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

Programme de recherche et développement en méthodologie : réalisations, 2020-2021

Date de diffusion : le 6 octobre 2021

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.

Programme de recherche
et
développement en méthodologie :

réalisations, 2020/2021

Le présent rapport fait la synthèse des réalisations en 2020-2021 du Programme de recherche et développement en méthodologie (PRDM) de la Direction des méthodes statistiques modernes et de la science des données de Statistique Canada. Ce programme comprend les activités de recherche et développement en méthodes statistiques susceptibles d'être appliquées à grande échelle aux programmes statistiques de l'organisme; ce sont des activités qui, autrement, ne s'exerceraient pas complètement dans le cadre des services réguliers de méthodologie offerts à ces programmes. Ajoutons que, dans le but de promouvoir l'utilisation des résultats des travaux de recherche et de développement, le PRDM comporte des activités de soutien aux clients pour la mise en application de travaux de développement antérieurs fructueux. Certaines activités de recherche exploratoire sont également rapportées. 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@statcan.gc.ca)

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

Programme de recherche et développement en méthodologie :

réalisations, 2020/2021

Table des matières

1 Modélisation, intégration des données et science des données

1.1 Estimation sur petits domaines

Les estimations classiques fondées sur le plan de sondages de paramètres de population, appelées 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 pour les petits domaines moyennant l'introduction d'hypothèses de modélisation. Ces dernières années, un regain d'intérêt a eu lieu à Statistique Canada pour l'étude et l'utilisation de méthodes d'estimation indirecte fondée sur un modèle pour les petits domaines. Les principaux objectifs du présent projet sont les suivants :

- élaborer de nouvelles méthodes d'estimation sur petits domaines qui tiennent compte des problèmes observés dans la production d'estimations sur petits domaines;
- étudier les propriétés des méthodes existantes sous différents scénarios afin de mieux comprendre comment et quand les utiliser;
- déterminer une méthodologie d'estimation sur petits domaines appropriée pour certaines enquêtes prometteuses;
- développer et tester des prototypes mettant en œuvre des méthodes nouvelles ou existantes susceptibles de profiter aux programmes statistiques.

Jusqu'à présent, des progrès ont été réalisés dans le cadre des sous-projets suivants.

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

Les outils de validation de modèles, comme les graphiques des résidus, sont souvent utilisés pour évaluer la plausibilité du modèle de Fay-Herriot. Des estimations de l'erreur quadratique moyenne (EQM) fondées sur des modèles sont ensuite utilisées pour évaluer les gains d'efficacité générés par les estimations sur petits domaines par rapport aux estimateurs directs. Toutes ces techniques sont utiles pour évaluer le rendement global des estimations sur petits domaines. Toutefois, les utilisateurs ne s'intéressent souvent qu'à leur domaine particulier, et un indicateur de la qualité des estimations pour leur domaine est plus pertinent de leur point de vue. L'EQM fondée sur un modèle atteint en partie cet objectif, mais intègre l'effet aléatoire local (l'erreur de modèle de couplage) qui intéresse les utilisateurs d'un domaine particulier. L'EQM fondée sur le plan de sondage serait plus pertinente pour ces utilisateurs, mais on connaît la grande instabilité des estimations sans biais de l'EQM fondée sur le plan de sondage (par exemple, Rao, Rubin-Bleuer et Estevao, 2018). Dans le cadre de ce projet, nous avons élaboré et étudié deux nouveaux diagnostics locaux pour l'évaluation d'estimations sur petits domaines.

Progrès :

Un document a déjà été rédigé et soumis à *Techniques d'enquête*. Nous avons reçu les commentaires des examinateurs cette année et révisé le document en conséquence. Le document a été accepté et sera publié dans un numéro de *Techniques d'enquête* de 2021 (Lesage, Beaumont et Bocci, 2021).

SOUS-PROJET : Modèles et estimation robustes sur petits domaines avec application à l'aide des données de l'Enquête sur la population active

Le modèle de Fay-Herriot, qui comporte des erreurs d'échantillonnage normales et des effets aléatoires, est largement utilisé dans l'estimation sur petits domaines pour améliorer les estimations d'enquête directes. Ghosh, Myung et Moura (2018) ont proposé des lois a priori *t* pour les effets aléatoires et une loi a priori de Jeffreys modifiée au moyen d'une méthode hiérarchique bayésienne (HB). Dans le cadre de ce projet, nous évaluerons les modèles HB avec des distributions normales et *t* basées sur le modèle de You et Chapman (2006), nous comparerons les divers modèles HB sur petits domaines et étudierons les effets de la modélisation robuste de l'estimation sur petits domaines (EPD) avec la distribution de la loi a priori *t* pour les erreurs d'échantillonnage et les effets aléatoires. Les modèles et les méthodes proposés seront programmés en R et comparés à l'aide des données de l'Enquête sur la population active (EPA).

Progrès :

Nous avons examiné le modèle You et Chapman (2006) et étudié deux types de modèles EPD robustes fondés sur la loi normale et la loi *t* pour des erreurs d'échantillonnage et des effets aléatoires. La spécification du modèle et les programmes ont été achevés et programmés. Les modèles et les méthodes ont été appliqués aux données de l'EPA aux fins d'évaluation et de comparaison. Un rapport de recherche est terminé (You, 2021a). Le modèle de loi normale et de loi t peut être étendu au modèle de variance d'échantillonnage de Sugasawa, Tamae et Kubokawa (2017) et You (2021b).

SOUS-PROJET : Estimation sur petits domaines au moyen du lissage et de la modélisation de la variance d'échantillonnage

Nous considérons le modèle de Fay-Herriot et étudions les méthodes des meilleurs prédicteurs linéaires sans biais empirique (MPLSBE) et HB pour l'estimation sur petits domaines avec lissage et modélisation de la variance d'échantillonnage. Nous étudions et proposons des méthodes de lissage et de modélisation de la variance d'échantillonnage et comparons les modèles de You et Chapman (2006), You (2016) et Sugasawa et coll. (2017) pour l'estimation du taux de chômage de l'EPA et d'autres problèmes d'estimation sur petits domaines.

Progrès :

Nous avons étudié les méthodes MLPSBE et HB avec lissage et modélisation de la variance d'échantillonnage et nous avons appliqué nos méthodes aux données de l'EPA. Nos résultats indiquent que le lissage de la variance d'échantillonnage peut améliorer l'efficacité et la précision de l'estimateur fondé sur le modèle. La modélisation de la variance d'échantillonnage peut également être utile et se révèle beaucoup plus efficace que l'utilisation d'estimations directes de la variance d'échantillonnage. Un document de recherche sur ce projet a été rédigé et paraîtra dans le numéro de décembre de *Techniques d'enquête* (You, 2021b).

SOUS-PROJET : Prototype d'estimation sur petits domaines

Un prototype SAS a été développé pour produire des estimations sur petits domaines. Ce prototype constitue la base de la mise en œuvre de nouvelles méthodes d'EPD au sein du système généralisé

d'estimation G-Est. Le prototype SAS est utilisé comme outil de production pour répondre à la demande croissante d'estimations à des niveaux désagrégés.

Progrès :

Nous avons mis en œuvre et testé une nouvelle méthode de sélection de modèles pour aider les utilisateurs à déterminer les principales variables auxiliaires du modèle. Il est basé sur une approche de sélection rétrograde utilisant un critère du chi carré pour déterminer les variables explicatives significatives. Cette amélioration permet de réduire considérablement le temps et les efforts requis pour construire un modèle approprié.

SOUS-PROJET : Estimation sur petits domaines des montants moyens des liquidités par division de recensement

Nous étudions l'estimation sur petits domaines des montants moyens des liquidités à l'échelon de la division de recensement. Les estimations d'enquête sur les montants moyens des liquidités à l'échelon de la division de recensement sont calculées à l'aide des données de l'Enquête sur la sécurité financière de 2016. Ces estimations d'enquêtes sont ensuite modélisées en fonction des caractéristiques socio-économiques des familles tirées du recensement de la population de 2016 au même niveau afin de produire des estimations sur petits domaines des montants moyens des liquidités à l'aide du modèle du modèle de Fay-Herriot.

Progrès :

Le modèle de Fay-Herriot a été validé avec soin et des estimations sur petits domaines des montants moyens des liquidités à l'échelon de la division de recensement pour 2016 ont été produites. Un rapport a été rédigé (Bocci, Morissette et Beaumont, 2020) et contient une description des modèles d'estimation sur petits domaines.

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

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

Progrès :

Dans le cadre de recherches antérieures, un estimateur de l'EQM sous le plan fondé sur une approche conditionnelle a été élaboré et un rapport interne préliminaire a été rédigé. Dans la dernière partie de cet exercice, une étude de simulation a été lancée pour évaluer l'approche conditionnelle proposée.

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

BIBLIOGRAPHIE

Ghosh, M., Myung, J. et Moura, F.A.S. (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44, 101-115.

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, 44, 151-166.

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

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Methodology Branch working paper, ICCSMD-2016-03-E.

You, Y., et Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

1.2 Estimation en temps réel au moyen de méthodes de séries chronologiques

SOUS-PROJET : Prévisions immédiates des indicateurs économiques

Nous visons à élaborer un indicateur précoce de l'activité économique canadienne en collaboration avec diverses équipes de Statistique Canada. L'objectif est de tirer parti des méthodes traditionnelles et nouvelles d'apprentissage automatique combinées à des données alternatives pour établir des prévisions immédiates des valeurs mensuelles des activités économiques.

Progrès :

Un environnement infonuagique de modélisation a été établi sur l'espace de travail d'analyse avancée de Statistique Canada. Une vaste base de données de séries chronologiques a été construite avec ElasticSearch et était composée de données économiques et de rechange de différentes fréquences. Des pipelines de traitement ont été construits en Python et en R pour permettre aux spécialistes des séries chronologiques d'accéder à la base de données pour la modélisation. Les résultats de la modélisation indiquent que certaines des sources de données alternatives utilisées peuvent améliorer le rendement des prévisions immédiates et fournir des données pertinentes et opportunes pour la modélisation macroéconomique future. Le rendement des modèles de séries chronologiques de type ARIMAX et PROPHET ainsi que différents modèles d'apprentissage automatique, y compris XGBoost, ont été comparés. Le modèle ARIMAX a obtenu des résultats légèrement meilleurs que PROPHET et le modèle XGBoost ajusté. Les premiers résultats indiquent qu'un modèle hybride (méta) combinant les trois modèles peut améliorer davantage le rendement d'ARIMAX.

Pour de plus amples renseignements sur l'espace de travail d'analyse avancée, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/plateforme-infonuagique.

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

L'augmentation de l'actualité des indicateurs statistiques est une priorité importante pour Statistique Canada et une option pour ce faire consiste à utiliser des modèles de séries chronologiques pour établir des prévisions immédiates des indicateurs économiques beaucoup plus tôt que le moment où le premier estimateur traditionnel est produit.

Progrès :

L'évaluation des modèles statistiques dans ce contexte s'est poursuivie en déterminant les modèles ARRMI et en indiquant les modèles d'espace comme les candidats les plus attrayants. Un exposé a été présenté au Comité consultatif des méthodes statistiques de Statistique Canada afin de recueillir des commentaires sur les progrès et les plans pour ce travail (Matthews, Patak et Picard, 2020).

Deux études de cas ont été menées en collaboration avec la Division de la science des données afin d'évaluer les méthodes prédictives pour la prévision immédiate des permis de bâtir et du produit intérieur brut mensuel. Ces études ont comparé des modèles plus traditionnels comme regARIMA à des modèles prédictifs d'apprentissage automatique, indiquant un rendement similaire et mettant en évidence non seulement les différences entre les méthodes, mais aussi les nombreuses similitudes (Matthews et Ritter, 2020). Un exposé sur ces travaux a été élaboré en vue d'être présenté au Symposium de 2021 sur les questions de méthodologie de Statistique Canada.

Une analyse approfondie des modèles à facteurs dynamiques a été effectuée afin de comprendre la théorie sous-jacente et le potentiel d'application dans le contexte de la prévision immédiate des indicateurs économiques canadiens. Cette méthode est largement utilisée dans d'autres organismes statistiques pour la prévision immédiate. Elle combine des techniques de réduction de la dimensionnalité pour générer des variables auxiliaires, couplées à des modèles linéaires dans un cadre d'espaces d'état pour générer les prédictions. Cette évaluation a été menée en partenariat avec le Dr Rafal Kulik, de l'Université d'Ottawa, ainsi qu'avec un étudiant diplômé, Ismael Zie Diamoutene, qui collabore également à cette analyse. Cette enquête a conclu que la puissance prédictive des modèles à facteurs dynamiques repose fortement sur la disponibilité de plusieurs variables auxiliaires fortement corrélées, qui sont encore identifiées pour les indicateurs économiques canadiens.

Des lignes directrices préliminaires ont été préparées et examinées par le Comité des méthodes modernes : Conversion de données en renseignements, afin de promouvoir la normalisation de la terminologie pour la production d'indicateurs avancés (y compris les prévisions immédiates fondées sur des modèles), de définir des critères à l'appui de la décision de diffuser de nouveaux indicateurs avancés, et de fournir des orientations sur les méthodes appropriées pour établir les prévisions immédiates, ainsi que les avantages et les inconvénients (Matthews, 2020b, 2021b). Des progrès ont été réalisés afin d'achever les lignes directrices qui seront intégrées à l'ensemble des politiques de Statistique Canada. On s'attend à ce que cette étape soit terminée au cours de la première moitié du prochain exercice.

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

1.3 Apprentissage automatique et activités choisies en science des données

SOUS-PROJET : Déterminer des carrefours de la COVID-19 à l'échelon de la région sociosanitaire

Dans les premiers instants de la pandémie de COVID-19, Statistique Canada a entrepris de collaborer avec l'Agence de la santé publique du Canada afin de déterminer et de prédire les points chauds de la pandémie de COVID-19 à l'échelon de la région sociosanitaire en utilisant des modèles d'apprentissage automatique. Le but principal du projet était d'élaborer un modèle de prévision pseudospatiotemporelle à partir de données socio-économiques et socio-démographiques, sur la santé, sur l'épidémiologie et sur la mobilité accessibles au public afin de détourner les ressources de la santé publique des régions à faible risque vers les régions à risque plus élevé.

Progrès :

Des modèles de prédiction des niveaux de risque supervisés (fondés sur l'apprentissage profond) ont été mis au point avec succès (Arim, Hennessey et Molladavoudi, 2021). Ils sont hautement personnalisables et permettraient la prédiction des risques avec diverses plages et fenêtres coulissantes de prévisions (jusqu'à plusieurs jours) au niveau de la région sociosanitaire. Un tableau de bord interactif a également été élaboré afin de permettre aux autorités sanitaires fédérales et provinciales de suivre les tendances des cas et des décès liés à la COVID-19 à l'échelon de la région sociosanitaire et de détourner les ressources de santé publique (par exemple, équipement de protection individuelle, travailleurs médicaux de première ligne médicaux, etc.) des régions à faible risque vers les régions à risque plus élevé, et de contenir plus rapidement les cas dans ces régions par l'isolement, la recherche de contacts et la quarantaine des contacts. La preuve de concept a pris fin avec succès à l'automne 2020. L'étape suivante consisterait à utiliser les modèles et le tableau de bord dans le contexte des unités sanitaires mobiles en collaboration avec Santé Canada.

SOUS-PROJET : Détection d'événements – Détection d'événements à partir d'articles de presse

L'objectif de ce projet est de déterminer les événements économiques liés à des entreprises spécifiques à partir d'articles de presse (provenant de différentes sources comme Factiva et Newsdesk) et de représenter les événements dans une chronologie. Le projet vise également à fournir une façon conviviale d'interagir avec les données en mettant en évidence les événements et les entreprises indiqués sur un tableau de bord.

Progrès :

Pour détecter les entreprises pertinentes dans les articles de presse, nous avons utilisé des algorithmes Named-Entity-Recognition (NER) basés sur BERT qui sont disponibles dans la bibliothèque SpaCy. Étant donné que ces modèles NER sont spécifiques à la langue, des modèles de traduction d'apprentissage profond ont été envisagés. La traduction des articles du français vers l'anglais nous a également permis d'appliquer l'algorithme NER à ces articles. Une approche de classification à étiquettes multiples a ensuite été utilisée pour détecter des événements spécifiques liés aux activités économiques. Pour le module frontal, nous avons utilisé les technologies Dash et Kibana pour concevoir des tableaux de bord. Les résultats sont prometteurs et les prochaines étapes seront axées sur la réalisation d'un plus grand nombre de tests d'utilisateurs.

Pour de plus amples renseignements, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/outil-infonuagique.

SOUS-PROJET : Classificateur de commentaires sur le recensement

À la fin des questionnaires du Recensement de la population, les répondants disposent d'une zone de texte où ils peuvent formuler des commentaires qui sont classés par domaine d'activité, comme l'éducation, le travail ou la démographie, et communiqués aux analystes experts correspondants. Les renseignements sont utilisés pour appuyer la prise de décision au sujet de la détermination du contenu pour le prochain recensement et pour surveiller des facteurs tels que le fardeau du répondant. Au cours d'un projet de preuve de concept, nous avons évalué si des techniques d'apprentissage automatique pouvaient être utilisées pour classer les commentaires des répondants au recensement en 16 catégories différentes.

Progrès :

Nous avons utilisé les données étiquetées provenant de la mise à l'essai du recensement de 2019 pour créer un algorithme de classification de texte semi-supervisé bilingue. Nous avons évalué la performance de quatre modèles différents : machine à vecteurs de support, réseau de mémoire à long ou à court terme simple et multiple, réseau de neurones à convolution multiple. Le projet de preuve de concept a été mené à bien et le classificateur sera utilisé en production à compter de mai pour classer des milliers de commentaires des répondants du Recensement de 2021 du Canada.

Pour de plus amples renseignements, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/recensement-commentaires-2021.

SOUS-PROJET : Indicateur de la Loi canadienne sur l'accessibilité

La mise en évidence des règlements à mettre en œuvre par une entreprise donnée comporte de lire des milliers de plans d'accessibilité longs dont le format n'est pas uniforme, et peut être très subjective. L'idée derrière cette preuve de concept est d'explorer de quelle manière le traitement du langage naturel peut être utilisé pour rechercher plus de 200 exigences réglementaires sur 10 à plus de 150 pages de plans d'accessibilité afin de mettre rapidement en évidence les règlements qu'une entreprise donnée doit mettre en œuvre.

Progrès :

À la fin de la preuve de concept, l'équipe a fourni un modèle de traitement de la langue naturelle qui utilisait une couche d'enrobage de mots préformés pour faire correspondre les éléments du plan d'accessibilité à son règlement. Parallèlement, le projet pilote a déterminé des mesures visant à accroître l'information utile correctement extraite des plans publiés. Il convient de noter que ce projet pilote n'est pas une évaluation exhaustive de tous les plans et règlements; pour appliquer les résultats du projet pilote à une évaluation des plans et des règlements de la *Loi canadienne sur l'accessibilité*, il faudra déployer des efforts considérables pour perfectionner le modèle choisi. Pourtant, ce travail montre que c'est faisable.

SOUS-PROJET : Modélisation thématique dynamique

Ce projet a examiné l'application éventuelle de méthodes de modélisation thématique à la Base canadienne de données des coroners et des médecins légistes (BCDCML) pour détecter les changements dans les profils de décès. La BCDCML contient des données non structurées sous forme de variables en texte libre, appelées textes narratifs, qui fournissent des renseignements détaillés sur les circonstances entourant les décès déclarés. Le but de ce projet est d'appliquer des techniques d'apprentissage automatique à la variable de textes narratifs pour découvrir des structures sémantiques cachées au sein de la BCDCML et d'analyser ces structures de manière dynamique (au fil du temps) afin de détecter les textes narratifs émergents sur les décès.

Progrès :

Une première preuve de concept pour élaborer un système dynamique de modélisation thématique a été conçue, mise en œuvre et déployée à l'aide des données de la BCDCML de la Colombie-Britannique. Le système comprend un modèle, un pipeline d'ingestion et de traitement de données et des outils de visualisation de données. Le modèle regroupe les textes narratifs sur les décès en grappes ayant des structures sémantiques similaires (sujets) tout en utilisant une approche de modélisation de sujet dynamique bayésienne novatrice dans laquelle les sujets précédemment appris sont utilisés comme lois a priori bayésiennes pour les sujets actuels. Pour interpréter, visualiser et analyser les résultats de modélisation, une interface utilisateur de tableau de bord a également été déployée. Les prochaines étapes sont les suivantes : (1) adapter ce modèle aux autres provinces canadiennes; et (2) élaborer des mesures d'indicateur afin de détecter les changements importants dans différents sujets.

SOUS-PROJET : Apprentissage de stratégies optimales d'atténuation des effets de la COVID-19 en utilisant l'apprentissage par renforcement

Un environnement de simulation orientée agents a été créé pour représenter la population canadienne pendant la pandémie de COVID-19. L'apprentissage par renforcement a été utilisé pour apprendre les comportements des agents qui réduisent au minimum la propagation de la COVID dans l'environnement de simulation. L'environnement de simulation a été conçu à l'aide de données disponibles gratuitement (de Statistique Canada, l'Institut canadien d'information sur la santé et l'Agence de la santé publique du Canada). L'objectif est d'optimiser les stratégies d'intervention non pharmaceutique mises en œuvre et de déterminer l'ensemble optimal de comportements de la population contribuant à réduire au minimum la propagation d'une infection dans l'environnement de simulation.

Progrès :

Une population comportant 50 000 agents a été construite et 100 simulations ont été effectuées lors de l'application de l'apprentissage par renforcement. Les comportements des agents ont été optimisés afin de réduire au minimum le nombre d'infections dans l'environnement de simulation. Les comportements ont été analysés et des idées intéressantes ont été découvertes. De plus, des scénarios incluant la « fatigue associée à la COVID » et la non-conformité ont été étudiés. Le projet a été achevé avec succès et a mené à une publication en tant que chapitre d'un prochain livre sur la modélisation mathématique pour la COVID-19 (Denis, El-Hajj Drummond, Abiza et Gopaluni, 2021).

Pour de plus amples renseignements, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/inp.

SOUS-PROJET : Évaluation des stratégies de réouverture des immeubles de bureaux pendant la COVID-19 à l'aide de bandits manchots

Le retour au travail et la « réouverture » à la suite des fermetures d'entreprises et d'industries liées à Covid ont été envisagés. Ce projet a été réalisé en collaboration avec l'Agence de la santé publique du Canada et a consisté à créer un environnement de simulation axé sur des agents pour des immeubles de bureaux de différentes tailles, comportant un nombre varié d'étages et d'ascenseurs, afin de représenter une journée de travail selon divers scénarios. Plus de 200 000 scénarios de travail différents ont été explorés, comprenant la présence de mécanismes de détection, le contrôle de la capacité des employés, la répartition des employés sur les étages, la taille des réunions, l'utilisation des ascenseurs ainsi que l'horaire de travail. Nous tirons parti de la théorie de la décision séquentielle pour contrôler l'échantillonnage de scénarios de simulation à l'aide de bandits manchots. Grâce à cette approche, chaque scénario a été évalué en fonction de la capacité de réduire au minimum le nombre d'événements d'infection survenant dans le milieu de travail. Il s'agit de la première méthode de modélisation à envisager des scénarios de retour au travail pour un environnement de travail dans un immeuble de bureaux, et qui a donné lieu à des recommandations partagées à l'ensemble du gouvernement concernant les décisions relatives au retour au travail dans les immeubles de bureaux.

Progrès :

Ce projet a été achevé avec succès. Plus précisément, un environnement de simulation fondé sur des agents représentant différents immeubles de bureaux de taille différente a été créé. Plus de 200 000 scénarios pour des immeubles à bureaux distincts ont été simulés, comprenant une vaste gamme de facteurs et de variables propres au milieu de travail des immeubles de bureaux. Des bandits manchots ont été utilisés pour contrôler le processus d'échantillonnage. De plus, les principales conclusions et recommandations découlant de ce travail ont été communiquées aux organismes et aux ministères de l'ensemble du Canada. L'Agence de la santé publique du Canada et Statistique Canada visent à préparer et à publier un manuscrit.

SOUS-PROJET : Estimation des superficies consacrées aux cultures sur la plate-forme d'analyse des données en tant que service (ADS) à l'aide de l'imagerie satellite

S'appuyant sur le succès de deux preuves de concept antérieures, ce projet vise à construire un nouveau modèle d'apprentissage automatique qui prend comme entrée un nombre variable d'images satellites d'un champ de culture particulier au Canada et effectue une régression sur l'image, estimant la superficie de 46 types de cultures différentes. L'imagerie satellite est jumelée aux données sur l'assurance-récolte pour certaines provinces canadiennes, à savoir l'Alberta, le Manitoba et la Saskatchewan. Ce projet permettra à Statistique Canada d'effectuer l'estimation des superficies en culture en saison dès juin ou juillet. L'objectif est d'effectuer un essai parallèle pour la campagne de culture 2021 et de réduire ou même de remplacer les enquêtes coûteuses.

Progrès :

Grâce à l'espace de travail d'analyse avancée de Statistique Canada, nous avons effectué le prétraitement coûteux de sept années de données historiques, en utilisant l'imagerie Landsat-8. Les données prétraitées sont mises à disposition pour former un réseau neuronal, et la quantité massive de données a conduit au modèle le plus précis qui puisse être atteint. En parallèle, le prototype de « système de production » est en cours de développement, qui utilisera en fin de compte le modèle d'apprentissage automatique pour produire des estimations en continu pour les analystes agricoles.

Pour de plus amples renseignements, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/imagerie-satellite.

Pour de plus amples renseignements sur l'espace de travail d'analyse avancée, veuillez consulter : https://www.statcan.gc.ca/fra/science-donnees/reseau/plateforme-infonuagique.

SOUS-PROJET : Données fondées sur la disposition spatiale et extraction de contenu

Le format de document transférable (PDF) est le plus fréquemment utilisé par les entreprises aux fins de production de rapports financiers. L'absence de moyens efficaces d'extraction des données de ces fichiers PDF hautement non structurés d'une manière qui tienne compte de la mise en page présente un défi important pour les organismes statistiques pour analyser efficacement les données et les traiter dans cette riche source de données administratives en temps opportun. Cette recherche présente un nouvel algorithme de vision informatique, à savoir Données fondées sur la disposition spatiale et extraction de contenu, conçu et développé par Statistique Canada, qui utilise simultanément des données textuelles, visuelles et de mise en page pour segmenter plusieurs points de données dans une structure tabulaire. Cette solution modulaire proposée réduit considérablement les heures d'efforts manuels consacrés à la détermination des renseignements requis et à leur saisie en automatisant l'extraction des variables financières pour une variété de documents PDF.

Progrès :

Au cours des quatre phases de ce projet, un certain nombre de méthodes ont été testées et appliquées. Dans son état final, les Données fondées sur la disposition spatiale et extraction de contenu utilisent des pixels pour déterminer les divisions idéales de page et définir la page entière en plusieurs sous-sections rectangulaires. Une fois définies, ces sous-sections sont exploitées à l'aide de techniques graphiques transversales pour déterminer la position de chaque jeton dans une matrice tabulaire. Le produit a été présenté à Statistique Canada et à d'autres ministères et il est maintenant prêt à être mis en production.

SOUS-PROJET : Détection des chantiers de construction au moyen de l'apprentissage profond

L'objectif du projet est de démontrer la possibilité de réduire le coût d'identification et de catégorisation des chantiers de construction et de leurs propriétés en automatisant complètement le processus de détection. Grâce à l'extraction à partir d'images satellites à l'aide de l'apprentissage profond, cela permettrait une détection plus rapide (mensuelle) des chantiers de construction. Pour détecter les chantiers, nous avons adopté une approche supervisée fondée sur l'apprentissage automatique pour classifier la classe d'objets pour chaque pixel dans une image (segmentation sémantique).

Progrès :

Nous avons développé une structure de données et de métadonnées pour le déroulement des opérations collaboratives, un pipeline de traitement d'images en parallèle qui peut traiter 100 Go d'images pour créer des ensembles de formation sur l'espace de travail d'analyse avancée, accessibles aux collaborateurs.

Nous avons conçu plusieurs modèles d'apprentissage automatique en utilisant l'approche de segmentation sémantique (réseau U) de Ronneberger, Fischer et Brox (2015) pour la détection de construction basée sur les données d'image et évaluée sur des secteurs de validation et d'essai invisibles. Nous avons conçu un pipeline d'après-traitement pour traiter les prédictions du modèle à l'échelon des pixels et créer des polygones. Sur un ensemble d'essais, le modèle de début de construction a obtenu des résultats d'une qualité suffisante pour justifier la poursuite de cette preuve de concept plus loin.

SOUS-PROJET : Prévisions d'occupation des hôpitaux à court terme en lien avec la COVID-19 à Ottawa

Un modèle bayésien hiérarchique a été construit pour générer des prévisions locales d'occupation des hôpitaux en lien avec la COVID-19, fondées sur le chiffre quotidien des nouvelles admissions à l'hôpital liées à la COVID-19 et le chiffre quotidien du recensement à minuit des hôpitaux. Les deux phénomènes clés modélisés sont le retard aléatoire entre l'infection (non observée) et l'admission (observée) à l'hôpital, et celui entre l'admission à l'hôpital et le congé/décès (observé).

Ce projet de preuve de concept a été commandé par Santé Canada en septembre 2020. Les prévisions sous-provinciales d'occupation des hôpitaux en raison de la COVID-19 pourraient éclairer leur prise de décision relativement au déploiement des ressources dans le cadre de la réponse à la pandémie au Canada. L'objectif de ce projet était de déterminer la faisabilité de produire de telles prévisions avec suffisamment d'exactitude, en se fondant uniquement sur les données d'admission et d'occupation quotidiennes à l'hôpital, en utilisant Ottawa comme scénario d'essai. Ce modèle a été adapté de Flaxman et coll. (2020).

Progrès :

Un modèle a été élaboré pour générer des prévisions à court terme de l'occupation des hôpitaux d'Ottawa en lien avec la COVID-19, fondées sur le chiffre quotidien des nouvelles admissions à l'hôpital liées à la COVID-19 et le chiffre quotidien du recensement à minuit des hôpitaux d'Ottawa. Un pipeline de modélisation et de prévision a été mis en œuvre et déployé sur l'espace de travail d'analyse avancée de Statistique Canada (infrastructure informatique collaborative en nuage). À l'aide de données ouvertes déclarées pour la période allant du 10 février 2020 au 9 mars 2021, un test d'efficacité a été effectué sous la forme d'une série de simulations de prévisions, en tronquant les données de formation pour 15 lundis consécutifs (du 23 novembre 2020 au 1er mars 2021). Un modèle a été adapté à chacun des 15 ensembles de données tronqués, des prévisions ont été générées en conséquence pour une période (jusqu'à 21 jours) suivant immédiatement la date de troncation des données de formation, et les prévisions ont été comparées aux données déclarées. Le modèle a donné des résultats satisfaisants, malgré certaines limites, étant donné la simplicité et la portée limitée des données observées. Les travaux ultérieurs possibles comprennent l'expansion vers d'autres administrations, en attendant la disponibilité des données locales. Les travaux sont documentés dans Bosa et Chu (2020).

SOUS-PROJET : Modélisation épidémiologique de la COVID-19 pour l'estimation de la demande d'équipement de protection individuelle

La pandémie de SRAS-CoV-2 (COVID-19) a imposé une demande sans précédent sur le gouvernement du Canada pour qu'il fournisse des renseignements opportuns, exacts et pertinents afin d'éclairer l'élaboration de politiques portant sur une foule d'enjeux, y compris l'acquisition d'équipement de protection individuelle (EPI) et le déploiement des EPI dans les provinces et les territoires. Des modèles épidémiologiques peuvent être utilisés pour extrapoler la trajectoire des épidémies, selon différentes hypothèses futures, permettant aux décideurs d'envisager un éventail de scénarios. L'un des objectifs de ce projet consiste à améliorer le modèle compartimental dynamique stratifié selon l'âge élaboré par l'Agence de la santé publique du Canada (ASPC) en collaboration avec Statistique Canada (Ludwig et coll., 2020) en mettant en œuvre d'autres scénarios de pandémie. Dans ce modèle, la population est attribuée à divers compartiments et traverse le modèle à un rythme défini.

Progrès :

Un modèle de vaccin a été conçu selon lequel les phases de vaccination ont été divisées en quatre compartiments : 1) première injection sans avoir encore développé d'immunité; 2) première injection et développement d'une immunité partielle; 3) deuxième injection et toujours une immunité partielle; 4) deuxième injection et développement d'une immunité maximale. Le passage d'un compartiment vaccinal à l'autre est contrôlé par le nombre de doses disponibles que chaque province peut distribuer ce jour-là. Conformément à la ligne directrice du Comité consultatif national de l'immunisation (ASPC, 2021), le modèle simule le processus de vaccination qui consiste à donner la priorité aux personnes âgées, puis à passer vers la population plus jeune, tout en permettant la distribution d'un petit nombre de doses au groupe d'âge moyen pour expliquer l'effet de la vaccination des travailleurs de la santé. Différents flux d'un compartiment à l'autre ont été simulés en tenant compte de l'efficacité du vaccin, du taux de vaccination et du niveau d'immunité.

Un modèle pour les variants préoccupants a été élaboré et mis à l'essai pour modéliser l'effet que peut présenter un variant spécifique (B117) dans les provinces canadiennes. D'abord, les constatations de Davies et coll. (2021), qui ont indiqué que l'infectiosité du B117 peut être plus élevée que celle du variant de type sauvage, ont été ajoutées au modèle. De plus, Challen et coll. (2021) ont indiqué que le B117 augmente le risque d'admission à l'hôpital (c.-à-d. cas graves), et cette information a également été incorporée au modèle.

Pour obtenir plus de renseignements, communiquez avec :
Saeid Molladavoudi (613-294-7418, saeid.molladavoudi@statcan.gc.ca**)**.

BIBLIOGRAPHIE

Challen, R., Brooks-Pollock, E., Read, J.M., Dyson, L., Tsaneva-Atanasova, K. et Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. BMJ, 372.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A., Russell, T.W., Tully, D.C., Washburne, A.D. et Wenseleers, T. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. Science, 372(6538).

Flaxman, S., Mishra, S., Gandy, A. *et al.* (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature **584**, 257–261. https://doi.org/10.1038/s41586-020-2405-7

Ludwig A., Berthiaume P., Orpana H., Nadeau C., Diasparra M., Barnes J., Hennessy D., Otten A. et Ogden N. (2020). Assessing the impact of varying levels of case detection and contact tracing on COVID-19 transmission in Canada during lifting of restrictive closures using a dynamic compartmental model. Canada Communicable Disease Report, 46(11/12):409-421.

Public Health Agency of Canada (PHAC). (2021). Recommendations on the use of COVID-19 vaccines. National Advisory Committee on Immunization (NACI): Statements and publications. https://www.canada.ca/content/dam/phac-aspc/documents/services/immunization/national-advisory-committee-on-immunization-naci/recommendations-use-covid-19-vaccines/recommendations-use-covid-19-vaccines-en.pdf

Ronneberger, O., Fischer, P. et Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

1.4 Couplages d'enregistrements

Le couplage d'enregistrements joue un rôle important dans la production de statistiques officielles. Il est cependant sujet aux erreurs, car il se fonde souvent sur des quasi-identificateurs non uniques qui sont enregistrés avec des variations et des erreurs typographiques. Le projet porte sur la production et l'utilisation de données couplées, y compris sur l'estimation exacte des erreurs de couplage.

SOUS-PROJET : Estimation des erreurs de couplage

L'estimation précise des erreurs de couplage est essentielle pour les applications, y compris l'estimation du nombre de faux négatifs attribuables au groupage. Ce projet évalue ces erreurs sans aucune vérification manuelle avec une extension du modèle de Blakely et Salmond (2002), pour tenir compte de l'hétérogénéité de l'enregistrement (voir les détails dans Dasylva et Goussanou, 2020). Ce modèle implique un mélange fini, où chaque composante est la somme d'une variable de Bernoulli avec une variable de Poisson indépendante. L'estimation est fondée sur la maximisation numérique d'une vraisemblance composite.

Progrès :

Le modèle a été utilisé pour estimer le nombre de faux négatifs attribuables au groupage lors du couplage de deux sources sans doublons, dont un dossier et un registre ou un recensement, avec une couverture presque complète. La méthodologie proposée a été évaluée avec succès dans une étude empirique, qui est décrite dans un article devant être publié dans *Techniques d'enquête* (Dasylva et Goussanou, 2021). Cela présente un intérêt au moment de coupler les données fiscales ou sur la mortalité et le recensement (Blakely et Salmond, 2002; Statistique Canada, 2017).

Une nouvelle procédure d'estimation a été élaborée pour tenir compte de certaines limitations de la précédente. En effet, cette procédure utilisait l'espérance-maximisation et était lente à converger. En outre, le nombre de composantes du mélange devait être sélectionné manuellement. La nouvelle procédure utilise plutôt une procédure de Newton-Raphson plus rapide et sélectionne automatiquement le nombre de composants en fonction du critère d'information d'Akaike. Il fournit également des intervalles de confiance du rapport de vraisemblance fondés sur un sous-ensemble des observations. Cette procédure peut être utilisée pour estimer les taux d'erreur à la fin d'un couplage, quelle que soit la méthode choisie (par exemple, probabiliste, déterministe, hiérarchique, etc.), à condition que la décision de coupler deux enregistrements ne dépende d'aucun autre enregistrement. Avec la méthode probabiliste, elle peut également être utilisée pendant le développement pour sélectionner automatiquement les poids de couplage et les seuils (compte tenu des taux d'erreur cibles), sans avoir à préciser la manière dont les variables sont corrélées. Par conséquent, elle simplifie grandement le processus d'estimation des erreurs et élimine les hypothèses d'indépendance conditionnelle qui sont une source possible de biais de corrélation (Blakely et Salmond, 2002; Newcombe, 1988, p. 149).

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

BIBLIOGRAPHIE

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

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Statistics Canada (2017). *2016 Census of Population Income Reference Guide*. Catalog no. 98-500-X2016004.

1.5 Recherche qualitative

SOUS-PROJET : Perceptions de la sensibilité des données

Le Centre de ressources en conception de questionnaire (CRQE) entretient un lien permanent avec les Canadiens lors des interviews cognitives. Ce lien est un excellent véhicule permettant de recueillir des données sur la sensibilité perçue de diverses sources de données d'enquête et non d'enquête et de contribuer en partie au projet d'échelle de sensibilité du Cadre de nécessité et de proportionnalité.

Progrès :

Une première ébauche d'un court questionnaire a été élaborée et administrée aux personnes qui participent aux interviews cognitives habituelles du CRQE. À la suite de cette première phase, on a obtenu des résultats d'enquête préliminaires, rédigé un rapport interne (Reicker, 2020) et recommandé des modifications au questionnaire. Le questionnaire mis à jour a ensuite été utilisé dans la collecte pour une brève deuxième phase. Les résultats de cette deuxième phase devraient être communiqués en 2021.

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

2 Confidentialité / Accès

La recherche sur la confidentialité à Statistique Canada continue de se concentrer sur le développement de nouvelles méthodes et idées qui offrent d'autres formes 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 Méthodes axées sur la perturbation

SOUS-PROJET : Ajustement tabulaire aléatoire

L'ajustement tabulaire aléatoire (ATA) 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 :

Les données de l'Enquête annuelle sur la recherche et le développement dans l'industrie canadienne (RDIC) ont été diffusées en décembre 2020, selon la méthode de l'ATA. Cette diffusion s'appuie sur la première utilisation réussie de la méthode de l'ATA, en 2019, pour la diffusion des données de l'Enquête sur l'innovation et les stratégies commerciales (ENSI). En 2020, l'accent a continué d'être mis sur l'élaboration d'idées visant à permettre l'application de l'ATA à une plus grande variété de produits de Statistique Canada, à mesure que l'organisme s'efforçait de réduire sa dépendance à l'égard des stratégies de suppression.

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 prototype d'ATA a été conçu, mais cette stratégie comportait des problèmes d'optimisation, car les problèmes mathématiques peuvent rapidement exploser à un point où celle-ci n'est pas pratique. Les travaux futurs examineront certains compromis qui pourraient permettre de résoudre les problèmes plus facilement, et sur l'application pour les petits programmes et études utilisant un environnement informatique à haute puissance, comme la technologie infonuagique et les serveurs à haute puissance.

SOUS-PROJET : Utilisation de la perturbation pour les fichiers de microdonnées à grande diffusion (FMGD) du recensement

L'application éventuelle de la méthode de la randomisation a posteriori (MRP) (Gouweleeuw et coll., 1998) dans le cadre de la stratégie de protection du contrôle de divulgation des FMGD du recensement canadien a été étudiée dans le but de réduire le nombre de suppressions pour certaines variables clés et d'accroître l'utilité globale des fichiers. La MRP utilise des méthodes de perturbation, plutôt que des techniques de suppression pour protéger les données, au moyen de divers échanges de valeurs variables.

Progrès :

Au moyen d'enquêtes et d'analyses approfondies, plusieurs options de mise en œuvre de la MRP ont été examinées. Une approche privilégiée a été peaufinée. Elle pourrait permettre d'améliorer l'utilité des fichiers des FMGD à l'aide de la MRP. Compte tenu de la nature confidentielle des méthodes de protection appliquées aux FMGD, nous n'aborderons pas d'autres détails, comme la manière dont elles pourraient être appliquées précisément ou si, en fait, la MRP sera appliquée aux FMGD du recensement canadien.

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

BIBLIOGRAPHIE

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. et de Wolf P.-P. (1998) Post *Randomisation for Statistical Disclosure Control: Theory and Implementation,* Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 463-478.

2.2 Expansion de l'accès afin d'inclure les données sur les entreprises

SOUS-PROJET : Fichiers sur les entreprises à grande diffusion

Les fichiers de microdonnées à grande diffusion (FMGD) sont l'une des nombreuses solutions d'accès aux données de Statistique Canada. Les microdonnées originales sont transformées au moyen de méthodes de confidentialité afin de réduire suffisamment le risque de divulgation pour qu'elles puissent être divulguées en toute sécurité au grand public. Statistique Canada n'a jamais diffusé de FMGD pour une enquête auprès des entreprises. Des facteurs relatifs aux données au niveau de l'entreprise apportent un niveau supplémentaire de protection de la confidentialité par rapport aux données au niveau de la personne. Un document décrivant les techniques ainsi que les défis liés aux risques et à l'utilité a été élaboré (Gilchrist, 2020) et comporte une application à une base de données économiques.

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

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

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

2.3 Données synthétiques

La Directive sur le gouvernement ouvert du gouvernement du Canada vise à s'assurer que les Canadiens ont accès à la plus grande quantité possible de renseignements et de données gouvernementales. Une solution pour les données ouvertes est les données synthétiques. Une version synthétique d'une base de données traiterait des questions de confidentialité avec les données personnelles tout en conservant la plus grande valeur analytique possible. Les méthodes utilisées par Statistique Canada pour créer des données synthétiques ont été documentées par Kenza Sallier comme étant la présentation gagnante au prix pour les jeunes statisticiens de 2020 (Sallier, 2020).

L'un des défis avec les données synthétiques est la terminologie et la nomenclature utilisées pour décrire les différents types de données synthétiques ainsi que les diverses méthodes et outils disponibles. La dernière élaboration d'un guide remonte à 2002 et celui-ci est en cours de révision afin d'inclure bien des méthodes modernes qui sont axées sur les fichiers synthétiques, qui préservent le contenu analytique plutôt que de créer de faux fichiers simples. Ce guide peut contribuer à une partie des travaux en cours d'élaboration par la Commission économique des Nations Unies pour l'Europe (CEE-ONU) et le Groupe de haut niveau pour la modernisation des statistiques officielles. Statistique Canada a également contribué à ce groupe en partageant ses expériences (Sallier, 2021).

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

À mesure que Statistique Canada acquiert une plus grande expertise dans la production de fichiers de données synthétiques de haute valeur analytique, nous commençons à relever de nouveaux défis avec la synthèse de données qui préservent les structures hiérarchiques sous la forme de familles où les enregistrements sont couplés et partagent des traits communs qui doivent être préservés. Ces défis se posent également lorsque l'on synthétise d'autres données structurées comme les données d'entreprise. Un document est en cours d'élaboration et se concentrera sur un exemple concret de l'application de cette stratégie aux données sur le revenu des familles.

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

2.4 Cryptographie

SOUS-PROJET : Classification de texte privé sur les données de lecteurs optiques à l'aide du chiffrement homomorphique

Avec la fermeture imminente des centres de données à Statistique Canada, l'organisme est sous pression pour déterminer comment migrer les projets comportant des données sensibles vers le nuage. Une solution consiste à utiliser des techniques de protection de la vie privée pour protéger les données privées de façon cryptographique. Une de ces technologies est le chiffrement homomorphique, qui permet d'effectuer des calculs sur des données chiffrées sans d'abord les déchiffrer. L'une des sources de données sensibles de StatCan sont les données de lecteurs optiques, qui proviennent des principaux détaillants et qui sont utilisées pour calculer l'indice des prix à la consommation, entre autres. L'étape de prétraitement

assisté par apprentissage automatique, qui consiste à classifier les descriptions de produits dans les codes du Système de classification des produits de l'Amérique du Nord, doit être effectuée sur le nuage et protéger les descriptions sensibles qui doivent être chiffrées.

Progrès :

Deux projets de preuve de concept ont été lancés et menés à bien cette année. Des recherches ont été menées sur la faisabilité de la mise en œuvre de l'analyse statistique (moyenne de calcul, totaux et variance) et de la modélisation de réseau neuronal (formation d'un modèle pour classifier les produits et générer des prévisions de produits) à l'aide de données de lecteurs optiques. Un modèle informatique a été conçu et un programme de mise en œuvre du modèle a été élaboré (Zanussi, Santos et Molladavoudi, 2021). À l'avenir, nous voulons évaluer la faisabilité d'un traitement chiffré pour le couplage d'enregistrements à l'aide d'une intersection d'ensembles privés. Lorsque tous les partenaires seront prêts, le projet pourra être prolongé et mis en production.

Pour obtenir plus de renseignements, communiquez avec :
Zachary Zanussi (613 298-1808, zachary.zanussi@statcan.gc.ca)

3 Théorie et cadre

3.1 Inférence provenant d'échantillons non probabilistes

Les bureaux nationaux de statistique souhaitent de plus en plus produire des statistiques officielles à l'aide d'échantillons de données non-probabilistes, comme les mégadonnées ou les données provenant d'enquêtes Web menées auprès de volontaires. En fait, Statistique Canada a récemment mené plusieurs enquêtes en ligne auprès de volontaires, appelées enquêtes par approche participative, pour évaluer les répercussions de la pandémie de COVID-19 sur différents aspects de la vie de la population canadienne. On veut utiliser des échantillons non probabilistes en raison de leur faible coût, du fardeau de réponse réduit, et d'un délai d'exécution rapide puisqu'ils permettent de produire des estimations peu après avoir déterminé les besoins en données. Toutefois, les échantillons non probabilistes sont bien connus pour produire des estimations qui peuvent être entachées d'un biais de sélection important. Beaumont et Rao (2021) discutent de cette importante limite, avec une illustration, et décrivent certains remèdes qui impliquent l'intégration des données de l'échantillon non probabiliste avec les données d'un échantillon probabiliste. Renaud et Beaumont (2020) décrivent quatre initiatives de recherche récentes visant à tirer parti des données d'échantillons non probabilistes.

La manière d'obtenir des estimations adéquates et de faire des inférences valides à partir d'échantillons non probabilistes est une question importante qui nécessite encore des recherches et des expérimentations. Les quatre sous-projets suivants ont tenté de répondre à cette question.

SOUS-PROJET : Réduction du biais des estimateurs d'échantillons non probabilistes par des méthodes de pondération par score de propension avec application aux données d'approche participative

L'objectif de ce projet est d'étudier et d'élaborer des méthodes de pondération par score de propension, qui combinent les données d'un échantillon non probabiliste qui contient des variables d'intérêt et des variables auxiliaires avec les données d'un échantillon probabiliste qui contient les mêmes variables auxiliaires (ou un sous-ensemble d'entre elles). La pondération par score de propension consiste à modéliser la probabilité de participation à l'échantillon non probabiliste.

Progrès :

Nous avons d'abord examiné un modèle logistique (voir Chen, Li et Wu, 2019) et développé une procédure de sélection de variables basée sur un critère d'information d'Akaike modifié (AIC). Notre critère AIC modifié tient compte de la structure des données et du plan de sondage probabiliste, qui peut être complexe. Nous avons également élaboré une méthode simple pour former des strates a posteriori homogènes. De plus, nous avons étendu l'algorithme de l'arbre de classification et de régression (CART) (Breiman, Friedman, Stone et Olshen, 1984) à cette structure de données et nous avons mis au point une procédure d'élagage qui, encore une fois, tient compte du plan de sondage probabiliste. Notre nouvel algorithme s'appelle nppCART. Quelques détails sur la procédure d'élagage sont donnés dans Beaumont et Chu (2020). Nous avons également mis au point un estimateur de la variance bootstrap, qui reflète deux sources de variabilité : le modèle d'échantillonnage probabiliste et le modèle de participation. Nos méthodes sont actuellement évaluées à l'aide de données de l'approche participative et de données de l'Enquête sur la population active (EPA). Un document a été rédigé, qui est sur le point d'être achevé.

SOUS-PROJET : Approche bayésienne approximative pour améliorer les estimateurs d'un échantillon probabiliste à l'aide d'un échantillon non probabiliste supplémentaire

Une méthode bayésienne pour combiner les données d'un échantillon probabiliste et non probabiliste a récemment été diffusée dans le Journal of Official Statistics (Sakshaug, Wisniowski, Ruiz, et Blom, 2019). Ces auteurs ont examiné l'estimation des paramètres du modèle lorsque la variable dépendante *y* et le vecteur des variables explicatives ***x*** sont observés dans les deux échantillons. Ils ont utilisé l'échantillon non probabiliste comme moyen de déterminer la moyenne a priori des paramètres du modèle, en supposant que le plan de sondage probabiliste est ignorable. L'objectif de ce projet est d'étendre leur méthode à l'estimation de paramètres de population finie pour un plan de sondage probabiliste qui n'est pas nécessairement ignorable, et de voir si nous pouvons obtenir des estimations fondées sur un modèle qui sont plus efficaces que les estimations pondérées standard.

Progrès :

Nous avons fait l'extension à l'estimation de la moyenne d'une population finie pour un plan de sondage probabiliste non ignorable. Nous avons également mené des expériences de simulation préliminaires qui ont été résumées dans You (2021c). Nous prévoyons terminer l'étude de simulation et rédiger un document qui sera présenté au Symposium de 2021 de Statistique Canada.

SOUS-PROJET : Estimation d'erreur quadratique moyenne pour les estimations d'échantillons non probabilistes à l'aide de techniques d'estimation sur petits domaines

Les données administratives et d'autres sources de données non probabilistes sont de plus en plus prises en compte par les bureaux nationaux de statistique comme un moyen d'obtenir directement les renseignements recherchés sur une population. Toutefois, les estimations de ces sources non probabilistes sont souvent assujetties à un certain nombre d'erreurs et il faut élaborer des indicateurs de leur exactitude.

Progrès :

Nous avons mis au point un estimateur de l'erreur quadratique moyenne conditionnelle (EQM) des estimations de l'échantillon non probabiliste qui s'appliquent lorsqu'un échantillon probabiliste avec les mêmes variables est également disponible. Notre estimateur conditionnel de l'EQM est obtenu en utilisant les techniques d'estimation sur petits domaines. Nous avons évalué la méthode à l'aide des données du programme Longitudinal Social Development Data (programme longitudinal d'élaboration de données sociales). Le fichier du programme est constitué de dossiers administratifs et permet d'estimer les caractéristiques de la main-d'œuvre. L'Enquête sur la population active (EPA) sert d'échantillon probabiliste. Nous avons commencé à écrire un article qui résume les constatations.

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

Nous considérons le problème où un échantillon non probabiliste est disponible qui contient un vecteur de variables auxiliaires, **x**, pour chaque unité de l'échantillon. Nous supposons que cet échantillon non probabiliste couvre une partie importante de la population. Un échantillon probabiliste contenant ***x*** ainsi que la variable d'intérêt *y* pour chaque unité de l'échantillon est également disponible. L'indicateur de participation à l'échantillon non probabiliste est disponible dans l'échantillon probabiliste. Ce scénario est

pertinent pour une enquête sur le trafic postal menée par La Poste en France. Alain Dessertaine a proposé un prédicteur pour ce scénario. Nous avons mis au point des estimateurs de variance, y compris un estimateur de variance bootstrap, pour évaluer la qualité du prédicteur proposé. Les détails sont donnés dans un rapport interne en ébauche.

Progrès :

La collaboration avec La Poste, l'École d'économie de Toulouse et l'Université de Besançon s'est poursuivie et l'objectif est de rédiger un document commun. Les résultats de ce projet seront d'abord présentés lors d'une session invitée au Colloque francophone sur les sondages à l'automne 2021.

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

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 Cadre pour l'apprentissage automatique

Statistique Canada utilise des techniques d'apprentissage automatique pour résoudre des problèmes de données à grande échelle. 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. Statistique Canada a élaboré un cadre adapté à l'utilisation responsable des techniques d'apprentissage automatique, y compris des lignes directrices pour la construction et la mise en œuvre de processus éthiquement et méthodologiquement solides. Il s'articule autour de 4 thèmes: respect des personnes, respect des données, application rigoureuse et méthodes éprouvées et un certain nombre d'attributs pour chaque thème. Il est aligné avec la Directive sur la prise de décision automatisée et son outil d'évaluation de l'incidence algorithmique, élaborés par le Secrétariat du Conseil du Trésor (2020).

Progrès :

Suite à des mises à l'essai et une revue par la gestion, le cadre pour l'utilisation des processus d'apprentissage automatique de façon responsable a été officiellement adopté en juillet 2020 par Statistique Canada. Une liste de contrôle accompagnant les lignes directrices a été finalisée et est utilisée pour aider à évaluer les processus d'apprentissage automatique responsables. Un processus d'évaluation des applications en apprentissage automatique qui passeront à la production a été élaboré et celui-ci comprend une revue indépendante par des experts, une documentation adéquate et le remplissage de la

liste de contrôle reliée au projet. La méthodologie du projet ainsi que la revue effectuée par les experts sont présentées devant un comité de revue scientifique qui émettra des recommandations sur la méthodologie proposée. Lors de la dernière année, sept projets ont été évalués à l'aide de ce cadre. Les prochaines étapes de ce projet incluent : l'élaboration d'un tableau de bord pour consigner les revues, le développement d'un gabarit de la documentation qui devrait être fournie aux réviseurs et une nouvelle revue de littérature concernant les concepts d'explicabilité et d'interprétabilité pour être à jour dans ce domaine.

Pour obtenir plus de renseignements, communiquez avec :
Keven Bosa (613-863-8964, keven.bosa@statcan.gc.ca).

3.3 Recherche d'indicateurs de qualité

Afin de fournir aux utilisateurs des indicateurs de qualité pour les programmes qui combinent les sources de données administratives, le Secrétariat de la qualité a entrepris de mettre au point un indicateur composite qui combine des indicateurs de qualité liés aux différentes étapes du traitement des données (couplage d'enregistrements, imputation, géocodage, etc.) en un seul indicateur. L'objectif est de donner une vision globale de la qualité d'une estimation en tenant compte de plusieurs facteurs susceptibles d'introduire des erreurs. Un premier programme publiera ces indicateurs et des estimations à l'été 2021. Ce projet sera présenté dans le cadre de l'European Establishment Statistics Workshop en septembre 2021 (Beaulieu et Lebrasseur, 2021).

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

4 Soutien (Centres de ressources)

4.1 Centre de ressource 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 de méthodes de couplage d'enregistrements, y compris de recommander des logiciels, des méthodes et des travaux de collaboration sur les applications de couplage d'enregistrements afin d'évaluer des méthodes nouvelles de couplage ou de développer des méthodes améliorées. 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 :

Nous avons continué à soutenir l'équipe de développement de G-Link et à faire le suivi des enjeux soulevés. 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 transmises au moyen de requêtes sur G-Link_info.

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, repérage et traitement des enregistrements orphelins et pseudoclés intégrées. Se sont ajoutés des travaux visant à explorer le couplage d'enregistrements dans le nuage en utilisant la plateforme Databrics et à intégrer un outil de vérification manuelle (évaluation de la qualité) et à corriger et à 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@statcan.gc.ca).

4.2 Systèmes généralisés

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

Progrès :

Un grand nombre de dossiers de soutien ont été traités par l'équipe de projet, principalement sur G-EST, BANFF et G-SAM. La plupart d'entre eux ont été résolus par des suggestions sur la façon d'appliquer correctement les systèmes. Toutefois, plusieurs d'entre eux ont exigé une plus grande participation. Deux nouvelles versions des systèmes généralisés ont été diffusées cette année. Le G-EST 2.03.002 a été diffusé, y compris des améliorations du rendement pour l'étalonnage et le traitement en parallèle de la variance d'échantillonnage (Statistique Canada, 2020). La version 2.08 de BANFF a également été publiée, y compris des améliorations au processeur BANFF et des corrections de bogues pour améliorer la localisation des erreurs (Statistique Canada, 2021).

Deux cas de soutien ont été déterminés qui comportaient des problèmes de traitement avec de grands ensembles de données complexes dans l'application SEVANI (partie de G-EST) pour estimer la variance due à l'imputation. Par conséquent, des travaux d'élaboration ont été menés pour construire un prototype permettant d'omettre des calculs inutiles pour des cas précis et pour un traitement parallèle. Ces prototypes sont en cours d'essai et feront partie d'une version future de G-EST.

L'élaboration d'ImpACT, un système permettant de visualiser l'imputation, a été poursuivie et le système a été utilisé pour évaluer les stratégies d'imputation de deux projets de Statistique Canada. Ce travail a été présenté au Comité d'examen scientifique de la Direction des méthodes statistiques modernes et de la science des données (Gray, 2020a). Il est prévu d'élargir cet outil afin d'y inclure d'autres visualisations et l'application à d'autres programmes.

Les membres de l'équipe ont participé à la formation au moyen de cours officiels avec l'institut de formation de Statistique Canada, ainsi que de séminaires à l'intention de statisticiens récemment recrutés et d'autres présentations ponctuelles à des analystes et à d'autres organismes. L'équipe a également contribué à l'organisation de l'atelier de la CEE-ONU sur la révision de données et a présenté les travaux relatifs à l'outil ImpACT (Gray, 2020b). Cette présentation a suscité l'intérêt de collègues internationaux et devrait mener à des collaborations.

Une initiative majeure pour les systèmes généralisés est un exercice de planification à long terme pour examiner l'évolution à venir des systèmes, y compris la possibilité d'utiliser des outils libres, l'inclusion de méthodes modernes, et de réexaminer l'approche pour gérer les développements futurs. Dans le cadre de ce travail, un certain nombre d'exigences générales pour les systèmes généralisés ont été décrites (Matthews, 2020a) et des travaux ont été entrepris pour définir les exigences opérationnelles pour chaque système à court, à moyen et à long termes. Ces exigences s'inscriront directement dans le plan d'évolution à long terme de chaque système qui sera élaboré au cours de la prochaine année en partenariat avec l'équipe de technologie informatique (Matthews, 2021a).

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

4.3 Centre de ressources en conception de questionnaire

Le Centre de ressources en conception de questionnaires (CRCQ) est le centre d'expertise de Statistique Canada en matière de conception et d'évaluation de questionnaires. Le CRCQ fournit des services de

conseils et de soutien et mène des projets et des études relatifs à l'élaboration, à la mise à l'essai et à l'évaluation de questionnaires d'enquête. Le CRCQ joue un rôle essentiel dans la gestion de la qualité et répond aux exigences des programmes de l'ensemble de Statistique Canada en consultant les clients, les répondants et les utilisateurs de données et en procédant à l'essai préliminaire des questionnaires d'enquête.

Alors qu'une grande partie du travail du CRCQ est effectuée selon le principe du recouvrement des coûts, cette équipe est souvent sollicitée, de manière ponctuelle, pour fournir des évaluations d'expert et des services de consultation relativement à un large éventail d'enquêtes. Le groupe propose également des cours sur la conception de questionnaires.

Progrès :

Le CRDQ a effectué de nombreux examens de questionnaires d'enquête. Bien que la plupart de ces examens mettaient en cause des questionnaires de Statistique Canada, plusieurs ont été menés pour des enquêtes effectuées par d'autres organismes gouvernementaux, comme la Banque du Canada, le Bureau du surintendant des faillites, le Bureau canadien de l'ombudsman de la responsabilité des entreprises, Affaires mondiales Canada et d'autres.

En réponse aux calendriers d'élaboration très rapides du questionnaire pour les nouvelles enquêtes relatives à la COVID-19, le CRDQ a mis au point et mis en œuvre un nouveau groupe de recherche qualitative, qui a permis de tester rapidement ces nouveaux outils de collecte de données.

Le CRDQ a continué d'expérimenter la recherche sur les méthodes mixtes. De plus, certaines recherches et expérimentations ont commencé avec des méthodes qualitatives asynchrones. Le groupe a également contribué à diverses initiatives de consultation ministérielle.

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

4.4 Secrétariat à 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 de différents programmes de Statistique Canada ou d'autres organismes.

SOUS-PROJET : Renforcement des capacités avec les 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 (le Cadre d'assurance de la qualité et les 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. Au niveau des partenaires nationaux, des exposés officiels sur les pratiques de gestion de la qualité ont été présentés à trois organisations, en plus d'organiser un certain nombre d'ateliers et de séminaires. Des documents sur la qualité des données et les bonnes pratiques de gestion de la qualité ont été fournis à l'Initiative de formation sur la littératie des données de Statistique Canada. Des discussions ont eu lieu au sein du Data Governance Standardization Collaborative ainsi que du Groupe de travail sur la qualité des données. Ce dernier groupe, coprésidé par Statistique Canada, vise à définir un cadre de qualité des données applicable à tous les organismes du gouvernement du Canada dans le cadre de la mise en œuvre de la Stratégie de données. La validation de la qualité d'un processus statistique effectué ainsi que la validation de la qualité d'une source de données utilisée par un autre organisme fédéral ont également été achevées. À l'échelle internationale, le Groupe d'experts des Nations Unies sur les Cadres nationaux d'assurance de la qualité a continué de participer à la préparation de la mise en œuvre du United Nations National Quality Assurance Framework Manual for Official Statistics (Nations Unies, 2019).

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

On trouvera des détails sur ce sous-projet à la section 3.3 (Recherche sur les indicateurs de qualité).

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

Le Centre de ressources pour l'assurance de la qualité (CRAQ) vise à mener des activités de recherche et de développement sur les méthodes statistiques d'assurance et de contrôle de la qualité, afin d'améliorer la qualité sortante des opérations de collecte et de traitement des données d'enquête au sein du bureau. Cela comprend l'offre de services méthodologiques pour G-Code, qui est utilisée à Statistique Canada pour créer des bases de données de codage pour le traitement des données. La recherche sur l'assurance et le contrôle de la qualité est souvent de nature générique et porte sur des questions d'efficience et d'automatisation, qui sont fréquemment appliquées à de nombreuses étapes des opérations d'enquête.

Progrès :

L'équipe de soutien méthodologique a aidé l'équipe de développement et a fait le suivi des commentaires des utilisateurs pour aider à déterminer les idées d'améliorations potentielles pour le G-Code. Le CRAQ a également fourni un soutien aux utilisateurs internes et externes du G-Code (utilisateurs internationaux) lorsque l'aide, les commentaires et les suggestions concernant le G-Code étaient nécessaires.

Au cours de l'année, les travaux ont porté sur la mise en œuvre d'une nouvelle version de G-Code (version 3.2), qui comprenait l'ajout de capacités d'apprentissage automatique (XgBoost et FastText). L'équipe du CRAQ a 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). Les nouveaux algorithmes ont été largement utilisés pour coder l'industrie et la profession pour le Registre des entreprises, l'Enquête sur la population active et de nombreux autres programmes/enquêtes (y compris le Recensement de la population). En outre, ces nouvelles fonctionnalités ont été présentées à des organismes externes (Bureau de la statistique de l'Australie et Statistics New Zealand). Dernièrement, l'équipe du CRAQ a aidé à l'intégration de Pytorch dans G-Code. PyTorch est une bibliothèque d'apprentissage automatique libre basée sur la bibliothèque Torch, utilisée pour des applications telles que la vision informatique et le traitement du langage naturel, principalement développé par le laboratoire de recherche sur l'intelligence artificielle de Facebook.

Le CRAQ a également travaillé à la mise en œuvre d'une variété de contrôles de qualité sur les processus de codage pour l'Enquête sur la population active, l'Enquête sur la santé dans les collectivités canadiennes, l'Enquête sur les postes vacants et les salaires et le Registre des entreprises.

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

4.6 Centre de ressources en analyse de données et consultation

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

Des services de consultation ont été offerts à la demande de clients internes et externes. Les questions variaient en complexité et comprenaient l'utilisation des poids bootstrap d'enquête, les valeurs *p*, la construction d'intervalles de confiance et les tests d'hypothèses avec les données d'enquête, l'estimation de la *taille de l'effet* avec des données d'enquête, l'analyse avec des données couplées, l'ajustement d'un modèle de régression logistique, etc. Nous avons également aidé nos clients à mettre en œuvre des méthodes dans les logiciels SUDAAN, SAS, STATA et R.

Prestation de formation et de matériel de formation

Le CRAD a conçu et présenté la première session du cours de modélisation statistique, intitulée « Modélisation statistique à l'aide de données d'enquête complexes – Régression linéaire ». Ce nouveau cours a été offert virtuellement par le Groupe de travail sur le développement de talent statistique aux employés de la Direction des méthodes statistiques modernes et de la science des données de Statistique Canada (Mach et Michaud, 2020)

Le document « Visualisation des données – Pratiques exemplaires » a été achevé et est disponible à l'interne dans les deux langues officielles (CRAD, 2020a). Il est destiné à servir d'outil aux analystes et aux équipes de diffusion de Statistique Canada.

Collaboration

Nous avons collaboré à l'élaboration de stratégies de mesure pour trois projets : i) Projet de mesure du rendement en santé mentale en milieu de travail, ii) Index du renouvellement de la fonction publique au-delà de 2020, et iii) Stratégie d'accessibilité pour la fonction publique du Canada.

Les trois projets utilisent les données recueillies par le Sondage auprès des fonctionnaires fédéraux (SAFF) pour mesurer des variables latentes, comme les facteurs de risque psychologiques, les comportements, etc. Les modèles de mesure ont été élaborés à l'aide de l'analyse factorielle et de la modélisation d'équations structurelles, comme l'ont discuté CRAD (2020b) et Blais et coll. (2020 et 2021).

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

4.7 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.

Les projets se répartissent en plusieurs sous-catégories, l'accent étant mis sur les thèmes suivants :

- Consultation et formation en séries chronologiques (y compris l'élaboration et la prestation de formations)
- Soutien et amélioration du système de traitement des séries chronologiques
- Développement et soutien pour la désaisonnalisation et l'estimation des tendances-cycles
- Modélisation et prévision, en particulier dans le contexte de l'estimation en temps réel

SOUS-PROJET : Consultation et formation sur les méthodes en séries chronologiques

Le Centre de recherche et analyse sur les séries chronologiques est chargé d'élaborer et de dispenser une formation sur les méthodes en séries chronologiques, y compris la désaisonnalisation, le rapprochement et la modélisation en séries chronologiques, à l'intention des participants internes de Statistique Canada et d'autres organismes. De plus, le Centre fournit des conseils et des consultations sur les projets de séries chronologiques en général pour Statistique Canada, y compris les demandes provenant de l'intérieur et de l'extérieur de l'organisme.

Progrès :

Avec le passage soudain au travail à distance, un nombre réduit de cours ont été donnés. Toutefois, le contenu et l'organisation des cours qui ont été offerts ont dû être modifiés afin de tenir des séances plus

courtes, mais plus fréquentes afin d'offrir le matériel. Plus précisément, des cours sur l'étalonnage (H-0436), le ratissage (H-0437) et l'ajustement saisonnier (H0434) ont été offerts au cours de l'année par l'entremise du centre de formation de Statistique Canada (Statistique Canada, 2021). De plus, un cours sur la modélisation et la prévision des séries chronologiques (H-0433) a été offert aux participants d'Immigration, Réfugiés et Citoyenneté Canada par l'entremise d'un format à distance. Les membres du Centre ont également participé à des activités de sensibilisation et de formation à d'autres groupes de Statistique Canada sur des thèmes touchant les séries chronologiques dans le cadre de la formation des recrues récentes et des agents financiers.

Le Centre a également participé à l'examen de documents de revues de référence concernant la pandémie de COVID-19, en particulier sur la prévision du nombre de cas pour certains pays, et à des discussions informelles avec des statisticiens participant à la modélisation statistique liée à la pandémie. Les membres du Centre ont également participé au Groupe de haut niveau sur l'apprentissage automatique de la Commission économique des Nations Unies pour l'Europe, contribuant à une comparaison entre l'apprentissage automatique et les méthodes de prévision traditionnelles (Picard, 2020). Le Centre a également examiné la méthodologie derrière plusieurs nouveaux produits statistiques diffusés par Statistique Canada, à savoir l'indicateur économique provincial (Statistique Canada, 2021b) et les estimations de la surmortalité (Statistique Canada, 2020), afin de valider les caractéristiques des séries chronologiques dans l'élaboration de ces indicateurs.

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

Le système de traitement des séries chronologiques est une application sur SAS personnalisable qui permet d'appliquer des techniques de séries chronologiques, y compris la validation des fichiers, l'application de stratégies de révision ainsi que des techniques d'ajustement saisonniers et de rapprochement utilisées abondamment dans la production d'estimations désaisonnalisées pour les programmes essentiels à la mission de Statistique Canada. Conçu il y a plus de 10 ans, le système est dans un état relativement mature et stable. Toutefois, il doit être constamment mis à jour afin d'élargir ses fonctionnalités et d'aborder les nouveaux besoins des programmes de l'organisme. À plus long terme, une nouvelle version du système devra peut-être être mise au point pour soutenir le traitement et permettre une certaine flexibilité dans l'utilisation de nouvelles techniques disponibles à partir de logiciels libres.

Progrès :

Des corrections mineures ont été apportées à la version 3.08 du système de traitement des séries chronologiques (STSC) décrit dans Ferland (2019) ainsi qu'à un certain nombre d'outils de soutien pour l'analyse et le développement de systèmes. Un outil qui est utilisé comme élément clé de l'assurance de la qualité a notamment été élargi pour inclure des diagnostics améliorés sur la saisonnalité résiduelle et les statistiques sur les irrégularités relatives à différentes étapes du processus de désaisonnalisation, afin de permettre un dépannage plus efficace et l'élaboration de solutions.

Des recherches ont été menées pour évaluer les outils disponibles afin d'appliquer l'analyse comparative au moyen d'outils libres, y compris ceux disponibles dans R. Une comparaison a été faite entre la fonctionnalité actuelle de G-Series, et un ensemble logiciel R appelé tempdisagg, en termes de méthodes disponibles et de leur paramétrage (Picard, 2021). Statistique Canada s'est joint au Seasonal

Adjustment Centre of Excellence d'Eurostat à titre d'organisme partenaire pour participer aux discussions et à l'élaboration d'outils connexes.

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

Analyse et évaluation des nouvelles méthodes et techniques de désaisonnalisation ainsi que consultation et centralisation de l'expertise dans l'application de la désaisonnalisation.

Progrès :

Le Centre de recherche et analyse en séries chronologiques a apporté un appui considérable à la désaisonnalisation afin d'assurer la qualité des résultats lors des chocs économiques sans précédent liés à la pandémie de COVID 19. De nombreux échanges ont eu lieu avec des représentants des bureaux nationaux de statistique, soit par des échanges de courriels, des conversations individuelles, ou soit par des vidéoconférences en petits groupes. Les échanges comprenaient des membres d'Eurostat, de l'Office for National Statistics, du United States Census Bureau et du Bureau of Labor Statistics, de Statistics Norway, de l'INSEE (Institut national de la statistique et des études économiques), de Statistics Israel, du Bureau de la statistique de l'Australie, de Statistics New Zealand et d'autres. Ces échanges ont été extrêmement utiles pour comparer les approches à court et à long terme et les mettre en opposition, et étaient presque universellement conformes à l'approche suggérée par Eurostat dans Eurostat (2020). Le Centre a communiqué les conclusions et les recommandations de ces consultations avec les personnes-ressources pertinentes de l'organisme, en mettant à jour périodiquement les données afin de s'assurer que l'information était largement disponible.

Pour ce qui est de la désaisonnalisation des programmes qui sont directement soutenus par le Centre, une stratégie d'assurance de la qualité pour la désaisonnalisation a été élaborée et mise en œuvre, y compris une communication accrue avec les analystes experts en la matière afin de s'assurer que des renseignements importants ont été communiqués aux spécialistes afin de fournir des conseils sur la désaisonnalisation. En outre, la stratégie d'examen annuel des paramètres de désaisonnalisation a été élaborée pour chaque programme en tenant compte du calendrier d'examen, ainsi que de l'ampleur des chocs dans les séries chronologiques. Cette stratégie globale a été documentée et partagée avec d'autres spécialistes de la désaisonnalisation lors d'un tour de table organisé par la section des statistiques gouvernementales de l'American Statistical Association, voir American Statistical Association (2021).

Le Centre de recherche et analyse en séries chronologiques a également tenu de vastes consultations sur l'adaptation saisonnière au sein de l'organisme pour des programmes qui ne reçoivent pas l'appui officiel de l'équipe. Des consultations ont été menées auprès des groupes produisant des statistiques désaisonnalisées sur le commerce de services, le tourisme, le commerce selon les caractéristiques des exportateurs, le nombre d'inscriptions et de sorties dans le Registre des entreprises et divers éléments du système de comptabilité nationale. Des travaux d'exploration ont été effectués aux statistiques désaisonnalisées d'autres programmes. On a notamment analysé les statistiques sur les chargements de wagons et proposé des scénarios pour produire des estimations désaisonnalisées pour certaines séries chronologiques de haut niveau. Une analyse similaire a été effectuée pour la production d'électricité. Dans les deux cas, de nombreuses séries saisonnières ont été déterminées et une décision est en cours quant à savoir si les estimations désaisonnalisées seront produites pour diffusion.

Des progrès continus ont été réalisés afin de mettre le tableau de bord de la désaisonnalisation à la disposition des analystes pour les programmes individuels, alors que deux enquêtes essentielles à la mission peuvent maintenant avoir accès à l'outil afin de comprendre et expliquer les résultats désaisonnalisés. On a également élaboré un plan visant à augmenter ce nombre à quatre autres enquêtes essentielles à la mission au premier trimestre de l'exercice à venir. Le tableau de bord est présenté dans Matthews (2019).

De plus, les méthodes d'estimation de tendances-cycles ont été évaluées dans le contexte de la pandémie de COVID-19. En particulier, compte tenu de la nature aiguë des chocs économiques, les tendances-cycles peuvent présenter une impression trop en douceur de l'économie, de sorte qu'un certain nombre d'autres mesures ont été déterminées, y compris la rupture de la série à un moment donné, et l'introduction d'effets aberrants pour modéliser les chocs plus directement. Ces méthodes seront évaluées plus en détail en tenant compte des avantages et des inconvénients de leur application éventuelle dans certains programmes.

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

On trouvera des détails sur ce sous-projet à la section 1.2 (Estimation en temps réel au moyen de méthodes de séries chronologiques).

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

BIBLIOGRAPHIE

Matthews, S. (2019). De-Mystifying Seasonal Adjustment: A visual tool to understand the process. Presentation to the Seasonal Adjustment Practitioners Workshop, United States Census Bureau.

Eurostat (2020). Guidance On Time Series Treatment In The Context Of The Covid-19 Crisis. https://ec.europa.eu/eurostat/documents/10186/10693286/Time_series_treatment_guidance.pdf.

Ferland, M. (2019). What's new in the Time Series Processing System – v3.08. Internal Document, Statistics Canada.

Statistics Canada, (2020). Excess mortality in Canada during the COVID-19 pandemic. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00076-eng.htm.

Statistics Canada (2021). Workshops, training and references. https://www.statcan.gc.ca/eng/wtc/training.

Statistics Canada (2021b). Experimental indexes of economic activity in the provinces and territories, December 2020. https://www150.statcan.gc.ca/n1/daily-quotidien/210413/dq210413d-eng.htm.

4.8 Confidentialité

Une partie du rôle et de la responsabilité de Statistique Canada continue d'être la sensibilisation et le soutien aux stratégies en matière de confidentialité. Parmi les activités menées tout au long de l'année, mentionnons la consultation avec Emploi et Développement social Canada (EDSC) sur les statistiques sur la transparence salariale, la présentation à l'atelier sur la phase 2 de l'initiative des lacunes statistiques du G20 sur les stratégies d'accès et les ateliers sur la confidentialité pour Santé Canada et EDSC.
D'autres activités de recherche et développement reliées à ce thème sont rapportées dans la section 2.

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

4.9 Communautés de pratique en science des données

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

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

À travers différentes activités reliées à l'apprentissage automatique réunissant de 50 à 80 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 :

Malgré le travail à distance, la Communauté a organisé plusieurs présentations touchant à plusieurs sphères de l'apprentissage automatique comme par exemple, une série de quatre présentations sur les applications d'apprentissage automatique développées par Statistique Canada pour aider les agences de première ligne à évaluer des scénarios de propagation de la COVID-19 et s'y préparer. La Communauté a également continué l'activité 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 mis à jour la liste des projets existants en méthodologie qui explore ou utilise l'apprentissage automatique et continue sa collaboration avec les autres communautés de pratique dont une nouvelle collaboration avec le groupe de travail sur le développement citoyen. Finalement, la communauté continue de relayer de l'information à ses membres à propos de diverses activités externes pertinentes, reliées à la science des données.

SOUS-PROJET : Communauté de pratique (CdP) d'analyse de textes sur l'apprentissage automatique

La CdP d'analyse de textes sur l'apprentissage automatique est un lieu interministériel centralisé qui permet aux praticiens de diverses compétences de discuter des applications pratiques du traitement du langage naturel (TLN). Divers spécialistes du GC se réunissent pour apprendre, discuter et adopter les applications éthiques du TLN. Les réunions mensuelles rassemblent environ 50 à 60 participants pour

partager les solutions et les problèmes de chacun. Environ la moitié des participants proviennent de ministères fédéraux à l'extérieur de Statistique Canada.

Progrès :

Tout au long de 2020-2021, des spécialistes de différents domaines de Statistique Canada ou d'autres ministères comme le Bureau du surintendant des institutions financières, l'Agence des services frontaliers du Canada, Immigration, Réfugiés et Citoyenneté Canada ou l'Agence du revenu du Canada, ont présenté leurs solutions de TLN de haute qualité. Chaque présentateur a illustré sa méthodologie moderne pour traiter rapidement sa source de données, qu'il s'agisse de données d'enquête, de données administratives ou de rapports publics. Les discussions qui ont suivi la présentation ont mis en évidence les concepts complexes que les participants doivent comprendre. Les participants provenaient de plus de 20 ministères différents.

La pandémie de COVID-19 et le passage vers le télétravail ont entraîné un revers inattendu, mais nous en avons profité pour renforcer notre cadre de téléconférence. Ce qui a fait appel à une communauté interministérielle plus inclusive.

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

5 Recherche divisionnaire et autres activités

5.1 Division des méthodes de la statistique économique

SOUS-PROJET : Analyse longitudinale pour le projet « Toking 9 to 5? Examining the Impact of Cannabis Legalization on Workplace Cannabis Use and Perceptions among Canadian Workers »

Le projet *Toking 9 to 5? Examining the Impact of Cannabis Legalization on Workplace Cannabis Use and Perceptions among Canadian Workers* est une étude longitudinale de quatre vagues à cohortes multiples des Instituts de recherche en santé du Canada. Un des principaux objectifs de cette étude est d'examiner l'incidence de la légalisation du cannabis sur les tendances longitudinales de la consommation du cannabis au travail, la perception du risque, les normes et la disponibilité, et si ces tendances au niveau des habitudes diffèrent selon l'âge, le sexe, le rôle des sexes sur le marché du travail, les groupes professionnels et la géographie. L'idée de Carrillo et Karr (2013) est séduisante pour s'attaquer à ce problème puisqu'ils ont proposé une méthode d'analyse marginale des enquêtes à cohortes multiples. Toutefois, le travail de Carrillo et Karr se limitait à des enquêtes *probabilistes,* alors que l'enquête *Toking 9 to 5 n'est pas probabiliste*. Par conséquent, une évaluation ou une modification de ladite méthode pour les enquêtes non probabilistes est nécessaire avant qu'elle puisse être appliquée aux données de l'enquête *Toking 9 to 5*.

Progrès :

Comme il n'y a pas de poids d'enquête pour l'enquête *Toking 9 to 5*, la première étape du projet consiste à créer des pseudopoids, qui peuvent être utilisés avec la méthode de Carrillo et Karr (2013). Nous avons examiné la documentation disponible pour la création de pseudopoids pour les enquêtes non probabilistes (Valliant et Dever, 2011, Wang et coll., 2020a, b). Toutes ces méthodes nécessitent l'utilisation d'une enquête probabiliste comme référence. Nous avons examiné différentes possibilités d'utilisation des enquêtes comme référence. Nous avons déterminé que l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et que l'Enquête canadienne sur le tabac, l'alcool et les drogues (ECTAD) sont les meilleures options; ces enquêtes comportent de nombreuses variables en commun avec les participants du panel *Toking 9 to 5*, et les populations cibles peuvent être rapprochées. Nous avons harmonisé les variables entre les variables du panel *Toking 9 to 5* et celles de l'ESCC pour les deux premières vagues du panel.

SOUS-PROJET : Intégration des données grâce à la méthode LASSO adaptée aux résultats

Les données administratives, ou données d'échantillon non probabilistes, sont de plus en plus utilisées pour obtenir des statistiques officielles en raison de leurs nombreux avantages par rapport aux méthodes d'enquête. En particulier, elles sont moins coûteuses, fournissent une taille d'échantillon plus grande et ne dépendent pas du taux de réponse. Toutefois, il est difficile d'obtenir une estimation sans biais de la moyenne de la population à partir de ces données en raison de l'absence de poids de sondage. En raison du biais de sélection, la moyenne de l'échantillon d'une variable sur un échantillon non probabiliste n'est pas nécessairement une estimation précise de la moyenne de la population pour cette variable. Plusieurs approches d'estimation ont été proposées récemment à l'aide d'un échantillon probabiliste qui fournit des renseignements auxiliaires sur la population cible. Dans cette recherche, nous nous concentrons sur la méthode Chen, Li et Wu (2019). Ces auteurs ont développé une méthode d'estimation doublement robuste pour l'inférence avec des échantillons non probabilistes. L'objectif principal de ce projet est

d'évaluer la possibilité d'étendre l'approche de Chen, Li et Wu (2019) à la sélection de variables à l'aide de la technique LASSO.

Chen, Li et Wu (2019) ont examiné la situation dans laquelle les variables auxiliaires sont données. Toutefois, lorsque l'information sur ces covariables est de grande dimension, la sélection de variables n'est pas une tâche simple, même pour un spécialiste. L'objectif de ce projet était d'étendre l'approche Chen, Li et Wu (2019) en appliquant une méthode LASSO adaptative avec une pénalité (Shortreed et Ertefaie, 2017) pour effectuer une sélection de variables. Ce travail est pertinent pour plusieurs raisons. Des études ont montré que l'inclusion des variables dans le modèle de score de propension, qui influence la sélection dans l'échantillon non probabiliste et qui explique également la variable cible (Van der weele et Shiptser, 2011), conduit à des estimations non biaisées. Toutefois, l'inclusion de variables au modèle de score de propension qui influent sur la sélection dans l'échantillon non probabiliste mais non sur la variable cible, augmentera la variance des estimateurs par rapport aux estimateurs qui excluent de telles variables (Schistermann et coll., 2009; Schneeweiss et coll., 2009; Van der Laan et Gruber, 2010). Dans le même sens, l'inclusion de variables dans le modèle de score de propension, qui ne sont reliées qu'à la variable cible augmentera la précision de l'estimateur sans affecter le biais (Brookhart et coll., 2006); Shortreed & Ertefaie, 2017).

Progrès :

Nous avons élaboré une extension de la méthode LASSO adaptative dans le contexte de la combinaison d'échantillons non probabiliste et probabiliste (Bahamyirou, 2021). Nous avons également conçu un programme R, qui met en œuvre la méthode proposée ainsi que l'estimateur de Chen, Li et Wu (2019). Les résultats de ce projet ont été soumis à *La revue canadienne de statistique*.

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

BIBLIOGRAPHIE

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. et Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149–-1156.

Carrillo I.A., et Karr A.F. (2013). Combining cohorts in longitudinal surveys. *Survey methodology*, 39(1), 149-182.

Chen, Y., Li, P. et Wu, C. (2019). Doubly robust inference with Non-probability Survey samples. *Journal of the American Statistical Association* (published online).

Gruber, S., et van der Laan, M.J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *International Journal of Biostatistics*, 6(1), 18.

Schisterman, E.F, Cole, S. et Platt, R.W (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20, 488.

Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H. et Brookhart, M.A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20, 512.

Shortreed, S.M., et Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4), 1111–1122.

Valliant, R., et Dever, J.A. (2011). Estimating propensity adjustments for survey volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-137.

Van der weele, T.J., et Shipitser, I. (2011). A new criterion for confounder selection, Biometrics. 2011 Dec; 67(4): 1406–1413.

Wang L., Graubard B.I., Katki H.A. et Li Y. (2020). Improving external validity of epidemiological cohort analyses: a kernel weighting approach. *JRSS, A*, 183(3), 1293-1311.

Wang L., Graubard B.I., Katki H.A. et Li Y. (2020). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. https://arxiv.org/abs/2011.14850v1

5.2 Division des méthodes de statistique sociale

SOUS-PROJET : Intervalles de confiance pour des totaux estimés

L'objectif de la recherche était d'élaborer une méthode d'intervalle de confiance pour des totaux estimés ayant de bonnes propriétés de couverture pour la diffusion du questionnaire détaillé du Recensement de 2021. La plupart des estimations du questionnaire détaillé du recensement sont des estimations du nombre total d'unités qui présentent une caractéristique d'intérêt. La méthode habituelle de construction des intervalles de confiance a une couverture médiocre pour des totaux estimés quand la taille de l'échantillon est petite et que presque toutes les unités échantillonnées ont une valeur de 0 pour la variable d'intérêt ou que presque toutes les unités ont une valeur de 1.

Progrès :

On a calculé un intervalle de confiance Wilson pour des totaux estimés en imitant l'approche utilisée par Wilson (1927) pour les proportions et en l'adaptant aux données d'enquête complexes à l'aide de la méthode proposée par Kott et Carr (1997). Des études approfondies de simulation ont été entreprises pour déterminer comment l'intervalle proposé devrait être mis en œuvre pour l'estimation sur des domaines et le calage. La méthode proposée a également été mise à l'essai au moyen de simulations conçues pour reproduire l'environnement du questionnaire détaillé du recensement. Une approximation de l'intervalle Wilson modifié a été élaborée pour contourner les contraintes imposées par le Système de tabulation généralisée (GTAB), qui sera utilisé pour la production du questionnaire détaillé du recensement. Les résultats de simulation montrent que l'intervalle Wilson modifié pour des totaux estimés a de bonnes propriétés de couverture pour presque tous les scénarios envisagés. Les travaux sont documentés dans Hidiroglou (2020), Neusy (2021) et Savard (2020, 2021).

SOUS-PROJET : Combiner les données d'enquête avec les données recueillies selon une approche participative à l'aide de techniques d'estimation sur petits domaines

L'approche participative est une méthode de collecte à faible coût qui permet de recueillir des renseignements rapidement au moyen d'un questionnaire qui est ouvert à tout participant désireux de participer. Toutefois, les données obtenues par l'approche participative ne peuvent pas être utilisées directement pour tirer des inférences statistiques, car le processus n'est pas probabiliste. L'utilisation de méthodes d'estimation sur petits domaines (EPD) pour combiner les données recueillies selon une approche participative avec des données probabilistes a été mise à l'essai dans le but de réduire la variance.

Progrès :

Le modèle Fay-Herriot a été utilisé pour intégrer les données catégoriques sur la COVID-19 recueillies selon une approche participative et les sources de données d'enquête sur la COVID-19 provenant des premier et troisième cycles de la Série d'enquêtes sur les perspectives canadiennes (SEPC). Le modèle de Fay-Herriot a été mis à l'essai pour sa simplicité et son interprétabilité. Dans l'ensemble, le modèle de Fay-Herriot a montré un succès modéré dans l'intégration de données catégoriques recueillies selon une approche participative avec des données d'enquête. Lorsqu'un ajustement approprié du modèle a été obtenu, des améliorations significatives ont été observées au chapitre de la précision des estimations. Cela s'est produit même pour le troisième cycle de l'approche participative, caractérisé par une réduction spectaculaire du nombre de participants. Toutefois, un ajustement approprié du modèle n'a pas toujours été réalisable, ce qui pourrait indiquer la nécessité d'évaluer l'utilisation d'autres méthodes d'EPD.

Un rapport de recherche et un document de travail de la Division des méthodes de la statistique sociale portant sur ce projet ont été rédigés (Chatrchi et Ding, 2021).

SOUS-PROJET : Intégration des données d'enquête et des données recueillies selon une approche participative : Méthode d'appariement d'échantillons

L'appariement d'échantillons est une méthode relativement nouvelle proposée par Rivers (2007), qui consiste à utiliser des données non probabilistes comme bassin de donneurs et à imputer toutes les réponses pour un échantillon probabiliste à l'aide d'un ensemble de variables d'appariement. Cette méthode a été mise à l'essai dans le contexte des enquêtes sur la COVID-19, où les estimations imputées ont été comparées aux estimations provenant d'enquêtes probabilistes.

Progrès :

Tout d'abord, la méthode d'appariement d'échantillons a été mise à l'essai, comme le proposait Rivers : toutes les unités d'un échantillon probabiliste ont été imputées à l'aide de donneurs provenant d'une source de données non probabilistes. L'échantillon était constitué de répondants à une enquête utilisant les mêmes questions que les données non probabilistes. Ainsi, les estimations de l'enquête probabiliste et de l'expérience d'appariement d'échantillons ont été comparées. Ensuite, on a utilisé la méthode d'appariement d'échantillons pour imputer les non-répondants d'une enquête non probabiliste à l'aide de donneurs provenant d'une source de données non probabilistes. Les estimations de cette méthode et de l'enquête probabiliste ont été comparées.

Les deux étapes ont été réalisées à l'aide de la première vague de la Série d'enquêtes sur les perspectives canadiennes (SEPC), une enquête probabiliste, et de la première vague des projets par approche participative comme source de données non probabiliste. Puis, elles ont été effectuées de nouveau à l'aide de la troisième vague de la SEPC comme enquête probabiliste, et de la composante d'approche participative de la troisième vague de la SEPC comme source de données non probabiliste.

Le rapport de recherche est terminé et un document de travail de la Division des méthodes de la statistique sociale sur ce projet est en cours d'examen (Poirier, Chatrchi et Wang, 2021).

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

BIBLIOGRAPHIE

Kott, P.S., et Carr, D.A. (1997). Developing an Estimation Strategy for a Pesticide Data Program. *Journal of Official Statistics*, 13, 4, 367-383

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Wilson, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22, 209-212.

5.3 Division des méthodes d'intégration statistique

SOUS-PROJET: Outrepasser des erreurs dans les données utilisées en couplage d'enregistrement avec la conception et l'étude d'une nouvelle méthode de couplage par génération aléatoire de pseudo-clés et de processus de décision.

L'utilisation de pseudo-clés est courante lors des projets de couplage. Une pseudo-clé est définie à partir de portions de variables des deux tables à coupler de sorte qu'elle forme une nouvelle variable qui va servir de clé de couplage. Elle simule le rôle d'une vraie clé (comme le Numéro d'Assurance Sociale (N.A.S.)) en imposant l'unicité dans une même table, et formera des paires à partir des enregistrements qui ont la même valeur de pseudo-clé. En plus d'être utilisée en couplage déterministe, elle est aussi importante en couplage probabiliste où elles sont appliquées pour contribuer à former un ensemble de paires liées à l'étape du « blocking ».

Ce projet développe une méthode de couplage fondée sur la génération des pseudo-clés aléatoires par échantillonnage des variables et sur l'agrégation des résultats de classification des paires obtenues par l'application sur les tables des pseudo-clés générées.

Le premier but était d'effectuer de nouveaux développements ou fonctionnalités de la méthode en langage python. Un autre objectif était de tester l'efficacité de la méthode développée dans le contexte d'un appariement à 2 tables, selon différents cas de figure. Le dernier objectif était de commencer à concevoir un cadre théorique aux pseudo-clés et leurs indicateurs d'unicité.

Progrès :

Des développements de la méthode ont été effectués pour produire une panoplie d'indicateurs, référencés par Christen (2007), et d'outils de visualisation de la qualité du couplage à partir d'une table de vérité. Ces fonctionnalités ont permis de réaliser le deuxième objectif en faisant l'étude de l'impact de la variation du degré d'inclusion/exclusion des deux tables d'entrées sur les résultats de couplage. De même, une analyse de l'influence du paramètre seuil minimal pour un vote entre classifieurs a été produite grâce notamment à l'outil interactif de la Courbe « Receiver Operating Characteristic » (ROC) qui présentent ces résultats en comparaison avec des pseudo-clés individuelles. Enfin, le développement théorique a débuté avec la formalisation de notions permettant une définition mathématique des « pseudo-clés » en couplage, similaire à celles présentées dans Atencia, David & Euzenat (2014) et Pernelle, Saïs, Symeonidou (2013), ainsi que du caractère aléatoire utilisé dans cette méthode. Cette définition permet d'aboutir à une probabilité d'unicité théorique naïve, supposant l'indépendance des caractères et une distribution i.i.d des enregistrements, qui a été comparée expérimentalement avec les taux d'unicité réels des pseudo-clés.

Pour plus de renseignements, communiquez avec :
Gautier Gissler (613-294-2574, gautier.gissler@statcan.gc.ca).

BIBLIOGRAPHIE

Atencia, M., David, J. et Euzenat, J. (2014). Data interlinking through linkkey extraction. ECAI 2014: 15-20.

Christen, P., et Goiser, K. (2007). Quality and Complexity Measures for Data Linkage and Deduplication. Quality Measures in Data Mining, Studies in Computational Intelligence, vol. 43, Springer.

Pernelle, N., Saïs, F. et Symeonidou, D. (2013). An automatic key discovery approach for data linking. *Journal of Web Semantics*, Elsevier, 23, 16-30.

5.4 Centre de collaboration internationale et d'innovation en méthodologie

SOUS-PROJET : Suivi des non répondants pour les enquêtes auprès des entreprises

Au cours des deux dernières décennies, les taux de réponse aux enquêtes ont diminué régulièrement. Dans ce contexte, il est devenu de plus en plus important pour les organismes de statistique d'élaborer et d'utiliser des méthodes qui réduisent les effets néfastes de la non-réponse sur l'exactitude des estimations d'enquêtes. Le suivi des non-répondants peut être un remède efficace contre le biais de non-réponse, même s'il exige beaucoup de temps et de ressources. L'objectif de ce projet de recherche est de faire la lumière sur certaines questions pratiques concernant le suivi de la non-réponse. Par exemple, si l'on suppose un budget fixe de suivi, quelle est la meilleure façon de choisir les unités devant faire l'objet d'un suivi? Devrait-on faire le suivi de tous les non-répondants ou seulement d'un échantillon d'entre eux? Si l'on fait un suivi d'un échantillon, de quelle manière devrait-il être sélectionné?

Progrès :

Une étude de simulation a déjà été réalisée pour répondre aux questions susmentionnées. Pendant l'année en cours, nous avons ajouté des résultats théoriques pour mieux justifier nos résultats empiriques, et nous avons terminé un article qui a été soumis à une revue statistique à comité de lecture (Neusy, Beaumont, Yung, Hidiroglou et Haziza, 2021).

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

Dans les enquêtes 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 par les poids de sondage des paramètres de population finie. Le biais conditionnel est une mesure attrayante de l'influence qui tient compte du plan de sondage et de la méthode d'estimation. Il se définit comme l'espérance par rapport au plan de sondage de l'erreur d'échantillonnage conditionnelle à la sélection d'une unité donnée dans l'échantillon. L'estimation de ce biais est relativement simple pour les plans de sondage et les estimateurs simples. Pour les plans ou les estimateurs complexes en revanche, il peut être fastidieux de dégager une expression explicite du biais conditionnel. Dans de tels cas, l'estimation de la variance s'obtient fréquemment par des méthodes par répliques comme le bootstrap. Les méthodes bootstrap sont normalement mises en œuvre en produisant un jeu de poids bootstrap qui est mis à la disposition des utilisateurs de données d'enquête. Nous étudions la façon d'utiliser les poids bootstrap pour obtenir un estimateur du biais conditionnel. Cet estimateur peut servir à construire des estimateurs robustes des paramètres de population finie, qui sont moins négativement touchés par les unités influentes que les estimateurs pondérés ordinaires. Nous prévoyons d'évaluer dans une étude par simulation notre estimateur bootstrap du biais conditionnel.

Progrès :

La théorie a été élaborée dans des recherches antérieures. Cette année, nous avons mené une étude par simulation et rédigé un article (Beaumont, Bocci et Saint-Louis, 2021), qui a été présenté à une revue statistique à comité de lecture.

SOUS-PROJET : Développement d'un système prototype pour l'estimation robuste

Dans maintes enquêtes économiques et quelques enquêtes sociales, des variables asymétriques sont recueillies, ce qui peut créer la présence de valeurs aberrantes et d'unités influentes. Les méthodes classiques d'estimation peuvent produire des estimateurs hautement inefficaces dans ce scénario. L'idée avec une estimation robuste est de diminuer l'effet de ces unités de l'échantillon sur les estimations. Le biais conditionnel est utilisé comme une mesure d'influence. 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.

Progrès :

Un prototype élaboré en SAS comporte un grand nombre des spécifications pour l'estimation robuste formulées dans Estevao (2021). Il s'agit d'une série de macros pour les diverses fonctions liées à la production d'estimations robustes pour des domaines. Les estimations robustes au niveau des domaines n'ont généralement pas la propriété d'additivité des estimations standards pour les domaines. Il y a donc

une macro qui crée des estimations cohérentes pour des domaines 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 pour les domaines et de tous les totaux connus de variables auxiliaires.

Au cours de la dernière année, une méthode bootstrap pour l'estimation de la variance a été incluse pour produire des estimations de la variance pour les estimations robustes pour les domaines. Un guide d'utilisation avec des exemples (Estevao, 2021) est disponible décrivant les 6 premières des 9 fonctions principales. Il sera mis à jour prochainement pour inclure une description de toutes les fonctions.

Les macros compilées et le guide d'utilisation pour leur application sont disponibles grâce au soutien des membres du Centre de collaboration internationale et d'innovation en méthodologie.

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 la variance dans des enquêtes avec 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 est valide pour 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 :

Au cours de l'année précédente, 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 tiennent compte de la non-réponse. Au cours de l'année en cours, nous avons commencé des études par simulation pour évaluer la méthode bootstrap proposée et nous avons rédigé la partie théorique d'un article, que nous prévoyons présenter à une revue statistique à comité de lecture.

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

BIBLIOGRAPHIE

Beaumont, J.-F., 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). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

5.5 Division de la science des données

SOUS-PROJET : Travaux de la Division de la science des données

En septembre 2019, Statistique Canada a créé la Division de la science des données afin de fournir une expertise en science des données et d'agir comme porte d'entrée pour le traitement avancé des données et l'analyse des mégadonnées non structurées. Son équipe multidisciplinaire (scientifiques des données, méthodologistes, gestionnaires), composée de spécialistes dans les plus récentes techniques de source ouverte, de codage, de matériel et de services infonuagiques a été créée pour s'attaquer aux projets portant sur les données structurées et non structurées (textes, images, etc.). L'équipe fait également la promotion de pratiques et de méthodes scientifiques éthiques qui produisent une inférence statistique valide. Son énoncé de mission est de renforcer les capacités en science des données au sein de l'organisme en résolvant des problèmes concrets.

Progrès :

Au cours de la dernière année, l'équipe a mené de nombreux projets de science des données en utilisant différentes techniques de science des données : lecture et extraction de données des documents en format PDF, utilisation de l'imagerie satellite pour dériver des données de l'agriculture, exploration des techniques de protection de la vie privée, extraction de données à partir de sources de nouvelles, de commentaires ou de textes narratifs, modélisation de scénarios relatifs à la COVID-19 pour aider les organismes de première ligne, entre autres. Nous avons également mené des activités horizontales, comme le lancement en octobre 2020, du Réseau de la science des données pour la fonction publique fédérale, qui compte maintenant près de 2 000 membres et publie un bulletin mensuel. Le Réseau est ouvert à toute personne qui s'intéresse à la science des données au Canada, qu'il s'agisse de scientifiques des données ou de gestionnaires de scientifiques des données. Les cinq caractéristiques du Réseau sont les suivantes : gestion des talents, formation, transfert d'information, collaboration et services communs.

La Division a également travaillé à une première version de la stratégie relative à la science des données à venir de Statistique Canada, fondée sur les six piliers suivants : gouvernance, opérationnalisation, infrastructure et outils, intégration, personnes et culture et leadership. La vision est la suivante : l'organisme intègre des méthodes d'apprentissage automatique et d'intelligence artificielle, de nouveaux processus, de nouvelles technologies et normes, avec une gestion des données analytiques de longue date et une expertise en matière de protection de la vie privée, afin de fournir de meilleures perspectives sociales et économiques aux Canadiens et aux décideurs. Enfin, la Division de la science des données contribue à accroître la capacité de l'organisme en matière de science des données en dirigeant ou en appuyant les communautés de pratique connexes, en contribuant à l'initiative de littératie des données de Statistique Canada ou en proposant une formation appropriée en science des données à tous les niveaux d'employés au moyen de cours, de séminaires et de forums en ligne et officiels.

Pour de plus amples renseignements, veuillez consulter :

La page web du Réseau de la science des données pour la fonction publique fédérale :
https://www.statcan.gc.ca/fra/science-donnees/reseau
Formation de la littératie des données sur l'apprentissage automatique :
https://www.statcan.gc.ca/fra/afc/litteratie-donnees/catalogue/892000062021004

Ou communiquez avec :
Sevgui Erman (613-716-5906, sevgui.erman@statcan.gc.ca).

5.6 Revue Techniques d'enquête

Techniques d'enquête est une revue internationale accessible à www.statcan.gc.ca/techniquesdenquête qui publie 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 examinateurs, les rédacteurs associés et les lecteurs; la revue des commentaires des examinateurs et des révisions des auteurs; la mise en page et l'édition des manuscrits; la liaison avec la traduction et la diffusion; et la mise à jour d'une base de données des articles soumis.

Progrès :

Les numéros de juin et décembre 2020 (46-1 et 46-2) ont été publiés en versions PDF et HTML. Ces numéros comprennent respectivement 6 et 5 articles réguliers. D'avril 2020 à mars 2021, les pages de Techniques d'enquête ont été consultées 47 000 fois et près de 18 000 copies d'articles ont été téléchargées. Cinquante-sept articles ont été soumis pour publication.

En janvier 2021, Jean-François Beaumont, conseiller principal en statistique à Statistique Canada, a été nommé comme nouveau rédacteur en chef de Techniques d'enquête. Beaumont est associé à la revue depuis 20 ans; il a été rédacteur adjoint de 2000 à 2010 et rédacteur associé de 2010 à 2020. Il succède à Wesley Yung, rédacteur en chef depuis 2015. Respecter les attentes accrues des auteurs en terme de rapidité du processus éditorial, tout en conservant la même valeur fondamentale de rigueur scientifique, constitue une priorité du nouveau rédacteur en chef.

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

5.7 Transfert de connaissances – Formation statistique

Le Groupe de travail sur le développement de talent statistique, dont le mandat principal demeure la modernisation de la formation statistique au sein de l'organisme, a connu une année chargée et productive. Même s'il existe un programme d'études bien établi comportant des cours qui couvrent un éventail de thèmes statistiques, le Groupe continue de développer et d'établir la priorité de nouvelles

activités d'apprentissage qui peuvent être développées en temps opportun, en mettant l'accent sur l'apprentissage actif.

Au début de l'année, lorsque la pandémie a frappé et que bon nombre des programmes de Statistique Canada ont été temporairement suspendus, les employés ont dû s'adapter rapidement au travail à temps plein à domicile. Pour de nombreux employés, cette période s'est révélée idéale pour se concentrer sur la formation. Par conséquent, un vaste inventaire d'outils d'autoapprentissage disponibles en ligne a été rapidement élaboré et partagé avec les employés de la Direction. Les sujets abordés étaient variés, y compris les techniques d'échantillonnage, l'estimation de la variance, la science des données et l'apprentissage automatique, l'intégration des données, la modélisation et divers logiciels statistiques (R, SAS, Python) pour n'en nommer que quelques-uns. Cet inventaire continue d'être une excellente ressource de formation pour les employés à ce jour.

Étant donné que la science des données et la modélisation des données continuent d'être des domaines prioritaires au sein de l'organisme, les nouvelles initiatives sont restées principalement axées sur ces sujets. Deux cours sur la science des données ont été proposés par des professeurs d'université connus. Le premier était une reprise du cours mis à l'essai l'année précédente, qui fait maintenant partie de notre programme. Le deuxième ciblait les gestionnaires de l'organisme, et se concentrait sur ce que les méthodes peuvent et ne peuvent pas faire plutôt que sur la théorie mathématique derrière les méthodes. Il a suscité un grand intérêt et devrait faire partie de notre programme régulier pendant les années à venir. De plus, une nouvelle activité d'apprentissage sur la modélisation statistique a été mise à l'essai pour la première fois l'an dernier. Trois sujets ont été abordés, à savoir la régression linéaire, la régression logistique et la validation croisée. Les participants ont d'abord reçu la directive de regarder une vidéo expliquant la technique de modélisation. Puis, le groupe a eu une discussion approfondie avec un modérateur, et enfin il y a eu une démonstration de l'application de la méthode dans un programme existant de Statistique Canada. Cette combinaison s'est révélée très réussie et a vraiment permis une participation active de tous les participants. L'expérience sera répétée l'année prochaine. De plus, à mesure que l'utilisation de R devient plus répandue au sein de l'organisme, des séances supplémentaires du cours d'introduction interne élaboré l'an dernier ont été offertes et continueront de l'être.

Une autre nouvelle initiative réussie au cours de la dernière année a été le lancement de faits intéressants sur la méthodologie. Il s'agit de courtes vidéos, de 30 à 40 minutes, élaborées par des employés de Statistique Canada et portant sur un large éventail de sujets. D'autres sont en développement pour l'année prochaine. Enfin, il convient de noter que la plupart des cours en classe existants dans notre programme d'études ont connu une transition sans heurts vers un format virtuel.

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

5.8 Symposium international de 2021 sur les questions de méthodologie de Statistique Canada

Le Symposium international de 2021 sur les questions de méthodologie de Statistique Canada « Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société » se déroulera, pour la première fois, de façon virtuelle durant plusieurs semaines à l'automne 2021. Habituellement tenu dans la région de la capitale nationale du Canada, le Symposium de 2021 sera

accessible en ligne tous les vendredis du 15 octobre au 5 novembre 2021 inclusivement. Le symposium sera gratuit, il comprendra des séances plénières, des séances simultanées et des séances de présentation par affiche qui porteront sur une grande variété de sujets, et sera précédé de trois ateliers le jeudi 14 octobre.

Progrès :

Le comité organisateur a été très actif dans l'élaboration des différentes activités entourant le symposium. Comme il a lieu virtuellement, un nouveau format a été élaboré, des plates-formes possibles ont été évaluées et un formulaire de soumission de résumé a été mis en œuvre en ligne. Le comité a été occupé à organiser les activités habituelles, comme la publicité pour le symposium, l'élaboration des thèmes et l'annonce de demandes de propositions de communication, l'organisation de séances sollicitées, et l'obtention de conférenciers principaux et de présentateurs spéciaux. Le comité a également amorcé l'organisation des ateliers, ainsi que le groupe d'experts – déterminer les thèmes et les participants potentiels. Après la clôture des soumissions, nous élaborerons le programme complet, terminerons la sélection de la plate-forme, obtiendrons des diapositives des présentateurs, traduirons du matériel, élaborerons un environnement de test et tiendrons le symposium. Puis, les dernières étapes consisteront à recueillir des documents pour les délibérations, puis à coordonner et réviser leur traduction aux fins de publication électronique.

Pour de plus amples renseignements, veuillez consulter :
https://www.statcan.gc.ca/fra/conferences/symposium2021/index.

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

American Statistical Association (2021). Time Series and Seasonal Adjustment Estimation During the COVID-19 Pandemic. Discussion de type table ronde, organisée par la *Government Statistics Section*, https://community.amstat.org/governmentstatisticssection/professionaldevelopmentmentoring/virtualworkshoppracticumfall2020349.

Arim, R., Hennessey, D. et Molladavoudi, S. (2021). Data, data analytics and data science at Statistics Canada during Covid-19 pandemic. Présenté à la Nexus 2021 Data Science Conference, Université du Manitoba.

Bahamyirou, A. (2021). Data integration through outcome adaptive LASSO and a collaborative propensity score. Document interne, Statistique Canada.

Beaulieu, M., et Lebrasseur, D. (2021). « Measuring and Communicating Quality for Programs Using Administrative Data Sources Exclusively ». Document en préparation pour *European Establishment Statistics Workshop*.

Beaumont, J.-F., Bocci, C. et St-Louis, M. (2021). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. Soumis pour publication dans une revue en statistique avec comité de lecture.

Beaumont, J.-F., et Chu, K. (2020). Statistical data integration through classification trees. Article présenté au Comité consultatif sur les méthodes statistiques, juin 2020, Statistique Canada.

Beaumont, J.-F., et Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Blais, A.-R., Mach, L., Michaud, I. et Simard, J.-F. (2020). Analysis of the Public Service Employee Survey Items as Measures of the Psychosocial Risk Factors. Présentation au Comité directeur sur les mesures de performance en santé mentale en milieu de travail, 7 octobre, 2020.

Blais, A.-R., Michaud, I., Simard, J.-F., Mach, L. et Houle, S. (2021). Measuring Workplace Psychosocial Factors in the Federal Government. Soumis pour publication dans *Rapports sur la santé*.

Bocci, C., Morissette, R. et Beaumont, J.-F. (2020). Overview of small area estimation methods used for the estimation of mean liquid assets. Rapport interne, Statistique Canada.

Bosa, K., et Chu, K. (2020). Ottawa COVID-19 Hospital Occupancy Forecasts – Final Report. Rapport interne (partagé avec les collaborateurs de Santé Canada), Statistique Canada.

Centre de ressources en analyse de données (2020a). Data Visualization - Best practices. Document interne, Direction des méthodes statistiques modernes et de la science des donnees, Statistique Canada.

Centre de ressources en analyse de données (2020b). Measuring Culture of Accessibility, A Brief Introduction to Factor Analysis. Présentation au groupe de travail sur l'accessibilité, 1 mai, 2020.

Chatrchi, G., et Ding, A. (2021), Combining survey data with crowdsourced data using small area estimation techniques, Document de travail de la Division des méthodes de la statistique sociale, SSMD-2021-02E, Statistique Canada.

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

Dasylva, A., et Goussanou, A. (2021). Estimation des faux négatifs causes par le blockage. *Techniques d'enquête* (à paraître dans le numéro de décembre 2021).

Denis, N., El-Hajj, A., Drummond, B., Abiza, Y. et Gopaluni, K.C. (2021). Learning COVID-19 Mitigation Strategies using Reinforcement learning. Chapitre à paraître dans un livre publié chez Springer.

Estevao, V.M. (2021). Robust Estimation - Parameter Description and User Guide, Methodology Specifications. Document interne, Statistique Canada.

Gilchrist, P. (2020). Public Use Microdata Files and Business Data: A Summary of Challenges for Confidentiality. Document interne, Statistique Canada.

Gray, D. (2020a). Evaluation of some imputation methods for monthly industry surveys using the Evaluation Framework and Visualisation. Document interne présenté au Comité de revue scientifique de la Direction des méthodes statistiques modernes et de la science des données, Statistique Canada.

Gray, D. (2020b). Evaluating Imputation Methods using ImpACT: First Case Study. Presented at the *United Nations Statistical Commission and Economic Commission for Europe – Workshop on Statistical Data Editing.* https://unece.org/statistics/events/SDE2020.

Hidiroglou, M. A. (2020). Wilson Intervals for Proportions and Counts. Document interne, Statistique Canada.

Lesage, É., Beaumont, J.-F. et Bocci, C. (2021). Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot. *Technique d'enquêtes*, 47 (à paraître).

Mach, L., et Michaud, I. (2020). Statistical Modelling with Complex Survey Data – Linear Regression. Session 1 of *Statistical Modelling Course*, Groupe de travail sur le développement de talent en statistique, Direction des méthodes statistiques modernes et de la science des données, Statistique Canada.

Matthews, S. (2020a). Over-arching principals for the Statistical Generalized Systems. Document interne, présenté au Comité directeur des systèmes généralisés, Statistique Canada.

Matthews, S. (2020b). Strategy for Development of Advance Indicator. Document interne présenté au Comité des méthodes et des outils modernes — conversion des données en renseignements, Statistique Canada.

Matthews, S. (2021a). Updated Generalized Systems Evolution Plan. Document interne, présenté au Comité directeur des systèmes généralisés, Statistique Canada.

Matthews, S. (2021b). Guidelines for Advance Indicators of Official Statistics. Document interne présenté au Comité des méthodes et des outils modernes — conversion des données en renseignements, Statistique Canada.

Matthews, S., Patak, Z. et Picard, F. (2020). Time Series Modelling to Produce Economic Indicators in (near) Real-time. Article présenté au Comité consultatif sur les méthodes statistiques, Statistique Canada.

Matthews, S., et Ritter, C. (2020). Building a better nowcast – towards a real-time modelling environment. Présentation au Comité de recherché et développement, Statistique Canada.

Neusy, E., Beaumont, J.-F., Yung, W., Hidiroglou, M. et Haziza, D. (2021). Non-response follow-up for business surveys. Soumis pour publication dans une revue avec comité de lecture.

Neusy, E. (2021). Wilson Confidence Intervals for Proportions and Totals of Binary Variables Using Complex Survey Data. Document interne, Statistique Canada.

Picard, F. (2020). SARIMA models for early estimates of energy balance statistics. Article présenté au United Nations Economic Commission for Europe – High level group on Modernizing Official Statistics.

Picard, F. (2021). A comparison between PROC BENCHMARKING and the R package TEMPDISAGG. Document interne, Statistique Canada.

Poirier, G., Chatrchi, G. et Wang, Y. (2021) Integrating survey data and crowdsourced data: Sample Matching method, Document de travail, Division des méthodes de statistique sociale (revue en cours), Statistique Canada.

Reicker, K. (2020). Perceptions of data sensitivity, pilot study with questionnaire testing participants. Document interne, Statistique Canada.

Renaud, M., et Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: the Statistics Canada experience. Article présenté au Comité consultatif sur les méthodes statistiques, Statistique Canada.

Sallier, K. (2020). Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis. *Statistical Journal of the International Association for Official Statistics*, 36, 1059-1066.

Sallier, K. (2021). Statistics Canada's experience creating public synthetic datasets using the FCS and the Synthpop Package. Article présenté au United Nations Economic Commission for Europe – High level group on Modernizing Official Statistics.

Savard, S.-A. (2020). Intervalles de confiance pour les comptes estimés au questionnaire long du recensement. Document interne présenté au Comité de revue scientifique de la Direction des méthodes statistiques modernes et de la science des données, Statistique Canada.

Savard, S.-A. (2021). Intervalle de confiance pour les comptes estimés. Rapport interne, Statistique Canada.

Statistique Canada (2020). G-EST 2.03.002 User Guide. Document interne, Statistique Canada.

Statistique Canada (2021). BANFF version 2.08 User Guide. Document interne, Statistique Canada.

You, Y. (2021a). Evaluation of robust small area estimation using normal and t distribution models. Document interne du CCIIM, Statistique Canada.

You, Y. (2021b). Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling. *Techniques d'enquête*, 47 (à paraître dans le numéro de décembre 2021).

You, Y. (2021c). A Bayesian approach to regression estimation and prediction inference for survey data analysis. Document interne du CCIIM, Statistique Canada.

Zanussi, Z. Santos, B. et Molladavoudi, S. (2021), Supervised text classification with leveled homomorphic encryption. À paraître dans : Recueil du 63rd ISI World Statistics Congress, Virtual.

Zhao, Z. (2021). Exploring available tools to generate synthetic data with high analytical value: R packages synthpop and simPop. Rapport d'étudiant co-op, dirigé par H. Gauvin. Document interne, Statistique Canada.