsi que des analystes et des chercheurs des milieux universitaires ou des centres de données de recherche (CDR).

Progrès :

Consultations

Dans le cadre de son mandat, le CRAD fournit des services de consultation à la demande de divers clients. Les demandes émanent en majeure partie des analystes et des méthodologistes de Statistique Canada. Les consultations portent sur un éventail de sujets : pondération bootstrap d'enquête, construction d'intervalles de confiance, vérification d'hypothèses, analyse sur données couplées, estimation de modèles linéaires hiérarchiques avec données d'enquête, etc. Nous avons également aidé nos clients à appliquer les méthodes des logiciels SUDAAN, SAS, STATA et R. Des consultations externes ont eu lieu auprès de différents clients d'autres ministères et organismes fédéraux ou provinciaux et des chercheurs universitaires. Les demandes touchaient à divers thèmes liés à l'analyse des données d'enquête : vérification des différences entre médianes, ajustement de régression logistique, repondération de la non-réponse, degrés de liberté dans l'estimation de variance, etc. Nous avons également exercé un contrôle méthodologique sur une étude de comparaison des taux de promotion entre groupes. Enfin, l'équipe a donné des conseils d'expert aux analystes et aux chercheurs des centres de données de recherche (CDR). Les sujets abordés étaient notamment les suivants : combinaison de cycles d'enquête, utilisation de poids bootstrap, modèles multiniveaux, régression quantile, analyse d'ensembles de données couplées.

Services et matériel de formation

L'équipe a continué à exposer les thèmes d'analyse de données statistiques dans le cadre de l'atelier d'interprétation des données.

À l'occasion de la Conférence annuelle des analystes des CDR, nous avons présenté un exposé sur l'analyse avec données couplées.

Nous avons rédigé sur les pratiques exemplaires en visualisation des données un document qui se veut un outil à la disposition des analystes et des équipes de diffusion de Statistique Canada.

Nous avons produit deux fiches d'information pour répondre à des questions fréquemment posées :

- Pourquoi l'ajustement de Fay est-il nécessaire lorsqu'on utilise des poids bootstrap moyens ?
- Comment la pondération bootstrap d'enquête se compare-t-elle à la méthode bootstrap proprement dite ?

Collaboration avec les analystes

Deux articles ont paru dans le numéro de septembre 2019 de Rapports sur la santé. (https://www150.statcan.gc.ca/n1/pub/82-003-x/82-003-x2019009-fra.htm).

Pour obtenir plus de renseignements, communiquez avec :
Harold Mantel (613-863-9135, harold.mantel@canada.ca).

4.6 Centre de recherche et analyse en séries chronologiques

La recherche sur les séries chronologiques vise à maintenir un degré élevé d'expertise et à offrir les services de consultation nécessaires dans ce domaine, à concevoir et mettre à jour des outils de solution des problèmes que posent les séries chronologiques dans la vie réelle et à étudier les problèmes de l'heure sans solutions connues ou acceptables. Certaines des activités de recherche sur les séries chronologiques sont présentées à propos de l'estimation en temps réel à la section 1.2.

SOUS-PROJET : Désaisonnalisation et estimation de la tendance-cycle

Ce sous-projet consiste à évaluer les méthodes d'application de la désaisonnalisation et de l'estimation de tendance-cycle. Ces méthodes sont là pour appuyer l'analyse des données de séries chronologiques dans le cadre des statistiques officielles. Cette recherche permet au Centre de recherche et d'analyse en séries chronologiques de maintenir une expertise de pointe dans ce domaine.

Progrès :

Nous avons préparé et produit un exposé et un compte rendu comparant les méthodes de désaisonnalisation X-12-ARIMA et SEATS. Nous avons trouvé à des fins de comparaison des modèles analogues d'espace d'états, lesquels représentent un pas en avant dans la conception d'une désaisonnalisation par espace d'états (Matthews et Dochitoiu, 2019). À la suite de ces travaux, nous collaborons avec l'Université d'Ottawa à l'élaboration d'un cadre méthodologique d'application d'une désaisonnalisation aux propriétés souhaitables au moyen de ces modèles.

Le travail s'est poursuivi sur l'estimation de variance pour les estimations désaisonnalisées. Précisons qu'un exposé et un compte rendu ont été préparés et produits à l'occasion des Joint Statistical Meetings de 2019 (Verret et Dochitoiu, 2019). L'approche retenue consistait à appliquer une méthode bootstrap coordonnée à l'estimation de la variance de plan de sondage dans une enquête-entreprise. Plusieurs problèmes subsistent en ce qui concerne les composantes de la variance, mais ces travaux représentent un grand pas en avant dans l'établissement de méthodes d'estimation de la variance de désaisonnalisation aux fins des enquêtes comportant un plan type de sondage d'enquête auprès des entreprises.

Nous avons employé un tableau de bord de la désaisonnalisation dans les enquêtes économiques mensuelles aux fins de l'analyse et de l'interprétation des résultats désaisonnalisés par les méthodologistes et les spécialistes en la matière. Programmé dans

R-Shiny, cet outil automatisé produit un résumé interactif des résultats désaisonnalisés d'une série individuellement traitée selon une grille d'analyse. L'outil en question a été présenté à l'atelier des praticiens de la désaisonnalisation parrainé par le Census Bureau des États-Unis (Matthews, Ferland, Verret et Habli, 2019). Il a été éprouvé dans plusieurs enquêtes sur un certain nombre de mois et s'est révélé efficace pour qui veut analyser et comprendre ce que sont des résultats désaisonnalisés. Aucun nouveau travail n'est prévu sur ce tableau de bord dans le cadre des budgets de recherche, car l'outil ne devrait nécessiter que de légères adaptations pour convenir à un projet donné.

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

Ce sous-projet consiste à appuyer et à développer les outils d'application des méthodes des séries chronologiques, et notamment de désaisonnalisation. Le développement s'est poursuivi sur le système de traitement des séries chronologiques servant actuellement aux applications de production dans tout l'organisme et un soutien sur demande a été offert à chaque programme.

Progrès :

Il convient de noter que la version 3.08 du Système de traitement des séries chronologiques (STSC) (voir Ferland, 2019) a été rendue publique comme nouvelle offre enrichie de techniques disponibles dans le cadre de ce système. Les nouvelles caractéristiques sont notamment des outils sur mesure pour la production du tableau de bord de la désaisonnalisation, des capacités accrues d'étalonnage et des options d'établissement de prévisions et en particulier de prévisions immédiates. La nouvelle fonctionnalité d'étalonnage consiste en un nouveau module d'interpolation ajouté qui substitue aux points imbriqués de données manquantes des interpolations spline cubiques (en lissage linéaire) ou des interpolations linéaires (en segmentation de droite). Elle consiste aussi en plusieurs options de projection de ces mêmes données manquantes en début (avant la première valeur) et en fin (après la dernière valeur) de série. Ce module servira tout particulièrement à étalonner les variables des stocks (inventaires, soldes, etc.). Les options de prévision créées permettent de produire des prévisions et des intervalles prévisionnels à l'aide des modèles regARIMA, ce qui représente un important premier pas en avant dans la préparation aux méthodes de base de prévision immédiate.

SOUS-PROJET : Consultation et formation générales en séries chronologiques

Dans le cadre de son mandat, le Centre de recherche et d'analyse en séries chronologiques (CRASC) tient des consultations à la demande de divers clients à Statistique Canada. Les sujets les plus fréquemment traités sont ceux de la détermination des ruptures de série, de

l'application de la désaisonnalisation et de la modélisation de séries chronologiques dans diverses situations et des applications particulières d'étalonnage et de rapprochement.

Progrès :

De plus, des échanges officiels et officieux ont lieu au besoin avec d'autres organismes statistiques (Census Bureau et Bureau of Labor Statistics aux États-Unis, Eurostat, Statistique Norvège, Australian Bureau of Statistics, etc.) et des organismes universitaires (universités de Waterloo et d'Ottawa) dans le cadre d'une collaboration et d'une consultation sur des sujets d'actualité.

Plusieurs articles de revue ont été examinés au sujet de l'application de la désaisonnalisation. Un de ces articles traitait d'une technique appliquée d'étalonnage visant à préserver les mouvements dans des séries dérivées lorsque l'étalonnage se fait en plusieurs étapes.

Les cours en place ont été mis à jour et donnés au besoin à des participants tant de l'intérieur que de l'extérieur de Statistique Canada. Le cours récemment mis à jour sur la modélisation et la prévision en séries chronologiques a été présenté à nouveau et un volet sur l'apprentissage actif a été ajouté au cours sur la désaisonnalisation avec X-12-ARIMA, le but étant de permettre aux participants d'apporter des données à désaisonnaliser pour acquérir de l'expérience dans l'application pratique de la méthode en question. Des cours ont également eu lieu sur l'étalonnage, la méthode itérative du quotient et les composantes des séries chronologiques.

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

4.7 Communauté de pratique de la science des données

La Communauté de pratique d'apprentissage automatique de Statistique Canada a pour but de faciliter la collaboration et le transfert de connaissances ainsi que l'amélioration de nos opérations en apprentissage automatique à Statistique Canada.

SOUS-PROJET : Développement de la capacité en apprentissage automatique

À travers différentes activités reliées à l'apprentissage automatique réunissant de 30 à 50 personnes telles que des dîners-causeries, présentations, groupes de lectures, groupes de visionnement et le partage d'information sur un site développé et mis-à-jour par les membres, la Communauté, par sa présence active, continue de collaborer au développement des capacités en apprentissage automatique des employés de Statistique Canada.

Progrès :

La Communauté a organisé plusieurs présentations touchant à plusieurs sphères de l'apprentissage automatique comme par exemple, une introduction à l'architecture informatique, la classification automatique de marchandises, l'utilisation d'un paquet R pour créer des données synthétiques, une introduction à GitHub et à sa version locale GitLab, la modélisation de contenu par sujet ou encore la présentation d'un nouveau Cadre pour l'utilisation responsable des processus en apprentissage automatique. Elle a tenu deux groupes de lecture sur des articles touchant l'apprentissage automatique pour les Organismes nationaux de statistique ou encore l'utilisation de l'apprentissage automatique pour tenir compte de la non-réponse. La Communauté a également créé un groupe de visionnement de cours d'apprentissage automatique en ligne gratuits, permettant aux participants d'échanger sur les sujets couverts suite au visionnement. La Communauté a développé une liste des projets existants en méthodologie qui explore ou utilise l'apprentissage automatique et a participé à un exercice de centralisation des présentations de toutes les communautés reliées aux sciences des données à Statistique Canada.

Pour obtenir plus de renseignements, communiquez avec :
Yanick Beaucage (613-854-2397, yannick.beaucage@canada.ca).

5. Recherche divisionnaire et autres activités

5.1 Division des méthodes de la statistique économique

SOUS-PROJET : Évaluation des estimateurs de régression d'enquête assistés par modèle à l'aide de Lasso et d'arbres de régression

Dans ce projet, nous évaluons au moyen d'études en simulation les propriétés de différents estimateurs de régression d'enquête pour les totaux de population finie. Nous visons ici la méthode LASSO (McConville, Breidt, Lee et Moisen, 2017) et les arbres de régression (McConville et Toth, 2019). La méthode LASSO sert à choisir des variables explicatives appropriées pour l'estimateur de régression. Quant aux arbres de régression, ils permettent de choisir un ensemble de groupes appropriés et, dans ce cas, l'estimateur de régression obtenu se ramène à un estimateur de poststratification. Nous évaluons d'abord les méthodes dans le contexte d'un échantillon probabiliste et nous nous attachons ensuite à tout ce qui est échantillons non probabilistes.

Progrès :

Nous avons effectué une vaste étude en simulation. Les résultats en sont examinés et récapitulés dans Lundy et Rao (2019). Notre grande conclusion est que les arbres de régression donnent l'estimateur de régression le plus efficace avec une petite taille d'échantillon et un grand nombre de variables auxiliaires.

Pour obtenir plus de renseignements, communiquez avec :
Wesley Yung (613-951-4699, wesley.yung@canada.ca).

Bibliographie

McConville, K.S., Breidt, F.J., Lee, T.C.M. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5, 131-158.

McConville, K.S., et Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46, 389-413.

5.2 Division des méthodes de la statistique sociale

SOUS-PROJET : Réseaux neuronaux pour l'imputation de l'impôt sur le revenu

Le Fichier administratif principal du revenu des particuliers (FAPRP) est une base de données administrative de l'impôt des particuliers qui est tirée des feuillets T1 et autres feuillets d'impôt. Dans ce fichier, l'information partielle sur le revenu des déclarants non fiscaux est complétée par voie d'imputation. L'imputation est dans ce cas une démarche particulièrement complexe. Les données de l'impôt fédéral et provincial sur le revenu font l'objet d'une imputation par donneur en quatre cycles consécutifs. D'abord, le revenu imposable est imputé et sert à calculer les crédits d'impôt et à imputer l'impôt fédéral et provincial dans les cycles qui suivent. Le processus d'imputation fait appel à une gamme de variables auxiliaires comme les divers types de revenu et les caractéristiques démographiques pour la recherche de donneurs, ainsi qu'à des formules pour le calcul de l'impôt net et des crédits d'impôt au moyen de paramètres fiscaux d'apport. L'évolution du régime fiscal d'année en année doit être surveillée et prise en compte dans le processus d'imputation. Ce processus est onéreux et n'a pas toujours été actualisé comme il aurait dû l'être.

Vu les relations multiples et non linéaires entre les données d'entrée et les variables fiscales imputées, les réseaux neuronaux peuvent nous apporter une catégorie appropriée de modèles et constituer une bonne solution de rechange à l'imputation par donneur. Ils pourraient également rendre moins nécessaire une mise à jour annuelle des programmes d'imputation en fonction de l'évolution du calcul des crédits d'impôt. Ajoutons que le FAPRP est un fichier de taille comptant 30 millions d'enregistrements et qui nous donne beaucoup de données pour l'apprentissage. Le but du présent projet est d'évaluer l'utilisation des réseaux neuronaux pour l'imputation de l'impôt sur le revenu dans le FAPRP comme solution de rechange à l'imputation par donneur.

Progrès :

Avec les conseils de la Division de la Science des données, un plan a été proposé pour explorer XGBoost avant les réseaux de neurones. Les deux méthodes ont été étudiées plus avant pour savoir comment elles pourraient être utilisées dans le contexte de ce projet. Un premier modèle utilisant XGBoost a été développé, formé et testé en utilisant un sous-ensemble de données (300 000 enregistrements, soit ~10 % de l'ensemble de données complet). Une première tentative de réglage des paramètres a également été lancée mais n'a pas pu être achevée en raison de l'arrêt de travail en mars.

Dans nos futurs travaux, nous envisageons de comparer les résultats d'imputation par quatre méthodes (modèle de régression type, XGBoost, réseaux neuronaux et programme en place

d'imputation par donneur). Nous disposons des valeurs réelles des déclarations de revenus pour un sous-ensemble d'enregistrements en imputation, ce qui pourrait servir de point de comparaison pour faire voir quelle méthode donne les meilleurs résultats.

SOUS-PROJET : Modèle au niveau du domaine dans l'estimation sur petits domaines

La demande d'estimation sur petits domaines (EPD) par les utilisateurs des données de Statistique Canada a constamment augmenté ces dernières années. On a entrepris de mettre en place un système de production d'EPD au début des années 2000. Ce système, qui fait maintenant partie du logiciel G-EST, traite le modèle de base de Fay et Herriot (1979) au niveau du domaine avec de multiples options (différentes méthodes, par exemple, d'estimation des composantes de la variance, éventail de méthodes d'estimation, etc.). Le modèle de base au niveau du domaine pourrait ne pas donner de résultats satisfaisants si un simple modèle linéaire n'explique pas suffisamment la relation entre la variable d'intérêt et les covariables. Le modèle par morceaux au niveau du domaine est un modèle de base modifié où la variable auxiliaire propre au domaine reçoit une partition en intervalles et où un segment de droite distinct est ajusté à chaque intervalle. Ajoutons que le choix des covariables influe grandement sur la qualité des estimations et l'exactitude du modèle. En règle générale, les données auxiliaires doivent venir d'une source indépendante des estimations directes et elles doivent être disponibles à des niveaux géographiques appropriés. Dans la pratique, il se peut que le modèle soit ajusté par rapport à des covariables qui ne sont pas en corrélation étroite avec la variable d'intérêt ou qui posent des problèmes de couverture.

Ce projet de recherche a deux grands objectifs : 1) comparer le modèle par morceaux au niveau du domaine au modèle de base et évaluer l'incidence du traitement par morceaux sur ses paramètres et ses résidus; 2) examiner le modèle avec des covariables qui ne sont pas en corrélation étroite avec la variable dépendante ou qui posent des problèmes de couverture (données agrégées sur les dépenses, par exemple). Les modèles en question sont mis à l'essai sur des données de l'Enquête sur les voyages des visiteurs (EVV) et les données sur les paiements servent d'information auxiliaire dans ce cas.

Progrès :

Nous avons examiné l'incidence du modèle par morceaux par rapport au modèle de base au niveau du domaine. Les résultats de l'analyse avec l'EVV et les données des paiements comme information auxiliaire ont semblé indiquer que le modèle par morceaux améliore l'ajustement du modèle. En particulier, le modèle modifié laissait des résidus ayant un comportement légèrement meilleur pour la normalité, l'homoscédasticité et la linéarité. Il a également livré des estimations plus fiables avec des coefficients de variation moindres en moyenne que ceux du modèle de base. Nous avons éprouvé le modèle par morceaux non seulement avec des

données sur les paiements, mais aussi avec des données agrégées sur les dépenses comme information auxiliaire. La version actuelle avec données agrégées sur les dépenses pose d'importants problèmes de couverture et soulève des questions de qualité avec pour résultat un nombre significativement réduit de domaines disponibles pour la modélisation et une corrélation plus faible avec les données de l'EVV. Ces constatations sont exposées dans un document interne. Selon les priorités et les ressources, nous essaierons et évaluerons les modèles au niveau du domaine tant avec les données sur les paiements qu'avec les données agrégées sur les dépenses soit comme variables séparées soit comme variable auxiliaire en combinaison.

SOUS-PROJET : Utilisation du Réseau routier national pour calculer la distance

Le Réseau routier national (RRN) est une série d'ensembles de données sous la responsabilité de Statistique Canada et disponible au portail des données ouvertes du gouvernement (ouvert.canada.ca). Ces ensembles contiennent les coordonnées en longitude et latitude des points du RRN, tout comme de l'information sur les points reliés par une route. Ainsi, ils peuvent servir à calculer la distance par route entre deux points géographiques (à l'intérieur du pays).

L'Enquête canadienne sur les mesures de la santé (ECMS) se reporte aux distances entre les adresses échantillonnées et les cliniques médicales en santé physique situées à chaque endroit échantillonné dans des modèles de non-réponse. En fait, la distance a servi à modéliser la non-réponse en clinique dans trois des cinq cycles de l'ECMS et la variable étroitement apparentée du code postal a été employée dans les deux autres. Il reste que l'ECMS a fait uniquement intervenir la distance en ligne droite (à l'aide de PROC GEODIST), car les données sur les adresses sont statistiquement sensibles. L'équipe de l'ECMS a commencé à concevoir une solution à l'aide des fichiers RRN. Toutefois, cette solution n'a pas été mise à l'essai à grande échelle en vue d'en déterminer la qualité et n'est pas pour l'instant généralisable à d'autres projets où on pourrait vouloir aussi exploiter les fichiers du RRN. Le présent projet vise à évaluer si la distance par route est préférable à la distance à vol d'oiseau dans la modélisation de non-réponse pour l'ECMS, ainsi qu'à documenter l'état du travail qui se fait là-dessus à la Direction de la méthodologie et ailleurs à Statistique Canada.

Progrès :

Nous avons comparé les modèles de correction de non-réponse à l'aide des données du cycle 5 de l'ECMS par les distances à vol d'oiseau et par la route. Nous avons présenté les résultats de ces comparaisons à une réunion de la communauté de pratique de l'analyse spatiale (Emond et Mather, 2019). Le code SAS a été généralisé de sorte que les macros puissent servir à d'autres projets. Les chercheurs se sont reliés et ont participé à cette communauté

de pratique où les projets et les idées sont partagés en analyse géospatiale. Les membres de cette collectivité et les comptes rendus des réunions (disponibles sur GCdocs) témoignent éloquemment des travaux de l'organisme dans ce domaine. Les résultats de recherche ont indiqué que, pour l'ECMS, il n'y a pas beaucoup de valeur ajoutée si on préfère la distance par la route à la distance à vol d'oiseau dans la modélisation de la non-réponse. L'enquête adoptera pour l'avenir la distance par route, mais nous ne prévoyons pas poursuivre les travaux dans ce domaine.

Bibliographie

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*.

Emond, N., et Mather, A. (2019). Distance by Road – CHMS, Présentation à the Spatial Analysis Community of Practice.

SOUS-PROJET : Reconnaissance optique de caractères dans le cadre de la modernisation du programme des dépenses des ménages

L'Enquête sur les dépenses des ménages collecte de l'information à propos des dépenses des ménages canadiens à l'aide de deux modes principaux, soit une entrevue rétrospective personnelle assistée par ordinateur et un journal de dépenses (format papier). Pour la composante journal, les répondants sont invités à transcrire leurs dépenses et/ou à inclure dans une pochette les reçus concernant leurs achats pendant une période d'une semaine. Ces reçus sont numérisés au bureau central et l'information est présentement saisie manuellement à partir des images des reçus. Pour chaque cycle de collecte, les données d'environ 30 000 reçus sont ainsi saisies de façon manuelle, ce qui représente un fardeau important pour l'agence. L'équipe en méthodologie a entamé un projet de recherche et développement concernant l'automatisation de la saisie de l'information contenue sur les reçus. En première phase, un algorithme d'apprentissage automatique (réseau de neurones) a été entraîné à la reconnaissance des logos des magasins, sous l'hypothèse que la disposition sur les reçus des éléments à saisir est différente pour chaque bannière. L'expérience s'est avérée très positive, l'algorithme réussissant à bien classifier les logos pour au-delà de 95 % des reçus de l'échantillon test. Afin d'optimiser ces résultats et de préparer la saisie des items et des montants sur les reçus, l'équipe aimerait effectuer de la recherche en apprentissage automatique pour la reconnaissance de caractères à partir des images des reçus. Ceci pourrait permettre d'extraire le nom des magasins pour lesquels le logo n'a pas été reconnu par notre algorithme et de saisir le reste de l'information d'intérêt. Le contexte se prête très bien à ce

genre de recherche puisque nous disposons d'une banque très importante d'images de reçus et d'information saisie à travers le temps (~10 ans), ce qui est idéal pour l'entraînement d'algorithmes. Bien que des logiciels soient déjà utilisés à l'interne pour la reconnaissance optique de caractères (dans le contexte de formulaires déjà prévus à cet effet), une consultation avec différents intervenants a montré qu'aucune approche n'existait pour la saisie d'information non structurée.

Progrès :

Une collaboration internationale a été établie avec une scientifique de données, Lan Benedikt du Data Science Campus, qui travaille sur un projet très similaire pour le Living Cost and Food Survey de l'Office for National Statistics (Royaume-Uni). Lan avait déjà développé une expertise sur l'implémentation des techniques d'extraction d'information de reçus (Benedikt, 2019) et, ici à Statistique Canada, plusieurs années de reçus saisis et codés sont disponibles, donc une entente a été établie et Lan s'est rendu au Canada pour travailler avec l'équipe pour une durée de 2 semaines (du 26 août au 6 septembre 2019). Les méthodes d'extractions ont été identifiées, les programmes ont été écrits ou modifiés et certains tests ont été effectués. Une présentation sommaire a été donnée à la gestion des deux équipes, à l'Accélérateur de la science des données, qui auront un projet très similaire dans le futur, et aux clients internes qui seraient impliqués dans l'implémentation d'une telle stratégie en production (Benedikt et Mayer, 2019). Concrètement, ce projet a permis d'acquérir des connaissances et programmes en traitement d'images, reconnaissance optique de caractères et des techniques d'extraction d'information (parsing) via soit un dictionnaire de mots ou des méthodes de regex. Étant donné la complexité et variabilité des reçus, l'apprentissage automatique a présentement été jugé non-pertinent pour cette composante du projet, mais reste très pertinent pour la composante d'auto codage des items des reçus. Les résultats préliminaires démontrent du potentiel concret pour l'utilisation des techniques de reconnaissance de caractères pour saisir l'information sur les reçus. Le projet a été présenté au Conseil de recherche et développement de Statistique Canada (Malo et Mayer, 2019) qui a accepté de financer la prochaine phase du projet qui aura lieu du 15 mars au 31 juillet 2020.

Bibliographie

Benedikt, L. (2019). Human-in-the-Loop AI in Government: A Case Study. Présentation donnée lors d'un séminaire méthodolohique de Statistique Canada, Septembre 2019.

Benedikt, L., et Mayer, E. (2019). STC and ONS-DSC collaboration: Automatic Capture and Coding of Shopping Receipts, Présentation à la direction, OID, ISD and DScD, septembre 2019.

Malo, D., et Mayer, E. (2019). Automation of the Capture of Shopping Receipts, Présentation donnée à la Comission de la recherche et du développement de Statistique Canada, décembre 2019.

SOUS-PROJET : Construction d'intervalles de confiance pour les différences de proportions

En 2017, le Comité des méthodes et des normes (CMN) a approuvé la recommandation faite par la Direction de la méthodologie d'adopter comme pratique exemplaire l'utilisation d'intervalles de confiance pour mesurer et déclarer la qualité des estimations. À la suite de cette recommandation, des recherches s'imposent en vue de la création d'un cadre destiné à appuyer l'emploi d'intervalles de confiance. Le présent projet vise à étudier des méthodes en ce sens et à faire valoir une diversité de situations où les analystes utiliseront des intervalles de confiance.

Progrès :

Il y a eu une présentation au Comité d'examen scientifique de la statistique sociale au sujet des lignes directrices de diffusion avec intervalles de confiance pour la déclaration de la qualité (Neusy et Baribeau, 2019). Le but est de progresser dans la voie menant à un ensemble de lignes directrices recommandées et approuvées. Le Comité a entériné un cadre général de règles de diffusion. Comme il n'a pas approuvé de seuils bien précis, des règles provisoires ont été élaborées et documentées pour les enquêtes du Centre d'intégration et de développement des données sociales (Neusy, 2019a). Une présentation a eu lieu au Groupe de travail sur les indicateurs de qualité – Estimation (Neusy, 2019b) en vue de coordonner les chevauchements entre les deux projets. Un court article a été rédigé pour The Survey Statistician au sujet de la déclaration de la qualité des estimations par intervalles de confiance (Neusy, 2020a). On a procédé à des simulations pour évaluer le rendement d'intervalles de confiance modifiés de Wilson dans le cas des estimations de domaine par échantillonnage aléatoire stratifié. Les résultats indiquent que les intervalles modifiés de Wilson (et de Clopper-Pearson) donnent de bons résultats dans le cas des estimations sur domaines de proportions. Les résultats ont été documentés (Neusy, 2020b). D'autres simulations visaient à évaluer le rendement des intervalles de confiance en cas de différence entre deux proportions. Elles ont fait intervenir des variables présentant des degrés divers de corrélation, de l'indépendance à la corrélation étroite. De même, des variables aux différences de 0 % à 20 % pour la population sous-jacente ont été mises en simulation. Le rendement des intervalles de confiance était inégal, bon dans certains scénarios et mauvais dans d'autres. Il faudra pousser la recherche pour pleinement analyser et comprendre les résultats des simulations dans le cas des différences de proportions.

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

5.3 Division des méthodes d'intégration statistique

SOUS-PROJET : Estimation des erreurs de couplage tenant compte du regroupement et de la mise en correspondance

Le problème général est celui de l'estimation des erreurs de couplage (plus particulièrement sous la forme de faux positifs) en fonction des poids de couplage. Il s'agit d'un prolongement du travail de Labrecque-Synnott (2019) où on s'est attaché à l'estimation fondée sur les poids sans tenir compte des contraintes de regroupement ni de mise en correspondance. Le but est à présent de tenir compte de ces contraintes, puisque la presque totalité des couplages dans l'Environnement de couplage de données sociales (ECDS) supposent une mise en correspondance individuelle des fichiers couplés.

Progrès :

Les groupes de liens possibles créés dans un projet de couplage probabiliste peuvent être soit simples (un seul individu du fichier A est mis en lien avec plusieurs individus du fichier B ou vice versa) soit complexes (liens entre plusieurs individus et du fichier A et du fichier B). Une solution théorique a été mise au point dans ces deux cas. Le problème d'estimation des erreurs compte tenu du regroupement et de la mise en correspondance peut être considéré comme un problème de conditionnement dans le cadre de la théorie du couplage d'enregistrements de Fellegi-Sunter (1969) pour laquelle des formules peuvent être dégagées. Selon la structure des événements intervenant dans le couplage, des simplifications peuvent permettre de rendre le tout réalisable dans la pratique. Ces éléments ont été programmés en SAS et mis à l'essai dans un des projets de couplage de l'ECDS. Dans le cas des groupes complexes comportant un grand nombre de liens, le calcul peut prendre un temps très long même avec les simplifications évoquées. On a entrepris de trouver une approximation appropriée à ces cas. Ce travail a été documenté dans Labrecque-Synnott (2020).

Bibliographie

Fellegi, I., et Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

SOUS-PROJET : Optimisation de la mise en correspondance individuelle dans le couplage d'enregistrements

La méthode gourmande fondée sur le poids total supérieur sert actuellement à la mise en correspondance individuelle dans G-Link. Avec cette stratégie, il n'est pas garanti que nous puissions parvenir à une solution optimale, parce que nous faisons des choix locaux. Une fois

le poids total de couplage attribué à chaque paire, l'identification des liens individuels peut se faire comme solution à un problème de programmation linéaire où la fonction objectif à maximiser est la somme des poids des paires en couplage avec pour contrainte que chaque unité du tableau A doit être mise en correspondance avec une seule unité du tableau B.

L'objectif de ce projet est :

- de documenter l'algorithme employé et de l'appliquer sur des données réelles;
- d'améliorer la méthodologie de la mise en correspondance un à un plus particulièrement en tenant compte du nombre de paires à considérer;
- d'implémenter dans G-coup l'algorithme de résolution.

Progrès :

Les progrès ont consisté en un examen de la documentation existante (Bertsekas, 1992; Chipperfield, Hansen et Rossiter, 2018; Hungarian Algorithm, 2013; Jaro, 1989; Jin, 2016; Lee, Xiong, Yu et Li, 2018; Sadinle, 2016; Sahu et Rudrajit, 2007), ainsi que des algorithmes disponibles appliqués dans un logiciel comme celui de Febrl et Relais.

Bibliographie

Bertsekas, D.P. (1992). Auction Algorithms for Network Flow Problems: A Tutorial Introduction.

Chipperfield, J., Hansen, N. et Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. Docklands: *Revue Internationale de Statistique*.

Hungarian Algorithm (2013). Site internet. 16 08 2019. <www.hungarianalgorithm.com/index.php>.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 406, 414-420. 14 05 2019. <https://www.jstor.org/stable/2289924>.

Jin, X. (2016). *Parallel Auction Algorithm for Linear Assignment Problem*. Projet. Stanford.

Lee, M., Xiong, Y., Yu, G. et Li, G.Y. (2018). Deep Neural Networks for Linear Sum Assignment Problems. *IEEE Wireless Communications Letters*, 7.6, 962-965.

Sadinle, M. (2016). *Bayesian Estimation of Bipartite Matchings for Record Linkage*.

Sahu, A., et Rudrajit, T. (2007). Solving the Assignment Problem Using Genetic Algorithm and Simulated Annealing.

SOUS-PROJET : Analyse spatiale

Ce projet de recherche a débuté cette année pour un travail en analyse spatiale dans le contexte du Registre des entreprises (RE). Deux objectifs étaient fixés pour l'année : (1) appliquer un algorithme de calcul de la distance par le réseau routier; (2) créer une cote d'accessibilité piétonnière/proximité pour les entreprises.

Distance du réseau routier ou distance en ligne droite : expérience avec le Registre des entreprises de Statistique Canada

Ces derniers temps, on s'est mis dans beaucoup d'enquêtes à mesurer ou à chiffrer la distance entre les entreprises pour produire des statistiques de proximité. Nombre d'enquêtes utilisent la distance comme simple mesure de l'accessibilité, du risque ou de la disparité. Jusqu'à présent, on a adopté dans la plupart des enquêtes la distance en ligne droite (euclidienne), parce qu'elle est facile à calculer. Cependant, la distance par le réseau routier est plus précise comme distance réelle entre les entreprises, bien que cette autre possibilité soit plus vorace en calcul et coûte aussi plus cher.

Progrès :

En 2019-2020, on a mis en place un programme avec l'algorithme de Dijkstra du chemin le plus court, SAS Proc OPNET, le Réseau routier canadien (RRC) et le Registre des entreprises, le but étant de calculer 906 780 919 distances par le réseau routier et de les comparer aux distances en ligne droite. Nombreuses sont les enquêtes qui ont exploité les capacités des serveurs en puissance pour calculer les distances du réseau routier, ce qui a créé une pénurie de ressources informatiques, d'où l'idée d'appliquer un modèle à la distance linéaire pour dégager une bonne estimation des distances du réseau routier. On a également entrepris de produire un article ou un rapport et une présentation à ce sujet.

Cote de proximité du Registre des entreprises

La croissance des données géocodées a suscité un afflux de nouvelles applications liées aux géodistances. Au début des années 2010, on a assisté à l'avènement du score de marche (indice d'accessibilité piétonnière), des cartes thermiques et autres applications d'analyse

spatiale. Le Walk Score (score de marche) est un type de modèle d'efficacité automatisé qui porte sur la commodité du lieu. Depuis 2016, le Registre des entreprises produit des données géocodées de localisation pour toutes les entreprises. Ce géocodage donne à l'utilisateur une représentation en latitude et longitude des centroïdes de côté d'îlot. Les coordonnées en latitude et longitude ont déjà servi au couplage des enregistrements, au calcul des distances dans les enquêtes sur les transports et à d'autres applications. En 2019, la possibilité de les employer pour créer une cote de proximité dans l'industrie (Système de classification des industries de l'Amérique du Nord ou SCIAN) a été examinée.

Progrès :

En 2019-2020, un cadre méthodologique a été conçu pour la cote de proximité du Registre des entreprises. La méthode consiste à extraire du Registre des entreprises les établissements actifs à adresse valide et à mettre en grappe toutes les unités se situant à une distance préétablie (1 km) en coordonnées de latitude et longitude. On obtient ainsi le nombre d'unités dans une zone de proximité d'un rayon préétabli pour tous les établissements. Les cotes de proximité sont mises à la disposition des utilisateurs.

SOUS-PROJET : INDICATEURS ÉCONOMIQUES : Évaluation de leur utilisation dans le modèle d'occupation des logements

Le modèle d'occupation des logements est utilisé dans le cadre du Recensement de la population pour prédire le statut d'occupation des logements privés au Canada comme étant soit occupé, inoccupé ou annulé (invalide). Le but de ce modèle est de permettre d'identifier le plus de logements inoccupés ou annulés possible pour que leur statut soit vérifié lors de l'opération de vérification des logements, et pour qu'ils soient par la suite retirés de la liste de logements pour lesquels un suivi de non-réponse est nécessaire.

Suite au déroulement de l'opération de vérification des logements de 2019, une hypothèse a été émise selon laquelle des indicateurs économiques reliés à la région des logements permettraient une meilleure identification des logements inoccupés ou annulés qu'elle contient, et donc leur utilisation dans le modèle permettrait d'améliorer celui-ci. Le but de ce projet de recherche était donc d'évaluer si des indicateurs économiques reliés à l'emploi pouvaient être obtenus et utilisés dans le modèle pour en augmenter son pouvoir prédictif. Ceci permettrait d'améliorer les prédictions du statut d'occupation des logements et donc de réduire le coût de l'opération de suivi des cas de non-réponse.

Progrès :

Le projet s'est concentré sur l'exemple concret de fermeture d'usine en évaluant comment l'information sur le Registre des Entreprises, plus précisément le statut d'activité de chacune des entreprises, la date à laquelle le statut a changé et le nombre d'employés affectés par ce

changement, pouvait être utilisée pour améliorer le modèle. Bien que les résultats ne se soient pas montrés concluants comparés à ceux du modèle initial, des recommandations ont été émises quant à la manière de mieux utiliser l'information du Registre des Entreprises si cette avenue devait être explorée de nouveau, et quant à l'apport potentiel de l'utilisation d'un modèle de régression logistique multiniveaux dans le cadre de prédictions pour le statut des logements (Legault, 2020).

Pour obtenir plus de renseignements, communiquez avec :
Michelle Simard (613-293-3192, michelle.simard@canada.ca).

5.4 Centre de collaboration internationale et d'innovation en méthodologie

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

Dans maintes enquêtes économiques et quelques enquêtes sociales, on observe des variables aux distributions asymétriques pour les unités de l'échantillon, ce qui peut créer des valeurs aberrantes et des unités influentes. Les méthodes classiques d'estimation peuvent produire des estimateurs hautement inefficaces là où les échantillons peuvent contenir des unités influentes. L'idée avec une estimation robuste est de diminuer l'effet de ces unités de l'échantillon sur les estimations d'intérêt. Nous employons le concept de biais conditionnel comme mesure de cette influence et de la contribution de chaque unité de l'échantillon à l'erreur d'échantillonnage d'un estimateur. 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 aidé Beaumont, Haziza et Ruiz-Gazen (2013) à concevoir un estimateur robuste. Ce travail intéresse un grand nombre d'enquêtes économiques et sociales de Statistique Canada.

Progrès :

Un prototype élaboré en SAS comporte un grand nombre des spécifications d'estimation robuste formulées dans Beaumont (2017). Il s'agit d'une série de macros pour les diverses fonctions liées à la production d'estimations robustes de domaine. On y trouve une macro pour le calcul des estimations classiques et des estimations robustes de domaine. Les secondes n'ont généralement pas la propriété d'additivité des premières. Il y a donc une macro qui crée en conséquence des estimations cohérentes de domaine à partir des estimations robustes en modifiant légèrement leurs valeurs. Une autre macro produit des poids par recalage pour les unités de l'échantillon comme garantie de reproduction des estimations cohérentes de domaine et de tous les totaux connus de variables auxiliaires. Ces macros peuvent être employées avec le soutien du Centre de collaboration internationale et d'innovation en méthodologie.

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

Dans les enquêtes-échantillons qui recueillent 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 attrayante mesure d'influence qui tient compte du plan de sondage et de la méthode d'estimation. Il se définit comme l'espérance par plan de sondage de l'erreur

d'échantillonnage conditionnée par la sélection d'une unité dans l'échantillon. L'estimation de ce biais est relativement limpide 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 (comme l'Enquête sur les dépenses des ménages), l'estimation de variance s'obtient fréquemment par des méthodes itératives comme le bootstrap. Les méthodes bootstrap s'appliquent normalement par la production d'un jeu de poids bootstrap qui est mis à la disposition des utilisateurs avec les données d'enquête. Nous étudions la façon d'utiliser les poids bootstrap disponibles pour obtenir un estimateur du biais conditionnel. Cet estimateur pourrait alors servir à construire des estimateurs robustes des paramètres de population finie qui seront moins négativement touchés par les unités influentes que les estimateurs pondérés ordinaires. Nous prévoyons évaluer dans une étude en simulation notre estimateur bootstrap du biais conditionnel.

Progrès :

Nous avons conçu un estimateur bootstrap du biais conditionnel à l'aide des poids bootstrap et avons rédigé un projet d'article (Beaumont et Bocci, 2020b). Nous avons lancé l'étude en simulation et obtenu des résultats préliminaires.

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

Le bootstrap sert fréquemment à l'estimation de variance dans des enquêtes au plan d'échantillonnage stratifié à plusieurs degrés. Il est souvent mis en œuvre par la production d'un jeu 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 la pondération bootstrap nécessaire. Elle trouve son emploi avec un échantillonnage stratifié avec remise au premier degré ou dans le cas de fractions de sondage au premier degré négligeables. Certaines enquêtes ne satisfont pas à ces conditions. Le but du présent projet est de proposer pour les plans d'échantillonnage à plusieurs degrés une méthode bootstrap qui s'appliquerait lorsque les conditions pour la validité de la méthode bootstrap de Rao, Wu et Yue (1992) ne sont pas réunies.

Progrès :

Nous avons conçu une méthode bootstrap simple qui est valide même avec des fractions de sondage non négligeables. Elle s'applique à tout plan d'échantillonnage à plusieurs degrés dans la mesure où une méthode bootstrap valide est disponible pour chaque degré de l'échantillonnage. Notre méthode s'applique aussi aux plans de sondage à deux degrés avec échantillonnage de Poisson au second degré. Nous employons ce plan pour établir des poids

bootstrap qui pondèrent la non-réponse. Un projet d'article est actuellement en cours de rédaction (Beaumont, 2020c).

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

Bibliographie

Beaumont, J.-F. (2017). Prototype d'estimation robuste, spécifications méthodologiques. Rapport interne, Statistique Canada.

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

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.

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.

5.5 Programme de développement

Le Groupe de travail sur le développement de talent statistique a été créé et il continue à ouvrir la voie à un programme statistique modernisé. Un certain nombre de cours nouveaux ou remaniés du programme ont été offerts pour la première fois, dont un cours consacré à l'estimation sur petits domaines. Le cours en science des données a été mis à l'essai l'an dernier et présenté avec succès par un éminent professeur d'université; il fait aujourd'hui partie du cursus normal. Un nouveau cours sur R a également été élaboré à l'interne et mis à l'essai. Un des buts pour l'avenir de la formation en statistique est de cultiver la diversité et l'efficacité de cette formation grâce à un accent sur les activités participatives où l'apprentissage actif est privilégié dans la mesure du possible. Les domaines devenus prioritaires sont notamment la science des données et l'apprentissage automatique, tout comme l'intégration des données et la modélisation statistique. Avec ces priorités à l'esprit, on a commencé à travailler sur un nouvel ensemble d'activités d'apprentissage en modélisation statistique.

Comme autres possibilités d'apprentissage, évoquons le grand nombre de cours en ligne recommandés, les webinaires et la communauté de pratique en science des données. Une nouvelle activité en cours d'élaboration, Ad-Lib, est une formation interactive qui se situe quelque part entre le groupe de lecture et le groupe de travail. Un thème y est donné comme point de départ dans une discussion qui évolue au gré des intérêts et des interventions des participants et d'un animateur pendant quelques heures.

Mentionnons enfin que 2019-2020 a été l'année où le programme des micromissions a été lancé à la Direction. Il s'agit d'affectations à court terme de l'organisme, c'est-à-dire d'une forme d'apprentissage actif où des employés de la Direction sont temporairement affectés à des projets à l'extérieur de celle-ci. Le but premier est de donner la possibilité au personnel d'apprendre et de connaître les priorités et les travaux qui ont cours à l'extérieur de la Direction et, dans certains cas, en dehors de Statistique Canada. La première année a été un succès et de nombreux employés y ont participé.

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

5.6 Publication – *Techniques d'enquête*

Techniques d'enquête est une revue internationale accessible à www.statcan.gc.ca/techniquesdenquete des articles dans les deux langues officielles sur divers aspects du développement statistique pertinents pour un organisme statistique. Son comité de rédaction comprend des leaders de renommée mondiale dans les méthodes d'enquête des secteurs gouvernemental, universitaire et privé. Le journal est publié en format HTML entièrement accessible et en PDF.

Les travaux liés aux processus éditoriaux et de production comprennent: la correspondance avec les auteurs, les arbitres, les rédacteurs associés et les abonnés; examen des commentaires des arbitres et des révisions des auteurs; reformater les manuscrits; édition de copies de manuscrits; liaison avec la traduction et la diffusion; et mise à jour d'une base de données des articles soumis. Cela fait partie des activités de transfert de connaissances.

Progrès :

Les numéros de juin et décembre 2019 (45-1 et 45-2) ont été publiés en versions PDF et HTML. Le numéro de juin 2019 comprend 10 articles, dont le Special Waksberg Invited Paper (Rubin, 2019). Le numéro de décembre 2019 comprenait 8 articles et une courte note. La revue a également publié un numéro spécial en mai 2019 dans le cadre d'une collaboration spéciale avec la Revue Internationale de Statistique en l'honneur du professeur J.N.K. Les contributions de Rao. Ce numéro spécial comprenait 8 articles et un court article du professeur J.N.K. Rao.

D'avril 2018 à mars 2019, les pages de *Techniques d'enquête* ont été consultées 27 000 fois et près de 6 000 copies d'articles ont été téléchargées à l'aide d'une méthodologie de métriques Web améliorée. Hormis les articles invités pour les numéros spéciaux, 31 articles ont été soumis pour publication.

En 2019, la publication de 3 numéros de la revue est prévue. En plus des deux numéros réguliers, un numéro spécial présentant certains articles présentés lors d'une conférence intitulée "Théorie et pratique contemporaines de l'échantillonnage des enquêtes : une célébration des contributions à la recherche par J.N.K. Rao" sera publié en collaboration avec l'*International Journal of Statistics*.

Pour obtenir plus de renseignements, communiquez avec :
Susie Fortier (613-220-1948, susie.fortier@canada.ca).

Bibliographie

Rubin, D.B. (2019). Le calage conditionnel et le sage statisticien. *Techniques d'enquête*, 45, 2, 199-210. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019002/article/00010-fra.pdf.

6. Documents de recherche parrainés par le Programme de recherche et développement en méthodologie

Arim, R., Bougie, E., Michaud, I., Tabuchi, T., Yung, W. et Kohen, D. (2019). A mixed-method exploration of the Public Service Employee Survey (PSES) items as potential measures of psychosocial factors in the workplace. Un rapport préparé pour le Workplace Mental Health Performance Measurement Project Group.

Beaulieu, M. (2020). Présentation au panel « Qualité des données dans l'ensemble du gouvernement : qu'est-ce que cela veut vraiment dire ? ». Conférence sur les données du Gouvernement du Canada.

Beaulieu, M., Chepita, R. et Fortier, S. (2019). Building a quality indicators framework in a multi-source environment. Presenté à l'*International Total Survey Error Workshop.*

Beaumont, J.-F. (2020a). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête,* 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-fra.pdf.

Beaumont, J.-F., et Bocci, C. (2020b). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. Document de travail, Direction de la méthodologie, Statistique Canada.

Beaumont, J.-F. (2020c). Bootstrap variance estimation for multistage sampling with application to nonresponse. Document de travail, Direction de la méthodologie, Statistique Canada.

Beaumont, J.-F., et Chu, K. (2020). Statistical data integration through classification trees. Document à présenter au Comité consultative sur les méthodes statistiques, juin 2020, Statistique Canada.

Beaumont, J.-F., Lesage, É. et Rao, J.N.K. (2020). Estimation of the design mean square error in small area estimation. Document de travail, Direction de la méthodologie, Statistique Canada.

Bocci, C., et Beaumont, J.-F. (2019a). Small area estimation methodology of the unemployment rate in special labour areas. Document de travail, Direction de la méthodologie, Statistique Canada.

Bocci, C., et Beaumont, J.-F. (2019b). Small area estimation of unemployment rate at the special labour area level. Document de travail, Direction de la méthodologie, Statistique Canada.

Buresi, G. (2019). Évaluation de la robustesse du modèle de Fay-Herriot pour les petits domaines. Document de travail, Direction de la méthodologie, Statistique Canada.

CEE-ONU (2019). Generic Statistical Data Editing Model (GSDEM) - Version 2.0, Modern Stats by HLG-MOS https://statswiki.unece.org/plugins/servlet/mobile?contentId=117771706#content/view/117771706.

Chu, K., et Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, mai 2019.

Colley, R.C., Christidis, T., Michaud, I., Tjepkema, M. et Ross, N.A. (2019). An examination of the associations between walkable neighbourhoods and obesity and self-rated health in Canadians. *Health Reports*, 30(9), 14-24.

Colley, R.C., Christidis, T., Michaud, I., Tjepkema, M. et Ross, N.A. (2019). The association between walkable neighbourhoods and physical activity across the lifespan. *Health reports*, 30(9), 3-14.

Dasylva, A., Goussanou, A., Ajavon, A. et Abousaleh, H. (2019). Revisiting the probabilistic method of record linkage. Document de travail, Direction de la méthodologie, Statistique Canada.

Do, Q. (2020). *Mixed Methods Research on International Engagement*.

Ferland, M. (2019). Time Series Processing System - What's new in V3.08. Document interne, Statistique Canada.

Gray, D. (2019). A Generalized Framework to Evaluate Imputation Strategies: Recent Developments. Présenté à la Joint Statistical Meetings of the American Statistical Association.

He, Y. (2019). Estimation of errors in privacy-preserving record linkage. Document de travail, Direction de la méthodologie, Statistique Canada.

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, 107-33. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Labrecque-Synnott, F. (2019). Estimation et propagation d'erreurs de couplage. Document de travail, Direction de la méthodologie, Statistique Canada.

Labrecque-Synnott, F. (2020). Weight-based linkage error estimation accounting for grouping and mapping. Document de travail, Direction de la méthodologie, Statistique Canada.

Lapointe, M.-A., et Mischler, L. (2019). Détecter la saisonnalité : Un enjeu actuel de l'analyse des séries chronologiques. Document de travail, Direction de la méthodologie, Statistique Canada.

Legault, J. (2020). Utilisation d'indicateurs économiques dans le modèle d'occupation des logements (DOM). Document de travail, Direction de la méthodologie, Statistique Canada.

Lesage, É., Beaumont, J.-F., et Bocci, C. (2020). Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu de modèle de Fay-Herriot. *Techniques d'enquête* (sous révision).

Li, K. (2019). Data Visualization for an Imputation Strategy Evaluation Tool. Rapport de stage, soumis à l'University of Ottawa.

Lundy, E., et Rao, J.N.K. (2019). Simulation study of model-assisted survey regression estimation. Document de travail, Direction de la méthodologie, Statistique Canada.

Matthews, S., et Dochitoiu, C. (2019). Comparaison des approches d'ajustement saisonnier à travers la représentation d'espace d'états. Compte rendu : Symposium 2019, Prévisions, Statistique Canada.

Matthews, S., et Patak, Z. (2020). Towards real-time estimation through time series modelling. Document de travail, Direction de la méthodologie, Statistique Canada.

Matthews, S., Patak, Z., Picard, F. et Mischler, L. (2020). Technical documentation of development of Nowcasting options in official statistics. Document de travail, Direction de la méthodologie, Statistique Canada.

Matthews, S., Ferland, M., Verret, F. et Habli, N. (2019). De-mystifying Seasonal Adjustment: A visual tool to understand the process. 3rd Seasonal Adjustment Workshop, Washington, D.C.

Matthews, S., Patak, Z., Picard, F. et Mischler, L. (2020). Technical documentation of development of nowcasting options in official statistics. Document de travail, Direction de la méthodologie, Statistique Canada.

Miller, J. (2019). Survival analysis of the Canadian Mortality Database linked to the Canadian Community Health Survey. Document de travail, Direction de la méthodologie, Statistique Canada.

Mischler, L. (2019). Reference Week Adjustment of Employment Insurance Statistics. 3rd Seasonal Adjustment Workshop, Washington, D.C.

Neusy, E. (2019a). Interim Release Guidelines for CSDID Surveys. DMSS Document interne, Statistique Canada.

Neusy, E. (2019b). Reporting Quality using Confidence Intervals. Présentation à la Working Group on Quality Indicators –Estimation.

Neusy, E. (2020a). Reporting the quality of estimates through confidence intervals. *The Survey Statistician*, Country Report, 81, janvier 2020.

Neusy, E. (2020b). Wilson Confidence Interval Simulations. Document interne, Statistique Canada.

Neusy, E., et Baribeau, B. (2019). Quality-based Release Criteria for Social Statistics Part 2, Présentation à la Scientific Review Committee on Social Statistics, septembre 2019.

Oyarzun, J., et Zhang, S. (2019). Business Register's Proximity Score. Présenté à Statistics Canada's Business Register Analysis Meeting, le 7 juin 2019.

Oyarzun, J., et Zhang, S. (2020). Road network distance vs straight-line distance: Statistics Canada's business register experience. Document de travail, Direction de la méthodologie, Statistique Canada.

Rancourt, E. (2019). The scientific approach as a transparency enabler throughout the data life-cycle. *Statistical Journal of the IAOS*, 35, 549-558.

Reedman, L. (2019). A Framework for Responsible Machine Learning Processes at Statistics Canada. Présenté au comité consultatif sur les méthodes statistiques de Statistique Canada, Document interne, octobre 2019.

Sallier, K. (2020). Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis, soumis pour publication (lauréate du prix des jeunes statisticiens de l'IAOS).

Statistique Canada (2019a). G-EST 2.03.001 Notes de version/G-Est 2.03.001 Avis de mise à niveau, https://www150.statcan.gc.ca/n1/fr/catalogue/10H0035.

Statistique Canada (2019b). Statistique Canada : lignes directrices concernant la qualité. Sixième édition. Statistique Canada, nº 12-539-X au catalogue. Ottawa, Ontario. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-fra.htm.

Statistique Canada (2020). G-Sam 1.03.001 Notes de version/G-Sam 1.03.001 Avis de mise à niveau, https://www150.statcan.gc.ca/n1/fr/catalogue/10H0031.

Verret, F., et Dochitoiu, C. (2019). Estimating the variance of seasonally-adjusted series of monthly Statistics Canada surveys. *Proceedings of the Joint Statistical Meetings of the American Statistical Association*.

You, Y. (2019). A summary report on "Supplementing small probability samples with nonprobability samples: A Baysian approach". Rapport interne du CCIIM, Statistique Canada.

You, Y. (2020a). EBLUP and Hierarchical Bayes small area estimation of LFS rates using area level models with sampling variances smoothing vs modeling. Rapport interne du CCIIM, Statistique Canada.

You, Y. (2020b). Report on hierarchical Bayes small area estimation of LFS totals using area level models. Rapport interne du CCIIM, Statistique Canada.