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

# Programme de recherche et développement en méthodologie

## Rapport annuel 2018-2019

Date de diffusion : le 15 novembre 2019

Statistique Canada Statistics Canada

Canada

## Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

**Courriel à** STATCAN.infostats-infostats.STATCAN@canada.ca

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

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

**Programme des services de dépôt**

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

## Normes de service à la clientèle

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

## Note de reconnaissance

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

Le présent rapport fait la synthèse des réalisations de 2018-2019 du Programme de recherche et développement en méthodologie (PRDM), parrainé par la Direction de la méthodologie de Statistique Canada. Ce programme englobe les activités de recherche et de développement qui ont trait à des méthodes statistiques susceptibles d'être appliquées à grande échelle aux programmes d'enquête de l'organisme; ce sont des activités qui, autrement, ne seraient pas entreprises dans le cadre des services de méthodologie offerts à ces programmes d'enquête. En outre, dans le but de promouvoir l'utilisation des résultats des travaux de recherche et de développement, le PRDM comprend des activités de soutien aux clients pour la mise en application de travaux de développement antérieurs fructueux. Des renseignements supplémentaires sur les projets décrits peuvent être obtenus auprès des personnes-ressources mentionnées. Pour en savoir davantage sur le PRDM dans son ensemble, communiquez avec :

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

# Programme de Recherche et Développement en Méthodologie

## Rapport annuel 2018-2019

## Table des matières

# 1 Projets de recherche

## 1.1 Recherche-développement – Estimation sur petits domaines

Les estimations classiques fondées sur les 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, on a observé à Statistique Canada un regain d'intérêt pour l'étude de méthodes d'estimation indirecte sur petits domaines reposant sur un modèle. Le système et la méthode sont documentés dans Hidiroglou, Beaumont et Yung (2018). Le but ultime est d'utiliser ces méthodes pour la production de statistiques officielles, lorsque cela est jugé approprié. Les principaux objectifs de ce projet sont les suivants :

i) élaborer de nouvelles méthodes d'estimation sur petits domaines qui tiennent compte des problèmes observés dans les enquêtes réelles;
ii) étudier les propriétés des méthodes existantes selon différents scénarios afin de mieux comprendre comment et quand utiliser ces dernières;
iii) déterminer une méthodologie d'estimation sur petits domaines appropriée pour certaines enquêtes candidates;
iv) mettre au point et tester des prototypes mettant en œuvre des méthodes nouvelles ou existantes susceptibles d'être utiles dans le cadre des programmes statistiques.

Des progrès ont été réalisés dans le cadre des sous-projets suivants. En voici une description.

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

Les outils de validation des 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'efficience générés par les estimateurs sur petits domaines par rapport aux estimateurs directs. Toutes ces techniques sont utiles pour évaluer le rendement global associé aux 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 (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 par rapport au

plan de sondage de l'EQM fondée sur le plan de sondage. L'objectif de ce projet est d'élaborer et d'étudier de nouveaux diagnostics locaux pour l'évaluation des estimations sur petits domaines.

**Progrès :**

Tout d'abord, nous avons calculé un intervalle sur l'effet aléatoire local afin de nous assurer que l'EQM fondée sur le plan de sondage pour le meilleur prédicteur est inférieure à celle rattachée à l'estimateur direct. Notre premier diagnostic évalue la probabilité conditionnelle que l'effet aléatoire local se situe à l'intérieur de l'intervalle. Si la probabilité est trop faible, il peut être préférable de choisir l'estimateur direct plutôt que l'estimateur sur petits domaines. Le deuxième diagnostic est la valeur P d'un test d'hypothèse fondé sur le plan de sondage selon lequel l'effet aléatoire local n'est pas supérieur à la limite de l'intervalle. Nous avons élaboré le fondement théorique et mené des études empiriques. Les résultats préliminaires ont été présentés lors du Colloque francophone sur les sondages en octobre 2018. Un article est en cours de rédaction (Lesage, Beaumont et Bocci, 2019).

**SOUS-PROJET :** Estimation du taux de chômage d'après l'EPA au moyen de modèles transversaux et de séries chronologiques

Le but de ce projet est d'étudier l'utilisation de méthodes reposant sur des séries chronologiques pour estimer les taux de chômage par RMR/AR dans l'ensemble du Canada, et pour étudier et comparer les estimations des taux de chômage d'après l'EPA selon le modèle transversal de Fay-Herriot (FH) et selon des modèles reposant sur des séries chronologiques.

**Progrès :**

Nous avons étudié les estimations du taux de chômage d'après l'EPA à l'aide du modèle de FH, du modèle spatial de FH (SFH), du modèle spatio-temporal de FH (STFH) et du modèle transversal/reposant sur des séries chronologiques. Pour le modèle spatial de FH, nous avons étudié les estimations d'après l'EPA selon la méthode SAR-SFH de Petrucci et Salvati (2006) et selon la méthode CAR-SFH de You et Zhou (2011). Pour le modèle transversal/reposant sur des séries chronologiques, nous avons étudié le modèle de You, Rao et Gambino (2003) et le modèle STFH de Marhuenda, Molina et Morales (2013). Nous avons également étudié le progiciel R SAE et avons fait une présentation (You, 2018) comportant une comparaison des estimations sur petits domaines des taux de chômage d'après l'EPA selon différents modèles utilisant les fonctions R/S-Plus. Nos résultats indiquent que le modèle de FH est très utile et que, lorsqu'on l'applique aux données de l'EPA, il permet d'améliorer les estimations directes d'après les données d'enquête en réduisant les erreurs relatives et en haussant la précision. Le modèle spatial de FH et le modèle reposant sur des séries chronologiques ne donnent pas nécessairement de meilleurs résultats que le modèle de FH lorsqu'on les applique à l'EPA.

**SOUS-PROJET :** Estimation du taux de chômage selon la méthode hiérarchique bayésienne (HB) avec différentes modélisations de la variance au niveau des RMR et selon la CNP-S4 à l'aide de la fonction R et du système G-EST

Le but de ce projet est d'étudier la faisabilité de produire des estimations sur petits domaines des taux de chômage à l'aide de la méthode de modélisation HB au niveau des RMR et selon la CNP-S4. On utilise les fonctions d'EPD S-Plus et on procède à des comparaisons portant sur les résultats obtenus par le système G-EST pour ce problème.

**Progrès :**

Nous avons obtenu des EPD des taux de chômage d'après l'EPA au niveau des RMR et selon la CNP-S4 pour 3220 petits domaines à l'aide du progiciel SAS G-EST et des estimations fondées sur les modèles de variance d'échantillonnage HB avec les fonctions R et S-Plus. Les estimations basées sur le modèle de FH avec les méthodes REML et ADM ont été obtenues au moyen du progiciel SAS G-EST et des fonctions R. Le progiciel SAS G-EST et les fonctions d'EPD R donnent de très bons résultats sur le plan de l'efficacité de calcul. Dans le contexte de la comparaison des estimations fondées sur des modèles et des estimations directes, il pourrait être nécessaire d'utiliser l'EBLUP étalonnée de You, Rao et Hidiroglou (2013) ou la procédure d'étalonnage HB (You, Rao et Dick, 2004) pour réduire le biais possible lorsque les petits domaines se situent à un niveau peu élevé et que le nombre de domaines est très élevé.

**SOUS-PROJET :** Évaluation et comparaison des résultats de la méthode de modélisation HB des totaux et des taux au niveau des RMR/AR

Le taux de chômage peut être modélisé directement dans le modèle de Fay-Herriot. C'est cette approche qui a été utilisée jusqu'à présent. Il est aussi possible de modéliser le numérateur et le dénominateur du taux de façon séparée, et le taux peut être calculé à partir de ces deux estimations totales distinctes en se fondant sur des modèles au niveau du domaine.

**Progrès :**

Les modèles au niveau du domaine ont été étudiés en vue de produire des estimations des totaux pour le taux de chômage, le chômage et le taux de participation à la population active. Pour le taux de chômage et le total du chômage, les estimations directes de la variance d'échantillonnage sont lissées, parce que l'échantillon est de petite taille dans le cas de nombreuses RMR/AR. Pour le taux de chômage, le modèle de FH avec lissage des estimations de la variance d'échantillonnage est utilisé; dans le cas du total du chômage, on a plutôt recours à un modèle log-linéaire non apparié avec lissage des estimations de la variance d'échantillonnage. Pour le total relatif à la participation à la population active, la taille de l'échantillon est relativement grande, de sorte que les estimations directes de la variance d'échantillonnage sont utilisées dans le modèle log-linéaire non apparié, avec une modélisation

de la variance d'échantillonnage. Les estimations fondées sur des modèles sont comparées aux estimations fondées sur les données de recensement, et les résultats démontrent que les modèles proposés améliorent sensiblement les estimations directes d'après l'EPA en ce qui touche la réduction du biais et la réduction du CV. Pour le taux de chômage, nous avons également calculé un estimateur de taux simulé d'après le total du chômage simulé selon la méthode HB et les totaux de participation à la population active obtenus avec la procédure d'échantillonnage de Gibbs. Nos résultats démontrent que la modélisation directe du taux d'après l'EPA donne de bien meilleurs résultats que le taux calculé selon la méthode HB à partir des échantillons totaux simulés au moyen du modèle HB lorsque les résultats sont comparés aux estimations fondées sur les données de recensement. Des précisions sont fournies dans un rapport de recherche (You, 2019).

**SOUS-PROJET :** Estimations sur petits domaines pour Affaires mondiales Canada (AMC)

La tâche convenue avec Affaires mondiales Canada est de déterminer la faisabilité de la production d'estimations annuelles sur petits modèles du taux de chômage et du nombre d'emplois pour des domaines donnés. Ces domaines désignent des régions métropolitaines de recensement (RMR) par profession (CNP-S4) ainsi que des RMR par secteur (SCIAN4). Les expériences ont porté sur des données relatives à l'année 2016. Si les estimations obtenues à partir de méthodes d'estimations sur petits domaines semblent prometteuses, cet exercice pourrait être reproduit et se révéler utile pour les années intercensitaires. Dans le cas des années de recensement, les taux et les nombres désirés peuvent être produits directement à partir des données de recensement. Pour ces expériences, nous avons utilisé le modèle de Fay-Herriot.

L'estimation sur petits domaines a pour but de produire une estimation sur domaine de meilleure qualité que l'estimation directe en combinant les données de l'Enquête sur la population active (EPA) et l'information externe provenant de tous les domaines. La première étape du processus d'estimation sur petits domaines consiste en fait à lisser les estimations directes de la variance. On y parvient en créant un modèle de lissage de variance qui utilise les sources auxiliaires déterminées. L'étape suivante comporte la modélisation des estimations directes des caractéristiques d'intérêt d'après l'EPA. Les prédictions de ces deux modèles sont alors intégrées pour produire une estimation sur petits domaines, de pair avec la mesure de la qualité connexe.

**Progrès :**

Ce projet a été mené à terme, et un rapport final a été soumis aux clients (Bocci et Beaumont, 2018). On a utilisé jusqu'à trois sources auxiliaires pour modéliser les estimations du taux de chômage et du nombre d'emplois d'après les données d'enquête. La production d'estimations sur petits domaines a comporté principalement la construction de modèles et

l'évaluation de leur rendement. Il fallait donc analyser des graphiques et d'autres diagnostics pour déterminer la pertinence de chaque modèle. Bien que le taux de chômage et le nombre d'emplois aient fait l'objet d'une enquête, seules des estimations sur petits domaines du nombre d'emplois pour les domaines d'intérêt ont été communiquées au client, parce que les diagnostics du taux de chômage n'étaient pas convaincants. Somme toute, les estimations sur petits domaines du nombre d'emplois semblent constituer une amélioration par rapport aux estimations d'enquête directes. Il semble que la qualité des données auxiliaires ait influé quelque peu sur les estimations sur petits domaines. En outre, les améliorations associées à l'utilisation de techniques d'estimation sur petits domaines étaient plus manifestes au niveau des RMR selon le SCIAN4 qu'au niveau des RMR selon la CNP-4.

**SOUS-PROJET :** Estimation sur petits domaines du salaire horaire moyen et médian annuel par région économique combinée à la profession

La tâche consistait à déterminer s'il était possible de produire des estimations sur petits domaines du salaire horaire moyen et médian annuel pour de petits domaines combinant région économique et profession. Les estimations directes du salaire horaire moyen (ou médian) annuel ainsi que les mesures de la qualité au niveau des domaines désirés peuvent être calculées d'après les données de l'Enquête sur la population active (EPA) de Statistique Canada. En général, ces estimations ne sont pas de bonne qualité pour des domaines assortis d'échantillons de petite taille. Par conséquent, des techniques d'estimation sur petits domaines ont été envisagées pour les domaines souhaités. Cela consistait à modéliser les estimations d'après l'EPA à l'aide de renseignements auxiliaires du Recensement de 2016, qui correspondent à un concept similaire du salaire horaire moyen (ou médian) au niveau du domaine. Si les estimations obtenues à partir de méthodes d'estimations sur petits domaines pour l'année 2016 semblent prometteuses, cet exercice pourrait être reproduit et se révéler utile pour les années intercensitaires. Dans le cas des années de recensement, les moyennes désirées peuvent être produites directement à partir des données de recensement.

**Progrès :**

Nous avons produit des estimations sur petits domaines du salaire horaire médian pour les années 2011, 2016 et 2017, et du salaire horaire moyen pour les années 2016 et 2017. Les résultats de notre enquête préliminaire ont fait l'objet d'une présentation plus ou moins formelle aux clients et à des spécialistes de la méthodologie des enquêtes en mars 2019. De plus, on a rédigé un document décrivant les principales étapes de production des estimations sur petits domaines dans le cadre de ce projet.

**SOUS-PROJET :** Élaboration d'une méthodologie d'estimation sur petits domaines pour l'Enquête mensuelle sur les industries manufacturières

En 2016, on a demandé aux méthodologistes de l'ICMIC de se pencher sur la possibilité de recourir à des techniques d'estimation sur petits domaines pour l'Enquête mensuelle sur les

industries manufacturières (EMIM). À partir de ces premiers travaux, une approche a été élaborée pour produire des estimations sur petits domaines mensuelles des ventes totales de biens fabriqués par région métropolitaine de recensement/agglomération de recensement (RMR/AR) et par industrie. Plus précisément, 12 RMR/AR ont été prises en compte dans le contexte des industries manufacturières, ce qui représente environ 320 domaines par mois. Depuis, l'EMIM a fait l'objet d'une refonte, et on prévoyait procéder à une révision historique au début de 2019.

**Progrès :**

Par suite de la refonte récente de l'EMIM, la méthode utilisée pour obtenir les estimations directes du total et les estimations directes correspondantes de la variance a changé. Cela a mené à une réévaluation de la méthode d'estimation sur petits domaines au cours du présent exercice. Par conséquent, les modèles d'estimation sur petits domaines et la stratégie connexe ont été modifiés. Conformément à la nouvelle stratégie, des estimations sur petits domaines mensuelles ont été produites pour les périodes allant de décembre 2012 à mai 2018. Un document décrivant la stratégie dans sa version la plus récente a été rédigé. Le changement le plus important apporté à la stratégie est que, dorénavant, aucun domaine n'est estimé à l'aide du système d'EPD. Les nouvelles estimations sont synthétiques.

En outre, nous avons fourni des spécifications et un code informatique aux méthodologistes de l'EMIM afin que ces estimations puissent être intégrées au système opérationnel utilisé pour produire les estimations directes de l'enquête (Bocci et Beaumont, 2019). Ces travaux se sont déroulés de septembre à décembre 2018. En mai 2019, nous produisons de nouveau les estimations sur petits domaines pour les périodes allant de décembre 2012 à mai 2018, étant donné que la révision historique prévue des estimations directes a été effectuée.

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

## Bibliographie

Marhuenda, Y., Molina, I. et Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

Petrucci, A., et Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics*, 11, 169-182.

You, Y., et Zhou, Q.M. (2011). Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé. *Techniques d'enquête*, 37, 1, 31-44. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11445-fra.pdf.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 1, 27-36. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003001/article/6602-fra.pdf.

You, Y., Rao, J.N.K. et Hidiroglou, M. (2013). De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot. *Techniques d'enquête*, 39, 1, 243-255. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11830-fra.pdf.

## 1.2 Recherche-développement – Couplage 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 l'estimation exacte des erreurs de couplage. Notons que le chiffrement des données rend la tâche encore plus ardue.

**SOUS-PROJET :** Estimation automatisée des erreurs de couplage

La déclaration des erreurs de couplage est une exigence importante selon le modèle du processus de couplage d'enregistrements de l'organisme (Statistique Canada, 2017) et les lignes directrices approuvées pour la déclaration de l'exactitude des couplages (Statistique Canada, 2019). Pourtant, elle continue de représenter un défi majeur qui limite la production et l'utilisation automatisées de données couplées. Les solutions précédentes comprennent des examens manuels (Newcombe, Smith et Howe, 1983; Dasylva, Abeysundera, Akpoué, Haddou et Saidi, 2016; Dasylva, 2018) et l'utilisation de modèles statistiques particuliers. Les examens manuels consistent en une inspection visuelle des paires d'enregistrements visant à déterminer leur état de correspondance. Ils sont coûteux en main-d'œuvre et intrinsèquement subjectifs (Newcombe et coll., 1983). Pour éviter ces problèmes, on peut utiliser diverses combinaisons de modélisations par mélange tirées de la littérature, y compris des mélanges loglinéaires (Fellegi et Sunter, 1969; Winkler, 1988; Winkler, 1993) avec ou sans l'hypothèse d'une indépendance conditionnelle ou des mélanges comportant des normales transformées (Belin et Rubin, 1995). Ces modèles ont toutefois leurs limites. En effet, le fait de supposer une indépendance conditionnelle (Fellegi et Sunter, 1969) conduit à des estimateurs biaisés parce que cette condition est rarement satisfaite en pratique. De plus, les mélanges loglinéaires avec interactions (Winkler, 1988; Winkler, 1993; Thibaudeau, 1993) ne sont généralement pas reconnus pour avoir une propriété d'identification. Un modèle ne possédant pas cette propriété produit des estimateurs biaisés, quelle que soit la taille de l'échantillon. Enfin, les mélanges à normales transformées tentent d'estimer les erreurs de couplage en modélisant la distribution du poids de couplage des paires (probabiliste) (Belin et Rubin, 1995). Il faut toutefois correctement séparer la distribution du poids appariée et la distribution non appariée. En outre, tous les modèles cités précédemment sont limités par le fait qu'ils portent sur une seule paire. Par conséquent, ils ne peuvent pas mesurer l'erreur de couplage au niveau de l'enregistrement, par exemple vérifier si un enregistrement comportant plusieurs liens adjacents a un lien positif véritable.

**Progrès :**

Un nouveau modèle statistique est proposé aux fins d'estimation automatisée des erreurs de couplage lors du couplage de deux registres complets sans enregistrements en double d'une

grande population iid avec $N$ individus. Il se fonde sur le concept de *voisin* d'un enregistrement donné, généralement défini comme étant un autre enregistrement qui forme une paire avec le premier enregistrement satisfaisant un certain critère : par exemple, la présence d'un lien au moyen d'une méthode de couplage qu'elle soit déterministe, hiérarchique, probabiliste, basée sur l'apprentissage automatique (Les solutions d'apprentissage automatique supervisé peuvent estimer les erreurs de couplage par validation croisée. Cependant, ces estimations sont souvent peu fiables en raison d'erreurs dans les données d'entraînement.) ou autre. Intuitivement, la distribution du nombre de voisins peut fournir de nombreux renseignements sur l'occurrence des erreurs de couplage. Ainsi, dans le cas considéré, ce lien peut être résumé comme dans le tableau suivant.

**Tableau 1**
**Nombre de voisins et erreurs de couplage**

| Nombre de voisins ($n_i$) | Faux négatifs | Faux positifs |
|---|---|---|
| 0 | 1 | 0 |
| 1 | ? | ? |
| Un grand nombre | ? | $\geq n_i - 1$ |

On utilise un modèle pour prédire le nombre de faux positifs et de faux négatifs quand ils ne sont pas entièrement déterminés par le nombre de voisins $n_i$.

Les paramètres du modèle sont liés aux mesures d'erreur ciblées (Statistique Canada, 2019) lors du couplage d'enregistrements voisins sans étape de résolution de conflit.

Le nouveau modèle présente de nombreux avantages. Il estime les taux d'erreur avec exactitude pour *toute* méthode de couplage, qu'elle soit probabiliste ou non (déterministe, hiérarchique, apprentissage machine, etc.). Il peut également fournir des estimations exactes des poids de couplage probabiliste, qui mènent à des décisions de couplage automatisé (c'est-à-dire un seuil unique et pas de zone grise si l'on utilise la méthode probabiliste) *optimales* entre *toutes les décisions possibles, quelle que soit la méthode de couplage.* De plus, le nouveau modèle est d'utilisation facile, car il ne tient pas compte des interactions entre variables de couplage, bien que le fait de préciser les interactions puisse augmenter l'efficacité des estimateurs. Les autres avantages comprennent la capacité d'estimer les faux négatifs attribuables à un blocage (définir un voisin uniquement à partir des critères de blocage) et les estimations au niveau de l'enregistrement des erreurs de couplage.

Le modèle a été appliqué avec succès dans des simulations et dans une étude empirique sur des données administratives. Il est décrit plus en détail dans le document de travail (Dasylva, Goussanou, Ajavon et Abousaleh, 2019).

On est en train de modifier la nouvelle méthodologie pour obtenir des estimateurs plus précis en prenant un échantillon d'estimation plus grand et pour répondre à des cas comprenant deux fichiers ou plusieurs fichiers comportant des doubles et un certain sous-dénombrement, y compris la résolution de conflits.

**SOUS-PROJET :** Codage automatisé utilisant une méthode de couplage d'enregistrements

Le codage consiste à attribuer le bon code à un champ à partir d'un ensemble de codes fini, selon certaines entrées. Il s'agit d'un problème de classification et d'une partie essentielle de la tâche de vérification de tout travail portant sur des fichiers de recensement, d'enquête ou administratifs. Certaines solutions ont recours à la méthode d'imputation par le plus proche voisin en reliant chaque nouvelle observation à un sous-ensemble d'observations codées (Wenzowski, 1988; Chu, Yeung et Dasylva, 2018). Ce sous-ensemble représente un bassin de donneurs dans lequel on sélectionne un donneur pour attribuer le code correspondant à la nouvelle observation. Cette méthodologie nécessite d'estimer l'erreur de codage et de régler les paramètres de couplage en fonction du taux d'erreur cible.

**Progrès :**

Une méthodologie d'estimation du taux d'erreur de codage et d'établissement des paramètres sous-jacents en fonction d'un taux d'erreur cible a été décrite dans Dasylva (2019). La méthodologie a donné de bons résultats dans les simulations, à en juger par la concordance entre le taux d'erreur cible et le taux atteint (Savard, 2019). Elle a également été appliquée dans une étude empirique utilisant les données du recensement de 2016 sur la langue maternelle, où elle a amélioré la solution proposée par Chu et coll. (2018) en établissant les paramètres de couplage en fonction du taux d'erreur cible. Comme dans les simulations, le taux d'erreur atteint était proche du taux ciblé.

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

## Bibliographie

Belin, T.R., et Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 430, 694-707.

Dasylva, A., Abeysundera, M., Akpoué, B., Haddou, M. et Saidi, A. (2016). Mesurer la qualité d'un couplage probabiliste par des vérifications manuelles. Recueil : Symposium 2016, Croissance de l'information statistique : défis et bénéfices, Statistique Canada.

Fellegi, I.P., et Sunter, A.B. (1969). A theory of record linkage. *JASA*, 64, 1183-1210.

Newcombe, H., Smith, M. et Howe, G. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medecine*, 13, 157-169.

Statistique Canada (2017). *Modèle du processus d'un projet de couplage d'enregistrements,* Nº 12-605-X au catalogue, Statistique Canada.

Statistique Canada (2019). Guidelines for Reporting the Record Linkage Accuracy, Statistique Canada.

Thibaudeau, Y. (1993). Le pouvoir discriminant des structures de dépendance dans le couplage d'enregistrements. *Techniques d'enquête*, 19, 1, 35-43. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1993001/article/14477-fra.pdf.

Wenzowski, M.J. (1988). ACTR un système généralisé de codage automatique. *Techniques d'enquête*, 14, 2, 317-326. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1988002/article/14586-fra.pdf.

Winkler, W.E. (1988). Using the EM algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter Model of Record Linkage. JSM 1993: Dans *Proceedings of the 1993 Joint Statistical Meetings, Survey Research Methods Section*, 8 au 12 août 1993, San Francisco, CA. Alexandria VA: ASA; 274-279.

## 1.3 Recherche-développement – Systèmes généralisés

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

- G-Est : un système généralisé d'estimation;
- G-Sam : un système généralisé d'échantillonnage;
- Banff : un système généralisé de vérification et d'imputation.

Outre le soutien et la formation liés aux systèmes généralisés, l'équipe se charge également de la recherche développementale liée à la visualisation des données, à l'estimation de la variance et à d'autres méthodes d'enquête liées à des processus d'enquête.

**SOUS-PROJET :** Soutien continu au système généralisé

L'unité Systèmes généralisés facilite l'utilisation des systèmes pour les enquêtes nouvelles et existantes ainsi que pour les programmes statistiques en cours de restructuration.

**Progrès :**

L'équipe de support des systèmes généralisés a fourni un support continu aux utilisateurs, mis à jour et présenté de la formation dans divers forums et rencontré des délégués internationaux pour discuter du développement actuel et futur des systèmes généralisés. Le groupe a rencontré des délégations d'Italie, du Japon, de Singapour et d'Irlande.

**SOUS-PROJET :** Développement de système généralisé - Exploration de méthodes supplémentaires

**Progrès :**

L'équipe de Banff a dirigé l'organisation d'un hackathon pour encourager l'exploration de nouvelles méthodes d'imputation. L'événement consistait en une journée dédiée au cours de laquelle les participants ont exploré un ensemble de données et ont été encouragés à trouver et à appliquer des méthodes d'imputation pour traiter la non-réponse. Le hackathon s'intitulait «Pouvez-vous battre BANFF» et, ultimement, les méthodes proposées ont été comparées à celles disponibles dans les procédures BANFF. En fin de compte, un certain nombre de méthodes ont été identifiées et la méthode la plus prometteuse était l'absence de forêt, disponible via le package R, *missForest*.

Par la suite, afin d'accroître les fonctionnalités du processeur Banff, l'équipe de développement de Banff a montré comment intégrer des programmes externes dans l'architecture de Banff. Compte tenu de la disponibilité et de la popularité des programmes de vérification de données externes à code source ouvert, ces options peuvent être attrayantes pour les utilisateurs actuels et futurs. Plus précisément, un flux de travaux contenant à la fois les procédures de

Banff standard et une étape d'imputation à l'aide du package R *missForest* a été développé et appliqué avec succès dans un environnement de test.

L'utilisation de R en général comme outil de recherche et de production a également été exploré. Les résultats de cette exploration ont été documentés et discutés avec le Comité consultatif sur les méthodes statistiques de Statistique Canada et lors d'une conférence internationale sur l'utilisation de R pour les statistiques officielles (Fortier et Thomas, 2018).

**SOUS-PROJET :** Développement de système généralisé - Cadre d'édition de données statistiques

**Progrès** :

Le système de vérification et d'imputation de Banff se compose de neuf procédures SAS personnalisées qui exécutent diverses tâches de traitement (vérification et imputation) de données statistiques. Ces procédures incluent un cadre de vérification des données qui permet aux sorties d'informations d'une procédure d'agir comme des entrées dans une autre, permettant ainsi des processus complexes de vérification de données, facilités par le processeur Banff. Ce travail a été présenté à l'atelier de la Commission économique des Nations Unies pour l'Europe sur la vérification de données statistiques (Neuchâtel, Suisse, septembre 2018).

En outre, l'équipe de recherche et développement de Banff a contribué à la version 2.0 du modèle générique d'édition de données statistiques (GSDEM) publié en juin 2019 (https://statswiki.unece.org/plugins/servlet/mobile?contentId=117771706#content/view/117771706), destiné à servir de référence à tous les statisticiens officiels dont les activités incluent la vérification de données.

**SOUS-PROJET :** Développement du système généralisé - Publication de G-Sam 1.02.001

**Progrès :**

Une nouvelle version de G-Sam a été développée, testée et publiée. La version 1.02.001 inclut une restructuration partielle et un recodage des trois modules (Stratification, allocation et sélection) afin d'améliorer l'efficacité et de faciliter le support technique et méthodologique. D'autres améliorations incluent la simplification des entrées pour l'utilisateur et l'amélioration des messages du journal SAS générés par G-Sam afin de mieux répondre aux besoins des utilisateurs.

**SOUS-PROJET :** Développement du système généralisé - Publication de G-EST 2.02

**Progrès :**

Une nouvelle version de G-EST a été développée, testée et publiée. Parmi les principales modifications apportées à la version 2.02, citons l'amélioration des performances pour

l'estimation de la variance d'échantillonnage à l'aide de la linéarisation de Taylor, la gestion des modèles de saut de questionnaire, une correction de bugs liée à l'estimation de la variance d'échantillonnage en présence de non-réponse et certaines modifications méthodologiques apportées aux exclusions d'étalonnage. Une deuxième version (version 2.02.008) a réintroduit le traitement parallèle pour la variance d'échantillonnage et une nouvelle version (3.2.1) de SEVANI.

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

## 1.4 Recherche prospective – Intégration des données

L'avènement du Web dans les années 1990 a ouvert la porte à de nouveaux modes de collecte de données pour les enquêtes, à savoir les vastes panels à participation volontaire sur le Web, et les mégadonnées. Les *panels à participation volontaire sur le Web* sont constitués de personnes qui utilisent régulièrement le Web et à qui on pose des questions sur divers sujets. Le terme générique *mégadonnées* désigne des ensembles de données tellement importants ou complexes qu'ils excèdent la capacité des applications de traitement des données classiques. Souvent, les panels sur le Web et les mégadonnées ne s'appuient pas sur des plans d'échantillonnage probabiliste.

Le projet vise principalement trois objectifs :

i) Évaluer la possibilité d'utiliser l'appariement d'échantillons ou d'autres techniques d'intégration des données pour certains programmes de Statistique Canada afin de réduire le fardeau des répondants ou les coûts de collecte des données.
ii) Élaborer ou adapter de nouvelles méthodes pour résoudre des problèmes pratiques.
iii) Élaborer et mettre à l'essai des prototypes qui mettent en application les méthodes les plus prometteuses.

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

Statistique Canada a récemment amorcé une phase de modernisation. L'un des éléments clés des différentes initiatives de modernisation en cours est d'utiliser davantage les diverses sources de données en les combinant. C'est ce que l'on appelle l'intégration des données dans la littérature statistique. Le but de ce projet était d'effectuer une revue exhaustive de la littérature sur les méthodes statistiques d'intégration des données.

**Progrès :**

La revue de la littérature a été menée durant l'été de 2018. L'examen a englobé les méthodes fondées sur le plan de sondage, comme celles reposant sur des bases de sondage multiples et l'étalonnage fondé sur le plan de sondage, de même que les méthodes fondées sur des modèles, comme l'appariement statistique, l'étalonnage dépendant d'un modèle et la pondération par l'inverse du score de propension. Cet examen a été présenté au Colloque francophone sur les sondages de Lyon, et un document a été rédigé et soumis pour publication (Beaumont, 2019).

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

On sait que les données provenant de sources non probabilistes, par exemple des panels sur le Web, sont assorties d'un biais de sélection. Une approche possible pour prendre en compte

le biais de sélection consiste à modéliser la probabilité de sélection dans l'échantillon non probabiliste, puis à pondérer chaque unité de l'échantillon par l'inverse de cette probabilité. Cette méthode est connue sous le nom de pondération par l'inverse du score de propension. Elle est particulièrement utile lorsque l'échantillon contient de nombreuses variables d'intérêt, car une même stratégie de pondération peut être appliquée à toutes ces variables.

Si les variables auxiliaires étaient connues pour l'ensemble de la population, le problème serait essentiellement identique à celui associé à la pondération de la non-réponse dans une enquête probabiliste. Toutefois, la plupart du temps, on ne dispose pas de cette information. Pour trouver une solution à ce problème, Chen, Li et Wu (2019) ont proposé une méthode entièrement paramétrique afin de combiner l'échantillon non probabiliste avec un échantillon probabiliste qui contient les valeurs de toutes les variables auxiliaires. Si l'ensemble de variables auxiliaires utiles est accompagné des interactions appropriées, cette méthode peut être appliquée directement. Le véritable défi consiste à choisir les variables auxiliaires et les interactions.

Les arbres de régression gagnent en popularité à titre de méthode non paramétrique permettant de choisir automatiquement les variables auxiliaires et de traiter les interactions dans le cas des problèmes de régression standard. Le but de ce projet est d'étendre l'utilisation des arbres de régression aux cas comportant la combinaison d'un échantillon probabiliste et d'un échantillon non probabiliste, et de rédiger un programme R qui met en application cette méthode.

**Progrès :**

Nous avons élaboré une extension des arbres de régression dans le contexte de la combinaison d'échantillons non probabilistes et probabilistes. Nous avons également mis au point un programme R qui met en application la méthode, et nous l'avons évalué au moyen d'une petite étude de simulation. Les résultats préliminaires semblent prometteurs. Les résultats de ce projet ont été présentés lors de l'assemblée de 2019 de la SSC (Chu et Beaumont, 2019).

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

**SOUS-PROJET :** Analyse des données de catégories obtenues par couplage probabiliste

**Description :**

En se fondant sur l'hypothèse selon laquelle les faux positifs constituent la seule source d'erreurs de couplage, on a utilisé la méthode de Chipperfield, Bishop et Campbell (2011)

consistant à inférer des données binaires au moyen de la régression logistique, et une étude de simulation a été menée en vue d'examiner l'efficacité de la méthode proposée. Les résultats des simulations ont montré que les estimateurs sont sans biais et présentent une plus petite variance que ce que l'on obtient à partir de données examinées manuellement. On a élaboré un ensemble de macros SAS pour analyser les données binaires par régression logistique.

Cette approche sera étendue aux données catégoriques, aux données des tableaux de contingence et aux données de dénombrement, et elle constituera une solution utile lors de l'intégration de multiples sources de données par couplage d'enregistrements.

L'objet de cette recherche est d'élaborer un prototype SAS pour analyser les données catégoriques couplées à l'aide d'un modèle logistique, d'un modèle log-linéaire ou d'un modèle de Poisson.

**Progrès :**

Une macro SAS a été élaborée et utilisée lors de l'analyse de données binaires par régression logistique en présence exclusivement d'enregistrements non couplés (faux négatifs). On utilise une approche de pseudo-vraisemblance qui attribue des coefficients de pondération aux enregistrements couplés. L'ajustement de la pondération en l'absence de couplage est efficace si la pondération est liée aux variables qui influent sur la probabilité qu'un enregistrement ne soit pas couplé ainsi qu'à la variable à l'étude.

Les résultats ont été présentés à la conférence du FCSM ainsi qu'à la conférence de l'INCASS, à Montréal, par Saïdi, Chu, Dasylva et Labrecque-Synnott (2018).

**SOUS-PROJET :** Estimation E-M des paramètres de couplage probabiliste sous une hypothèse de dépendance entre les champs d'un enregistrement

**Description :**

Il est bien connu que le non-respect de l'hypothèse d'indépendance conditionnelle du modèle de Fellegi-Sunter (F-S) engendre en général des estimations biaisées des distributions m (probabilités liées) et u (probabilités non liées). Cette limitation a des répercussions sur le processus de classification (détermination des seuils et de l'optimalité de la règle de décision de F-S) ainsi que sur les niveaux des erreurs pour les faux positifs et les faux négatifs. L'algorithme EM avec interactions entre les champs d'un enregistrement estime les paramètres de couplage et donne lieu à des améliorations substantives de l'efficacité du couplage.

L'objectif de ce projet est d'implémenter dans G-coup l'algorithme EM sans hypothèse d'indépendance conditionnelle pour le calcul des paramètres du couplage probabiliste. Nous travaillerons à l'amélioration du prototype développé en 2015 par Dasylva et collaborateurs.

**Progrès :**

Les macros SAS ont été élaborées en 2015 par Abel Dasylva et ses collaborateurs afin de calculer le poids EM à l'aide de modèles log-linéaires dans le cadre de la procédure PROC CATMOD, avec des termes d'interaction et des données manuelles facultatives, selon l'hypothèse de l'omission des résultats manquants (sous un modèle de données manquant au hasard). Les auteurs ont mené des tests en utilisant des données produites à partir de modèles théoriques.

L'équipe a mis en concordance et documenté les fichiers d'entrée et de sortie ainsi que les fichiers intermédiaires à partir des deux algorithmes. Un examen plus approfondi des besoins techniques entourant la mise en œuvre dans G-Coup a également eu lieu. Les objectifs de recherche dans le cadre de ce projet consistent à transformer des ensembles de données dans G-Coup pour l'application de nouvelles macros et de macros d'essai avec des données réelles en vue d'effectuer des comparaisons portant sur l'état de couplage véritable en se fondant sur des identificateurs uniques connus. Nous étudierons également la situation où des données manuelles sont disponibles, en précisant les interactions au niveau du nouveau modèle, les limites des paramètres et les diagnostics du modèle de sortie.

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

## Bibliographie

Chen, Y., Li, P. et Wu, C. (2019). Doubly robust inference with non-probability survey samples. Manuscrit non-publié.

Chipperfield, J.O., Bishop, G.R. et Campbell, P. (2011). Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées. *Techniques d'enquête*, 37, 1, 17-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11444-fra.pdf.

## 1.5 Recherche prospective – Approches non-probabilistes

**SOUS-PROJET :** Exploration sur l'utilisation de sources de données non-probabilistes

Nous avons récemment acquis des données d'un panel non probabiliste à participation volontaire contenant des variables similaires à quelques-unes de celles de l'Enquête sur la santé dans les collectivités canadiennes (ESCC). Il a été demandé aux membres du panel de répondre à un court questionnaire au moyen d'une application sur téléphone cellulaire. Ils recevaient en retour des récompenses de leur programme préféré. Il est bien connu que les estimations directement dérivées de données de panel font l'objet d'un biais de sélection. Ces données non probabilistes fournissent une occasion d'examiner l'efficacité des techniques d'intégration des données pour réduire le biais de sélection.

**Progrès :**

Nous avons évalué deux techniques : l'appariement statistique, également appelé appariement d'échantillons (Rivers, 2007), et l'étalonnage dépendant d'un modèle. L'efficacité des deux méthodes dépend de la force des variables auxiliaires disponibles dans les deux sources. Dans notre expérience, nous avons utilisé les variables suivantes : l'âge, le sexe, le niveau de scolarité, l'état matrimonial, et la région sociosanitaire. Nous avons ensuite comparé les estimations obtenues à l'aide de l'appariement statistique et de l'étalonnage aux estimations de l'ESCC. Nous avons observé que les deux méthodes réduisaient le biais substantiel observé dans le cas des estimations directes du panel. L'appariement statistique a semblé légèrement plus efficace que l'étalonnage. Cela peut tenir à sa nature non paramétrique, contrairement à l'étalonnage, qui repose sur un modèle linéaire. Cependant, un biais non négligeable a persisté. La présence de ce biais peut tenir à deux raisons : i) les variables auxiliaires utilisées n'étaient pas des valeurs explicatives assez fortes des variables de santé d'intérêt; ii) des erreurs de mesure étaient présentes dans les données du panel. Nos résultats ont été présentés aux conférences de la SSC et de l'INCASS à l'été de 2018. Un rapport décrivant nos conclusions a été rédigé et publié dans le cadre des travaux de la SSC (Chatrchi, Beaumont, Gambino et Haziza, 2018).

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

### Bibliographie

Rivers, D. (2007). Sampling for Web Surveys. *Proceeding of the Joint Statistical Meeting*, Salt Lake City, Utah.

## 1.6 Recherche prospective – Science des données

Dans le cadre du Symposium international de 2018 sur les questions de méthodologie, la Direction de la Méthodologie a organisé un panel de discussion intitulé « La science des données et l'apprentissage automatique pour les statistiques officielles – réussites, opportunités et défis ».

En 2018, les Directions de méthodologie et des études analytiques s'apprêtent à mettre sur pied un Centre d'excellence en science des données (CdESD) au nom du secteur. Le CdESD est hébergé dans la méthodologie et fonctionne en étroite collaboration avec les divers autres agents et groupes de science des données de l'organisation.

Le rôle principal du Centre d'excellence en science des données est d'offrir la rigueur scientifique nécessaire aux activités de science des données. Les méthodologistes travaillant dans le CdESD seront appelés à :

- Fournir des consultations et des conseils sur l'application des méthodes de science des données dans le contexte de la création, du traitement et de l'analyse des données pour les statistiques officielles;
- Contribuer directement aux projets pilotes ou exploratoires de science des données, en particulier ceux impliquant des techniques avancées liées au traitement et à l'analyse des données (c'est-à-dire l'analyse prédictive);
- Diriger et mener des recherches liées au soutien et à l'avancement des applications de la DS pour la production de statistiques officielles;
- Évaluer ou proposer des activités de formation pour renforcer les capacités en matière de compétences de DS dans l'ensemble de l'Agence;
- Offrir du soutien et des conseils quant aux composantes liées à l'éthique pour les recherches utilisant des méthodes d'intelligence artificielle.

L'un des projets menés conjointement avec divers partenaires est une revue en profondeur des techniques d'apprentissage automatique pour la classification des données. Cette revue traite de 4 sous-thèmes : les ensembles d'entainement dééquilibrés, le trasfert de l'apprentissage, la mise à niveau des hyperparamètres et l'evaluation de la qualité. Les réflexions et les résultats ont été présentées au Comité consultatif sur les methodes statistiques (Yeung, Chu, Laroche and Fortier, 2018).

D'autres utilisations de techniques et d'outils liés à la science des données sont décrites ailleurs dans ce rapport, dans les sous-sections pertinentes.

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

## 1.7 Recherche divisionnaire

**SOUS-PROJET :** Réévaluation de l'estimateur composite par régression pour l'Enquête sur la population active

L'Enquête sur la population active a commencé à utiliser l'estimateur composite par régression il y a près de 20 ans, après une analyse approfondie. Bien qu'il soit très efficace pour réduire la variance des estimations d'un mois à l'autre et de niveau, une récente analyse préliminaire montre qu'il produit peut-être un biais de taille inattendue dans certaines estimations. La source de ce biais n'est pas immédiatement évidente, mais il se peut qu'elle soit liée au problème de dérive examiné par Fuller-Rao (2001).

Le but du projet est d'étudier de possibles ajustements à l'estimateur composite par régression, y compris la réévaluation du facteur de combinaison utilisé dans la combinaison de deux estimateurs composites différents. Ce facteur a été défini en se basant sur des études empiriques réalisées par Gambino, Kennedy et Singh (2001) et le fait de répéter l'étude avec des données plus récentes pourrait s'avérer instructif. En outre, nous aimerions étudier des ajustements aux catégories de totaux de contrôle composites, étant donné que certaines des catégories contiennent très peu de répondants et devraient être combinées.

**Progrès :**

Les principaux articles sur l'estimateur composite par régression de Fuller et Rao (2001), Gambino et coll. (2001) et les références connexes dans ces articles ont été examinés en profondeur. L'examen d'un article plus récent de Preston (2015) a également été effectué. Les estimations mensuelles couvrant la période de janvier 1997 à décembre 2018 fondées sur la méthode de régression généralisée ont été comparées aux estimations issues de la méthode d'estimation composite par régression pour un grand nombre de séries de l'EPA. Les résultats provisoires indiquent qu'il y a quelques différences entre les estimations produites au moyen de ces deux méthodes. Les différences sont plus prononcées dans certaines séries que dans d'autres. À ce stade-ci, la cause de ces différences n'est pas encore connue. On croit que ces différences peuvent s'expliquer par les propriétés théoriques des deux méthodes, l'incidence de la non-réponse ou le choix du nombre de variables (et de totaux) de contrôle de calage. Des travaux sont en cours pour détecter et expliquer officiellement les causes potentielles de ces différences et, à une date ultérieure, étudier d'éventuelles solutions dans la mesure du possible.

Pour plus de renseignements, communiquez avec :
**Emmanuel Benhin** (613-862-7638, emmanuel.benhin@canada.ca).

**SOUS-PROJET :** Signaler la qualité en utilisant des intervalles de confiance

En 2017, le Comité des méthodes et des normes (CMN) a approuvé notre recommandation d'adopter comme pratique exemplaire l'utilisation d'intervalles de confiance pour mesurer et signaler la qualité des estimations. Le but du projet est d'effectuer de la recherche en vue d'appuyer l'utilisation d'intervalles de confiance pour signaler la qualité.

**Progrès :**

Au cours de la période, un ensemble de règles a été mis au point pour décider si une estimation et son intervalle de confiance pourraient être diffusés ou s'ils devraient être supprimés pour des raisons de qualité. Les règles proposées pour les enquêtes sociales ont été présentées au Comité technique de la Division des méthodes de la statistique sociale (DMSS) (Neusy, Boulet, Duggan, Mach, Mantel et Reedman, 2018).

Un texte normalisé pour les guides de l'utilisateur a été rédigé aux fins des enquêtes dans lesquelles on choisit d'utiliser des intervalles de confiance pour signaler la qualité. Ce texte a été utilisé dans l'Étude sur les transferts de fonds internationaux.

Les gestionnaires de la Division des enquêtes spéciales ont assisté à une présentation (Neusy, 2019a) sur les problèmes liés à l'utilisation des coefficients de variation (CV) et les avantages d'utiliser des intervalles de confiance pour signaler la qualité.

Une note interne documentant la dérivation de l'intervalle de Wilson pour les chiffres pondérés a été rédigée (Neusy, 2019b).

Pour plus de renseignements, communiquez avec :
**Elisabeth Neusy** (613-863-3513, elisabeth.neusy@canada.ca).

**SOUS-PROJET :** Validation de l'extension au Bootstrap de Rao-Wu proposée par Pérez-Duarte applicable aux échantillons tirés selon des plans proportionnels à la taille à un ou deux degrés à l'aide d'une simulation Monte Carlo

L'estimateur de variance provenant du Bootstrap de Rao-Wu, aussi appelé « rescaled Bootstrap », est utilisé par plusieurs enquêtes à Statistique Canada. Cet estimateur a fait ses preuves avec des échantillons provenant de plans de sondage stratifiés à un degré où les unités sont échantillonnées à l'aide d'un tirage aléatoire simple. Dans les cas où l'échantillonnage est à plusieurs degrés, l'estimateur de Rao-Wu est efficace si le premier degré est échantillonné à l'aide d'un tirage aléatoire simple à tous les degrés et si la fraction de sondage est petite. Pour les enquêtes utilisant un échantillonnage à plusieurs degrés où le

premier degré est échantillonné à l'aide d'un tirage proportionnel à la taille avec remise, une extension de l'estimateur de Rao-Wu existe.

Osiewicz et Pérez-Duarte (2012) proposent une modification à l'extension de l'estimateur de Rao-Wu permettant d'estimer les variances d'échantillonnage d'estimateurs provenant de plans de sondage proportionnel à la taille sans remise et de plans de sondage à deux degrés où le tirage au premier degré est sans remise et proportionnel à la taille suivant la méthode de Rao-Hartley-Cochran (RHC) et le deuxième degré est une sélection aléatoire simple sans remise.

Le but de cette recherche est de valider cette extension à l'aide d'une simulation Monte Carlo.

**Progrès :**

Pour les besoins de cette étude, huit plans d'échantillonnage ont été testés. Tous les plans utilisent la même population et un échantillonnage à deux degrés où l'échantillonnage au premier degré est un échantillonnage proportionnel à la taille sans remise de RHC et l'échantillonnage au deuxième degré est un échantillonnage aléatoire simple. Seules les fractions de sondage au premier et au deuxième degré varient. L'espérance Monte Carlo du CV de l'estimateur d'un total d'un plan à deux degrés utilisant l'échantillonnage probabiliste proportionelle à la taille (PPT) de RHC (Rao, Hartley and Cochran, 1962) ainsi que l'espérance Monte Carlo du CV de l'estimateur de variance proposé par Osiewicz et Pérez-Duarte (2012) ont été produites. À ce jour, le CV de l'estimateur de variance explose. Les raisons de cette explosion n'ont pas encore été étudiées. Un article interne décrivant la recherche et les résultats obtenus jusqu'à maintenant est écrit (Devin, 2018). Les programmes SAS sont aussi disponibles.

Pour plus de renseignements, communiquez avec :
**Nancy Devin** (613-618-1027, nancy.devin@canada.ca).

**SOUS-PROJET :** Données des compteurs électriques pour le Manitoba

Ce projet porte sur les données mensuelles de compteurs électriques au Manitoba. Il s'agit d'un projet qui exploite des mégadonnées et qui permettra une compréhension sans précédent de l'utilisation de l'électricité au Manitoba. L'un des objectifs de ce projet est de produire des estimations totales trimestrielles de la consommation électrique résidentielle pour les aires de diffusion du Manitoba.

**Progrès :**

En novembre 2018, une présentation a été donnée lors du Symposium international de 2018 sur les questions de méthodologie (Duddek, 2018), et le compte rendu correspondant a été

rédigé. C'était un résumé du travail accompli pour regrouper les données sur l'électricité à la géographie du Recensement de 2016, puis évaluer la couverture, entre autres objectifs.

Les données des compteurs électriques ont également été evaluées pour une utilisation à des fins d'imputation dans le contexte de l'Enquête sur les ménages et l'environnement (EME). Dans le cadre de cette étude, les donnéées ont été analysées en lien avec des données des degrés-jours de chauffage du site Web d'Environnement Canada. Ceci a permis l'élaboration d'un modèle de régression au niveau du compteur liant la consommation électrique trimestrielle à la météo. Nous avons alors pu effectuer la détection et le traitement des valeurs aberrantes pour nous occuper de certains cas particulièrement inhabituels de sous-dénombrement des kWh mesurés pour le premier trimestre de 2015.

Ce travail a été présenté au Comité technique des enquêtes auprès des ménages le 1er mars 2019, et est intitulé « L'utilisation des données de Manitoba Hydro pour évaluer l'imputation de la consommation d'électricité de l'Enquête sur les ménages et l'environnement » (Duddek, 2019).

Pour plus de renseignements, communiquez avec :
**Christopher Duddek** (613-862-9234, christopher.duddek@canada.ca).

**SOUS-PROJET :** Explorer l'utilisation des réseaux neuronaux pour classer automatiquement les logos sur les reçus de magasinage de l'EDM

Ce projet de recherche vise à évaluer la faisabilité de l'utilisation de reçus codés des cycles passés de l'Enquête sur les dépenses des ménages (EDM) pour former un algorithme de réseau neuronal afin de classer automatiquement les logos apparaissant sur ces reçus en fonction du nom du magasin. À l'heure actuelle, les renseignements pertinents sur les reçus de l'EDM tels que le nom du magasin, les articles achetés, la date, le total de l'achat, etc., sont tous saisis manuellement. Cependant, des méthodes de saisie automatisée pourraient potentiellement économiser du temps, des ressources et des fonds. Ce projet porte sur l'extraction automatique du nom du magasin en fonction de son logo. Deux stratégies seront mises à l'essai au moyen de réseaux neuronaux, qui sont un type d'algorithme d'apprentissage automatique. Premièrement, nous essaierons d'utiliser des techniques d'apprentissage par transfert en tirant parti d'algorithmes de réseau de neuronal formés d'avance. Ensuite, un algorithme de réseau neuronal convolutif personnalisé sera élaboré. L'ensemble de formation créé pour ce projet sera dérivé des bases de données et des reçus numérisés de l'EDM. Une technique de recadrage sera élaborée pour extraire les logos des magasins sur les reçus et une étiquette leur sera attribuée en fonction du nom codé de leur magasin figurant dans la

base de données de l'EDM. Enfin, après avoir formé les algorithmes, les résultats des deux stratégies seront évalués à l'aide d'une série test et des paramètres comme l'exactitude, la sensibilité et la spécificité. On présentera ensuite une conclusion sur la faisabilité de l'application de ces techniques pour automatiser le processus de saisie des reçus à l'avenir.

**Progrès :**

- Les reçus codés des EDM de 2015, 2016 et 2017 ont été utilisés pour créer l'ensemble de formation, de validation et de mise à l'essai. Seuls les reçus des 20 magasins les plus importants ont été conservés. Les 20 principaux magasins ont été choisis en fonction de la fréquence des reçus du magasin et du fait que leur logo était une image. Au total, nous avons recueilli plus de 40 000 logos étiquetés que nous avons répartis en ensembles de 60 %, 20 % et 20 %, pour la formation, la validation et la mise à l'essai, respectivement.
- Un algorithme de recadrage des logos a été mis au point pour recadrer automatiquement le logo dans le haut du reçu.
- Une technique utilisant un réseau neuronal préformé a été mise à l'essai, mais elle a entraîné un surapprentissage, c'est-à-dire que l'algorithme était très bon à prédire le nom du magasin pour l'ensemble de mise à l'essai, mais avait beaucoup de difficulté à prédire les noms de magasin pour les nouveaux reçus. Cette technique a été rapidement abandonnée puisque nous ne pensions pas qu'elle soit susceptible de produire des résultats de bonne qualité.
- Un réseau neuronal convolutif (RNC) personnalisé a été créé et formé. Cet algorithme prédit correctement le nom du magasin des 20 principaux reçus dans notre ensemble de mise à l'essai avec une précision de plus de 95 %. Les erreurs commises par l'algorithme étaient minimes et étaient surtout attribuables à des erreurs de classification au cours du codage manuel ou à la mauvaise qualité de certains reçus.
- L'utilisation d'un RNC personnalisé pour classer automatiquement les logos des reçus de magasinage a été jugée une réussite et démontre de premiers résultats prometteurs vers l'automatisation de la saisie de renseignements pertinents sur les reçus de l'EDM.
- Le projet a été documenté et presenté à divers groupes (Lee, 2018a) (Lee, 2018b) (Mayer, 2019a) (Mayer, 2019b).

Pour plus de renseignements, communiquez avec :
**Émilie Mayer** (613-220-1138, emilie.mayer@canada.ca).

**SOUS-PROJET :** Élaborer des indicateurs de qualité pour mesurer la précision des indices de prix

Pour les indices de prix, il est difficile de mesurer la précision des estimations pour diverses raisons. Souvent, on produit les estimations d'indices dans un cadre plus ou moins éloigné de

la méthodologie d'enquête traditionnelle que ce soit à cause du plan d'enquête utilisé ou des sources de données alternatives. Dans ce contexte, on ne peut pas toujours calculer directement les indicateurs traditionnels comme le CV. Pourtant, il y a une demande pour ce type d'information que ce soit par soucis d'informer les utilisateurs ou pour appuyer la planification à l'interne. Pour les indices de prix à la production et pour l'indice des prix à la consommation, nous souhaitons développer des indicateurs de qualités selon le cadre utilisé pour l'estimation. Nous produisons déjà des CVs et intervalles de confiance adaptés pour le plan d'enquête de certains indices. Nous souhaitons adapter cette approche à d'autres indices et développer de nouveaux indicateurs dans les cas où la méthode présente ne s'applique pas. Nous avons comme objectifs de faire une revue de littérature plus approfondie sur ce qui existe déjà et faire davantage d'exploration et de réflexion pour l'élaboration de nouveaux indicateurs.

**Progrès :**

Au cours de l'année fiscal 2018-2019, nous avons complété une revue de littérature faisant un survol des différents de cadres de qualité utilisés pour définir la qualité des indices de prix. La revue couvre des cadres utilisés à Statistique Canada et des cadres utilisés par différentes agences statistiques tel le Bureau of Labor Statistics et Eurostat, par exemple.

Nous avons porté un intérêt particulier aux indicateurs touchant à la précision des indices. Dans ce contexte, nous avons documenté différents cas où on estime une variance échantillonnale au moyen du bootstrap pour des indices de prix utilisant un plan de sondage avec probabilités proportionnelles à la taille. Nous avons calculé la variance selon l'approche proposée par Beaumont and Patak (2012). Nous avons également appliqué cette méthode d'estimation de variance probabiliste à un indice de prix basé sur un échantillon tiré à partir d'un plan non-probabiliste. Pour ce dernier, nous avons supposé différents plans et documenté les hypothèses requises.

Nous nous sommes également intéressés à l'estimation de la variance du modèle. Cette variance a une interprétation différente de la variance échantillonnale et elle présente un intérêt dans un contexte où les échantillons de prix proviennent d'un plan non-probabiliste. Nous avons documenté le processus d'estimation de la variance du modèle selon l'approche proposé par Zhang (2010). Nous avons appliqué l'approche proposée à l'indice des prix des services du commerce de détail. Nous avons également élaboré une série d'indicateurs pour évaluer la précision d'indices pour une série de domaines. Pour l'utilisation du taux de réponse, nous avons fait des simulations empiriques afin d'évaluer le risque de biais et établir des seuils d'acceptabilité.

Pour plus de renseignements, communiquez avec :
**Justin Francis** (613-863-0276, justin.francis@canada.ca)
**Jean-Sébastien Provençal** (613-513-9441, jean-sebastien.provencal@canada.ca).

**SOUS-PROJET :** Élaboration d'un prototype d'estimation robuste

Dans le cadre de procédures d'estimation classiques, la présence d'unités influentes dans l'échantillon peut avoir des répercussions négatives sur les estimations par domaine. Différents auteurs ont élaboré des méthodes d'estimation robustes pour atténuer l'instabilité de ces estimations. Dans le cadre de ce projet, nous visons à mettre au point un programme SAS pour mettre en application des méthodes permettant de minimiser l'influence estimative maximale qu'une unité peut avoir sur l'estimation robuste (Beaumont, Haziza et Ruiz-Gazen, 2013). L'influence est mesurée par le biais conditionnel. Ce programme nous permettra d'étudier l'efficacité des méthodes en les appliquant à certaines de nos enquêtes.

**Progrès :**

Nous avons créé un ensemble de macros SAS pour diverses fonctions associées à une estimation sur domaines robuste des totaux dans le cadre d'un échantillonnage aléatoire simple stratifié sans remise. Cela comprend les fonctions suivantes, qui peuvent être exécutées individuellement ou de haut en bas pour ce plan : (1) calcul des poids de sondage; (2) calcul des poids d'étalonnage à partir d'information auxiliaire pour une ou deux partitions de la population; (3) production d'estimations robustes des totaux par domaine; (4) création d'estimations par domaine cohérentes pour tenir compte des contraintes au niveau des variables; (5) calcul des poids réétalonnés pour s'assurer que les estimations pondérées standard sont identiques aux estimations robustes; (6) estimation non robuste des totaux par domaine avec estimations de la variance connexe.

Certaines fonctions, comme celle des poids d'étalonnage et celle de l'estimation de la cohérence, comportent de nombreuses options et caractéristiques afin de répondre à diverses exigences.

Chaque fonction peut être exécutée seule, pourvu que l'on fournisse les entrées nécessaires. Pour faciliter l'exécution des programmes, nous avons créé des structures de validation afin de vérifier les données d'entrée de chaque programme, de pair avec des notes, des avertissements et des erreurs.

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

## Bibliographie

Beaumont, J.-F., et Patak, Z. (2012). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *Revue Internationale de Statistique*, 80(1), 127-148.

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

Fuller, W., et Rao, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey.

Gambino, J., Kennedy, B. et Singh, M. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation.

Osiewicz, M., et Pérez-Duarte, S. (2012). Flexible variance estimation in complex sample surveys: Rescaled bootstrap in multistage, pps surveys – ÉBAUCHE.

Preston, J. (2015). Estimateur par la régression modifiée pour les enquêtes-entreprises répétées avec bases de sondage évolutives. *Techniques d'enquête*, 41, 1, 81-100. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14160-fra.pdf.

Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1962). On a simple procedure of Unequal Probability Sampling without replacement.

Zhang, L.-C. (2010). A model-based approach to variance estimation for fixed weights and chained price indices. *Official Statistics in Honor of Daniel Thorburn*, 149-166.

# 2 Activités de soutien

## 2.1 Confidentialité et contrôle de la divulgation

L'équipe a fourni un soutien et des conseils sur l'application des techniques de contrôle de la divulgation à la fois à l'interne et pour répondre à des demandes externes spécifiques.

Au cours de cet exercice, l'ajustement tabulaire aléatoire (ATA) a été utilisé pour la première fois dans les tableaux publiés de Statistique Canada. Statistique Canada a publié un ensemble complet de tableaux de sortie pour l'Enquête sur l'innovation et les stratégies d'entreprise (EISE) à l'aide de l'ATA pour protéger la confidentialité des valeurs apportées. Pour accompagner cette publication, un blogue intitulé « L'ajustement tabulaire aléatoire est arrivé! » et d'autres produits de communication ont été préparés pour annoncer cette nouvelle approche de la confidentialité. Le blog comprenait un lien vers la présentation technique de la méthode (Stinner, 2017). La méthode a également été testée avec plusieurs enquêtes différentes, à la fois comme preuve de concept des domaines d'application et pour identifier les problèmes associés à la mise en œuvre pratique des ATA.

L'équipe a également exploré l'utilisation de la synthèse de données avec une implémentation pratique pour la création d'un jeu de données à utiliser dans un concours ouvert d'analyse de données. Le travail et la méthode sont documentés dans Sallier et Girard (2018). Enfin, l'équipe a commencé à explorer le cadre de confidentialité différentiel (Girard, 2019).

Pour plus de renseignements, communiquez avec :
**Steven Thomas** (613-882-0851; steven.thomas@canada.ca).

### Bibliographie

Stinner, M. (2017). Contrôle de la divulgation et ajustement tabulaire aléatoire. Acte du recueil du Congrés annuel 2017 de la Société statistique du Canada, Winnipeg.

## 2.2 Centre de ressources en couplage d'enregistrements (CRCE)

Les objectifs du Centre de ressource en couplage d'enregistrements (CRCE) consistent à offrir des services de conseils aux utilisateurs externes et internes des méthodes de couplage d'enregistrements ce qui comprend la formulation de recommandations au sujet des logiciels et des méthodes à utiliser et des travaux concertés sur les applications de couplage d'enregistrements. Nous avons pour mandat d'évaluer diverses méthodes de couplage d'enregistrements et diverses progiciels de couplage d'enregistrements et, au besoin, de développer des prototypes de logiciels intégrant des méthodes non offertes dans les progiciels existants. Nous facilitons aussi la diffusion de l'information concernant les méthodes, le logiciel et les applications de couplage d'enregistrements aux personnes intéressées à l'intérieur et à l'extérieur de Statistique Canada.

**Progrès :**

Nous avons continué à soutenir l'équipe de développement de G-Coup et à travailler conjointement à solutionner toutes sources possibles, passées ou présentes, de corrections, de bogues ou d'améliorations pour G-Coup. Le CRCE a aussi offert un soutien aux utilisateurs internes et externes de G-Coup qui ont demandé de l'aide, ont formulé des commentaires ou ont présenté des suggestions.

En vue d'améliorer la classification de Fellegi-Sunter et de réduire le fardeau de l'examen manuel, la majorité du travail méthodologique du CRCE s'est concentré sur le l'intégration et la documentation dans la version 3.4 de G-Coup des seuils de poids automatisés avec une approche graphique et des techniques d'apprentissage automatique. La nouvelle version 3.4 de G-Coup contient plusieurs améliorations, y compris les fonctions d'importation et d'exportation, l'interface pour charger les données et les tables de conversion, d'exclusion et look-up ainsi que de nouvelles règles mixmatch de comparaison plus performantes pour des chaînes de caractères. Le CRCE a mis à jour le tutoriel et a contribué à l'élaboration du guide d'utilisateur de la version 3.4 de G-Coup.

L'inventaire des couplages d'enregistrements effectués au sein de la direction de méthodologie a été mis à jour en 2018 et les résultats présentés.

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

## 2.3 Centre de recherche et analyse en séries chronologiques (CRASC)

La recherche sur les séries chronologiques vise à maintenir un niveau élevé d'expertise et à offrir les consultations nécessaires dans ce domaine, de concevoir et de mettre à jour des outils en vue d'appliquer des solutions aux problèmes que posent les séries chronologiques dans des situations réelles, ainsi que d'étudier les problèmes courants pour lesquels il n'existe aucune solution connue ou acceptable.

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

- Consultation et formation sur les séries chronologiques (y compris l'élaboration et la prestation de formations);
- Soutien et amélioration du système de traitement des séries chronologiques;
- Désaisonnalisation et estimation de la tendance-cycle;
- Soutien relatif au logiciel G-Series pour l'étalonnage et le rapprochement;
- Modélisation et prévisions.

***Consultation et formation dans les séries chronologiques***

Dans le cadre de son mandat, le Centre de recherche et d'analyse en séries chronologiques (CRASC) offre des consultations à la demande de divers clients au sein de Statistique Canada. Les thèmes abordés le plus fréquemment durant la période de référence étaient la détection des ruptures dans les séries, l'application de la désaisonnalisation dans diverses situations (Système des comptes nationaux, estimations au niveau local, nouvelles enquêtes de la Division de la statistique du travail [DST], etc.) et les contextes particuliers pour l'étalonnage et le rapprochement. En outre, des échanges formels et informels sont effectués avec d'autres organismes statistiques (Bureau of Economic Analysis, United States Census Bureau, Bureau of Labor Statistics, Eurostat, etc.), des organisations universitaires (Université de Waterloo, Université d'Ottawa) pour collaborer et formuler des commentaires sur les sujets d'actualité.

**Progrès :**

Des lignes directrices sur la continuité des séries chronologiques ont été finalisées et entérinées par le Comité des méthodes et des normes (Statistique Canada, 2019a, 2019b).

Les experts du centre ont offert de la consultation selon les besoins, entre autres sur des mesures de volatilité, sur la desaisonnalisation de séries du flux et l'étalonnage des séries d'importation et d'exportation. Les cours exisitant ont été offerts selon la demande. Le cours nouvellement mis à jour sur la modélisation statistique et les prévisions a été offert pour une seconde fois.

L'impact sur la désaisonnalisation de l'utilisation de méthodes de perturbation pour le contrôle de la divulgation a été étudié et des suggestions ont été fournises afin d'inclure une composante d'autocorrelation dans la methode d'ajustement tabulaire aléatoire.

Les travaux recents de recherche et d'exploration ont été communiqués avec des pairs (Verret, 2018 et Matthews, 2018).

***Soutien et amélioration du système de traitement des séries chronologiques***

Ce thème inclut l'élaboration en continue et le soutien du système de traitement des séries chronologiques, actuellement utilisé dans les applications de production dans l'ensemble de l'organisme.

**Progrès :**

Un prototype de tableau de bord de désaisonnalisation a été developpé dans l'application R-shiny et sera bientôt deployé en mode pilote à quelques utilisateurs. Ce prototype facilite l'interpretation des données désaisonnalisées et indique rapidement des statistiques clés liées à ce processus.

Un module a été developpé et documenté pour l'étalonnage de stock. Une stratégie permettant de détecter la corruption de fichiers avant l'execution des modules d'extraction de tendance-cycle a été developpée.

***Désaisonnalisation et estimation de tendances-cycles :***

Ce thème traîte de l'analyse et l'évaluation de nouvelles méthodes et techniques pour la désaisonnalisation. Cela comprend des comparaisons entre les méthodes basées sur un modèle et X-12-ARIMA. Un objectif à long terme de cette recherche est de mettre au point des méthodes convenables pour l'ajustement de séries à fréquence élevée (quotidienne, horaire, etc.). On y explore aussi l'apprentissage automatique pour sélectionner des options à spécifier dans la désaisonnalisation, en vue de simplifier le soutien des processus de production, et d'accroître la capacité afin de produire rapidement des séries à volume élevé.

**Progrès :**

Une comparaison des méthodes X-12-ARIMA et SEATS a été documentée (Matthews et Dochitoiu, 2018). Des travaux préliminaires sur l'évaluation des tests de saisonnalité, et l'élaboration potentielle d'un test non paramétrique sont en cours Mischler, Lapointe et Patak (2019). D'autres projets de recherche sont en cours, entre autres sur le développement d'une stratégie pour produire des échantillons bootstrap longitudinaux afin de pouvoir estimer la variance désaisonnalisée par répliques (Verret, 2019); sur sur les modèles GARCH afin de se

pencher sur l'hétéroscédasticité saisonnière et de dériver des mesures d'irrégularité (en collaboration avec un chercheur académique) et sur une stratégie pour élaborer une désaisonnalisation préliminaire de courtes séries.

Dans le cadre du projet visant à explorer l'utilisation de méthodes d'apprentissage automatique pour la sélection des options de désaisonnalisation, un ensemble de données d'entrainement a été collecté et préparé. Une approche objective pour évaluer l'impact sur la qualité a été déterminée. Les premiers tests ont été effectués sur le paramètre de la longueur des filtres et l'approche est prometteuse.

### *Soutien relatif au logiciel G-Series pour l'étalonnage et le rapprochement*

**Description :**

Avec une publication récente de G-SERIES 2.0, l'objectif de ce travail se tourne vers le soutien et le développement futur. Un soutien est offert pour les utilisateurs pour les questions et les problèmes liés à la méthodologie. De nouvelles applications de la nouvelle fonctionnalité sont construites et mises à l'essai à mesure qu'elles sont repérées. La recherche et les plans pour la prochaine diffusion seront documentés.

**Progrès :**

L'équipe a examiné une approche d'étalonnage conditionnelle proposée par un utilisateur expérimenté et formulé des observations sur le document de travail. Cette approche sera évaluée au moyen d'une application à des données réelles et envisagée pour une implémentation dans une prochaine version de G-SERIES. Des améliorations ont également été apportées à la macro d'équilibrage afin de contrer une limitation identifiée dans SAS 9.4.

### *Modélisation et prévision*

**Description :**

Utilisation continue et accrue de SAS et d'HPF comme outil de prévision pour usage général. Actuellement utilisés dans la prévision des valeurs pour la non réponse critique (Enquête mensuelle sur les industries manufacturières [EMIM], Enquête sur les voyages internationaux [EVI]), pour repérer les ruptures dans les séries, et d'autres applications de modélisation. Exploration d'autres méthodes et outils disponibles dans SAS et d'autres progiciels produits par de tierces parties.

**Progrès :**

- L'équipe a procédé à la modélisation des données sur le tourisme comme preuve de concept afin de produire des indicateurs économiques avancés et a présenté ses

conclusions au symposium sur la méthodologie de Statistique Canada (Patak et Armenski, 2018). Ce travail a évalué plusieurs classes de modèles de séries chronologiques, l'utilité des composantes prédictives, l'automatisation de la sélection du modèle et la précision des prévisions à différents horizons.

- L'équipe a également entrepris des travaux de modélisation des données d'exportation afin d'atténuer le manque de données provenant d'une source externe (fermeture du gouvernement américain en décembre 2018). Des prévisions pour le mois à venir ont été générées pour une utilisation interne temporaire en l'absence de données de la source administrative affectée par la fermeture. Des prévisions et des indicateurs de qualité ont été générés et des consultations sur leur utilisation et leur interprétation ont été données (Leung et Dochitoiu, 2018).
- L'équipe a mené une analyse documentaire afin d'identifier les approches couramment utilisées dans les modèles de prévision de diverses organisations et les outils disponibles pour leur application. L'élaboration d'un document de directives pratiques est en cours dans le contexte du project d'exploration de l'estimation en temps quasi réel de Statistique Canada.
- L'élaboration de modèles d'espaces d'état a été mise au point par l'équipe en collaboration avec un professeur de l'Université d'Ottawa. Les travaux ont consisté à évaluer et à modifier le modèle structurel de base pour la désaisonnalisation. Une présentation a été préparée pour la conférenece *International Symposium on Forecasting* (à présenter en juin 2019).
- Des applications supplémentaires de modélisation d'espace d'états ont été développées dans le contexte de l'estimation sur petits domaines (Matthews et Dochitoiu, 2018) afin d'inclure des composantes de séries chronologiques et l'estimation des effets de rotation d'échantillon pour les plans d'enquête par panels rotatifs (Dochitoiu, 2018) ont été développées pour une évaluation ultérieure.

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

## 2.4 Secrétariat de la qualité

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

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

**Progrès :**

Un plan de rédaction a été présenté au Comité des méthodes et des normes de Statistique Canada, qui l'a approuvé. Un comité de lecture a formulé des commentaires et des suggestions à propos de la première ébauche, ce qui a été suivi de consultations ciblées auprès de certaines divisions de Statistique Canada afin de recueillir des commentaires sur des domaines de spécialisation précis. Ces commentaires ont été intégrés à l'ébauche de la nouvelle version, qui comprendra également des éléments liés à la protection des renseignements personnels, la transparence et l'éthique. La nouvelle édition sera diffusée en 2019-2020.

**SOUS-PROJET :** Travaux de recherche et de renforcement des capacités avec des partenaires internationaux

Le Secrétariat de la qualité a fourni des conseils et a pris des mesures de renforcement des capacités à l'intention de partenaires internationaux, principalement en présentant un aperçu général des pratiques de gestion de la qualité de Statistique Canada et des documents officiels liés à la qualité (Cadre d'assurance de la qualité et Lignes directrices concernant la qualité).

**Progrès :**

En 2018-2019, nous avons présenté des exposés à des délégations venant d'Indonésie et du Cameroun et nous avons dirigé un atelier d'une semaine à Santo Domingo, en République dominicaine. Lors de nos réunions avec des visiteurs internationaux, nous avons fait des présentations sur des sujets qui présentent un intérêt pour eux, plus précisément la gestion de la qualité et le Modèle générique du processus de production statistique. L'atelier tenu à Santo Domingo a aidé des participants de plus d'une douzaine de pays d'Amérique latine à préparer une feuille de route en vue de l'élaboration d'un cadre d'assurance de la qualité. Au cours de chacune de ces initiatives de collaboration, nous avons eu des discussions sur les défis modernes reliés à la qualité des données dans les organismes nationaux de statistique, notamment sur la façon d'indiquer la qualité dans le cas des données statistiques qui ne sont pas recueillies dans le cadre d'enquêtes par sondage.

Des travaux de recherche sur le developpement de nouvelles mesures ou indicateurs de qualité dans un contexte de données mixtes ont également débuté Reedman (2018) et Reedman et Windross (2018).

**SOUS-PROJET :** Participation aux travaux du Groupe d'experts des Nations Unies sur les cadres nationaux d'assurance de la qualité

Le Secrétariat de la qualité a assuré la coprésidence du Groupe d'experts des Nations Unies sur les cadres nationaux d'assurance de la qualité et a contribué à la préparation d'un cadre national d'assurance de la qualité actualisé en communiquant du contenu de notre cadre interne d'assurance de la qualité.

**Progrès :**

Une version provisoire du Cadre national d'assurance de la qualité (CNAQ) des Nations Unies actualisé a été rédigée et publiée en novembre 2018 à des fins d'examen par les pairs. Des recommandations ont ensuite été mises en œuvre, et cette version a été adoptée par la Commission de statistique des Nations Unies à New York en mars 2019. Le Secrétariat de la qualité a coprésidé l'activité parallèle de la Commission visant à discuter du CNAQ.

Pour plus de renseignements, communiquez avec :
**Ryan Chepita** (613-851-5340; ryan.chepita@canada.ca).

## 2.5 Centre de ressources en analyse de données

Le Centre de ressources en analyse de données (CRAD) est une équipe de statisticiens-conseils et de chercheurs au sein de la Direction de la méthodologie. Le CRAD a pour principaux objectifs de donner des conseils sur l'emploi approprié des outils et des méthodes d'analyse de données, et de promouvoir l'adoption de pratiques exemplaires en la matière. Les services du CRAD, axés principalement sur les données d'enquête, les données de recensement et les données administratives, sont offerts aux employés de Statistique Canada et à ceux d'autres ministères, ainsi qu'aux analystes et aux chercheurs du milieu universitaire ou des centres de données de recherche (CDR).

**Progrès :**

*Consultations*

Dans le cadre de son mandat, le CRAD a offert des services de consultation sur demande à divers clients, y compris à des analystes d'une douzaine de divisions de Statistique Canada. Les consultations couvraient, entre autres, des sujets relatifs à l'établissement d'intervalles de confiance, à la mise à l'essai d'hypothèses sur les différences entre les sous-populations, à l'analyse au moyen de cycles d'enquête combinés et à l'analyse de données de survie. Nous avons reçu plusieurs demandes concernant la comparaison de quantiles ou de leurs fonctions, un sujet qui est très rarement discuté dans les ouvrages publiés. Pour la comparaison des taux de croissance du patrimoine médian, nous avons proposé quelques méthodes, et l'une d'entre elles a été utilisée. De plus, nous avons aidé les analystes de Statistique Canada à intégrer les méthodes dans les logiciels R, SUDAAN, SAS SURVEY et STATA.

Nous avons également fourni des services à d'autres méthodologistes. Les consultations portaient, entre autres, sur les méthodes d'apprentissage automatique, l'autocodage, la classification de texte, les degrés de liberté de l'estimation de la variance et l'élaboration de modèles de régression. Nous avons aussi créé des exemples à partir de STATA pour un atelier sur l'analyse au moyen de données de l'ESG.

Des consultations ont également été offertes à l'extérieur de Statistique Canada, soit à divers clients d'autres ministères et organismes fédéraux et provinciaux. Les demandes portaient notamment sur l'analyse à partir d'échantillons non aléatoires, l'évaluation de la stabilité de l'analyse en composantes principales selon des données d'enquête et l'examen méthodologique d'une étude portant sur la comparaison de promotions entre des groupes.

Finalement, le groupe a donné des conseils spécialisés aux analystes et aux chercheurs des CDR. Les sujets comprenaient la combinaison de cycles d'enquête, l'estimation d'erreurs-types des estimations des recensements de 1991 à 2006 et l'utilisation de poids bootstrap pour effectuer l'analyse de données d'enquête.

*Prestation de formation*

L'équipe a présenté des séminaires portant sur l'analyse de données tirées d'une enquête à plan d'échantillonnage complexe et offert les cours du programme lié à leur champs d'expertise, incluant un cours sur l'analyse des données de survie. Des présentations spéciales à des fins de formation ont été developéée et présentées entre autres sur les cartes thermographiques, l'apprentissage automatique et l'utilisation adéquate de la valeur *p*.

*Collaboration avec les analystes*

L'article « Répartition du temps entre le sommeil, la sédentarité et l'activité : liens avec l'obésité et la santé chez les adultes canadiens », par Rachel Colley, Isabelle Michaud et Didier Garriguet, a été publié dans le numéro d'avril de *Rapports sur la santé* (*Rapports sur la santé*, vol. 29, nº 4, avril 2018, produit nº 82-003-X au catalogue de Statistique Canada).

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

## 2.6 Centre de ressources en conception de questionnaires (CRCQ)

Le Centre de ressources en conception de questionnaires (CRCQ) de la Direction de la méthodologie est le centre d'expertise de Statistique Canada en matière de conception et d'évaluation de questionnaires. 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 offre aussi des cours sur la conception de questionnaires.

**Progrès** :

Dans le cadre des initiatives de modernisation de Statistique Canada, le groupe a exploré diverses manières d'utiliser la technologie et des méthodes modernes pour optimiser certains de ses travaux opérationnels. Les résultats ont été résumés dans un exposé intitulé *Modernizing questionnaire testing - Using technology to find efficiency (while maintaining quality)* et présenté lors de l'atelier QUEST, une conférence internationale d'experts (Solomon, 2018). Bien que l'étude ait surtout mis l'accent sur les avantages persistants d'un véritable test en personne, elle a également permis d'identifier des cas ciblés ou des tâches secondaires pouvant être effectuées à distance ou avec un soutien local.

Le groupe a également contribué à diverses initiatives de consultation d'entreprise.

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

## 2.7 Transfert de connaissances – Formation en statistique

De nouvelles stratégies permettant d'atteindre notre objectif de renforcement des capacités et de développement des talents ont graduellement été mises en oeuvre. D'une manière générale, le curriculum est désormais divisé en blocs thématiques sous la responsabilité des centres de ressources appropriés (et les activités correspondantes sont rapportées dans leur section respective). Pour les méthodes d'enquête de base, des nouvelles versions de plusieurs des cours ont été développées. Une nouvelle version du cours d'échantillonnage incluant une présentation à faire par les participants a été offerte. Un nouveau cours sur l'estimation robuste a été offert. Un cours pilote d'introduction à la science des données a été ajouté au curriculum et enseigné par un professeur reconnu d'université. Un cours sur les estimations sur petits domaines est en développement. Les nouvelles stratégies de formation ont été résumées et discutées lors d'un panel au *Joint Statistical Meetings* (Fortier, 2018).

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

## 2.8 Transfert de connaissances – *Techniques d'enquête*

Techniques d'enquête est une revue internationale, disponible sur Internet à l'adresse www.statcan.gc.ca/Techniquesdenquete, dans laquelle sont publiés, dans les deux langues officielles, des articles sur divers aspects des méthodes statistiques susceptibles d'intéresser les organismes statistiques. Le comité de rédaction est formé de chefs de file de renommée mondiale du domaine des méthodes d'enquête issus des secteurs public, universitaire et privé. La revue est diffusée en format HTML pleinement accessible et en format PDF.

Les travaux associés aux processus de rédaction et de production comprennent la correspondance avec les auteurs, les examinateurs, les rédacteurs associés et les abonnés, l'examen des commentaires des examinateurs et des révisions des auteurs, le reformatage des manuscrits, la révision des manuscrits, la liaison avec les services de traduction et de diffusion, et la tenue à jour d'une base de données sur les articles soumis. Ces travaux font partie des activités de transfert des connaissances.

**Progrès :**

Les numéros de juin et décembre 2018 (44-1 et 44-2) ont été diffusés en versions PDF et HTML. La revue Techniques d'enquête de juin 2018 comprend 7 articles réguliers. Le numéro de décembre 2018 a été dirigé par deux redacteurs en chef invités Jean-François Beaumont et David Haziza. Il comprend 10 articles sélectionnés parmi l'ensemble des communications présentées lors du 9e Colloque francophone sur les sondages qui s'est déroulé à Gatineau du 11 au 14 octobre 2016.

D'avril 2018 à mars 2019, les pages de Techniques d'enquête ont été visionnées 27 000 fois et près de 6 000 copies d'articles ont été téléchargées selon une méthodologie améliorée de métriques web. En tout et en excluant les numéros spéciaux, 31 articles ont été soumis pour publication.

En 2019, la diffusion de 3 numéros de la revue est prévue. En plus des deux numéros régulier, un numéro spécial mettant en valeur certains articles présentés lors d'une conférence intitulée « Contemporary Theory and Practice in Survey Sampling: A Celebration of Research Contributions of J.N.K. Rao » sera publié en collaboration avec la Revue internationale de Statistique.

Pour plus de renseignements, communiquez avec :
**Susie Fortier** (613-863-9135, susie.fortier@canada.ca).

# 3 Documents de recherche parrainés par le Programme de recherche et développement en méthodologie

Beaumont, J.-F. (2019). Are probability surveys bound to disappear for the production of official statistics? *Techniques d'enquête* (soumis).

Bocci, C., et Beaumont, J.-F. (2018). Report on Small Area Estimation for the Global Affairs Canada Contract. Document de travail, Direction de la méthodologie, Statistique Canada.

Bocci, C., et Beaumont, J.-F. (2019). Small area estimation methodology applied to the Monthly Survey of Manufacturing-UPDATE. Document de travail, Direction de la méthodologie, Statistique Canada.

Chatrchi, G., Beaumont, J.-F., Gambino, J. et Haziza, D. (2018). An investigation into the use of sample matching for combining data from probability and non-probability samples. Proceedings of the Survey Methods Section, Statistical Society of Canada.

Chu, K., et Beaumont, J.-F. (2019). Formation of Homogeneous Self-Selection Propensity Classes for Non-Probability Samples via Probability Samples, SSC.

Chu, K., Yeung, A. et Dasylva, A. (2018). *Census Secondary Mother Tongue Write-in Autocoding: Prototype 2*. Présentation en PowerPoint.

Dasylva, A. (2018). Design-based estimation with record-linked administrative files and a clerical review sample. *Journal of Official Statistics*, 34, 41-54.

Dasylva, A. (2019). Autocoding as Record Linkage. Document de travail.

Dasylva, A., Goussanou, A., Ajavon, A. et Abousaleh, H. (2019). Revisiting the Probabilistic Method of Record Linkage. Document de travail.

Devin, N. (2018). Validation de l'extension au Bootstrap de Rao-Wu proposée par Pérez-Duarte applicable aux échantillons tirés selon des plans proportionnels à la taille à un ou deux degrés à l'aide d'une simulation Monte Carlo. Document de travail.

Dochitoiu, C. (2018). Modelling Sample Rotation Effects through unobserved components. Document de travail.

Duddek, C. (2018). Combiner les données du recensement et d'Hydro-Manitoba pour comprendre la consommation d'électricité résidentielle. Recueil : Symposium 2018, Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données, Statistique Canada.

Duddek, C. (2019). Using Manitoba Hydro data to evaluate imputation of electricity consumption in the Households and the Environment Survey, présentation au comité technique de la Division des méthodes de la statistique sociale (DMSS).

Fortier, S. (2018). Modern Statistical Training Program for a National Statistical Office. Présentation d'un panel à la 2018 Joint Statistical Meetings, Vancouver.

Fortier, S., et Thomas, S. (2018). (R)evolution of generalized systems and statistical tools at Statistics Canada, présenté à la *6th Conference on the Use of R in Official Statistics,* La Haye.

Girard, C. (2019). Making Head or Tail of Differential Privacy. Séminaire à l'interne.

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

Lee, B. (2018a). Receipt Logo Recognition Using Neural Network Algorithm, Simon Fraser University rapport de stage coopératif.

Lee, B. (2018b). Receipt Logo Recognition Using Neural Network Algorithm, présentation à l'interne à la Division des méthodes de la statistique sociale (DMSS).

Lesage, É., Beaumont, J.-F. et Bocci, C. (2019). Deux diagnostiques locaux pour évaluer l'efficacité de l'estimateur composite issu du modèle de Fay-Herriot. Ébauche, Statistique Canada.

Leung, J., et Dochitoiu, C. (2018). Use of time series forecasts as a contingency in the absence of expected data – Application to international trade aggregates. Document de travail.

Matthews, S. (2018). Current Challenges with Quality Assurance of Seasonal Adjustment. Présenté à la Second Seasonal Adjustment Practitioners Workshop, Washington, D.C.

Matthews, S., et Dochitoiu, C. (2018). Comparison of Seasonal Adjustment Approaches through State Space Representation. Document de travail.

Matthews, S., et Dochitoiu, C. (2018). State Space Modelling for Small Area Estimation with Time Series Components. Document de travail.

Mayer, É. (2019a). Using Convolutional Neural Networks to Automatically Classify Logos on Shopping Receipts (MLCoP), présentation à la communauté de pratique sur l'apprentissage automatique de Statistique Canada.

Mayer, É. (2019b). Using Convolutional Neural Networks to Automatically Classify Logos on Shopping Receipts, présentation au Symposium "Data Science and Statistics".

Mischler, L., Lapointe, M.A. et Patak, Z. (2019). A non-parametric test for detecting seasonality. Rapport de stage coopératif.

Neusy, E. (2019a). Reporting Quality Using Confidence Intervals. Présentation interne à la Special Surveys Division Managers Meeting.

Neusy, E. (2019b). Wilson Confidence Intervals for Weighted Counts. Document de travail.

Neusy, E., Boulet, C., Duggan, J., Mach, L., Mantel, H. et Reedman, L. (2018). Quality-based Release Criteria for Social Statistics, présentation au comité technique de la Division des méthodes de la statistique sociale (DMSS).

Patak, Z., et Armenski, T. (2018). Should I stay or should I go: A cross-border traveller's tale, Statistics Canada Symposium on Statistical Methodology, va paraitre dans le recueil de la conférence.

Reedman, L. (2018). A modest attempt at communicating about Quality. Présenté à la 2018 European Conference on Quality in Official Statistics, Krakow, Pologne.

Reedman, L., et Windross, M. (2018). *The NSO, the NSS and Beyond!* Présenté à la 2018 European Conference on Quality in Official Statistics, Krakow, Pologne.

Saïdi, A., Chu, K., Dasylva, A. et Labrecque-Synnott, F. (2018). Analysis of Binary Data Obtained by a Probabilistic Record Linkage, CANSSI Conference, Montréal.

Saïdi, A., Chu, K., Dasylva, A. et Labrecque-Synnott, F. (2018). Logistic regression with Linked Data, FCSM Conference, Washington.

Sallier, K., et Girard, C. (2018). Towards a successful implementation of Synthesis in a National Statistical Agency: A model for cooperation. Présenté à la 2018 Privacy in Statistical Databases conference, organisé par l'Unesco Chair in Data Privacy, Valencia, Espagne.

Savard, S.-A. (2019). Une approche de codage automatique avec données d'apprentissage basée sur la méthodologie du couplage d'enregistrements. Document de travail.

Solomon, J. (2018). Modernizing questionnaire testing - Using technology to find efficiency (while maintaining quality). Présenté à la QUEST Workshop, Allemagne.

Statistique Canada (2019a). Directive on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics Programs. Document de travail.

Statistique Canada (2019b). Guidelines on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics. Document de travail.

Verret, F. (2018). Some discussions on calendar effects in X12-ARIMA. Présenté à la Second Seasonal Adjustment Practitioners Workshop, Washington, D.C.

Verret, F. (2019). Variance Estimation for Seasonally Adjusted Estimates. Présenté à l'Economic Statistical Methods Division technical committee.

Yeung, A., Chu, K., Laroche, R. et Fortier, S. (2018). Exploring modern coding methods. Document interne. Présenté au Comité consultatif sur les méthodes statistiques de Statistique Canada.

You, Y. (2018). Area level modeling approaches to small area estimation using R and S-Plus with applications. Présentation divisionnaire du Centre de collaboration internationale et d'innovation en méthodologie (CCIIM).

You, Y. (2019). Hierarchical Bayes small area estimation of LFS status using linear and non-linear area level models. Rapport de recherche du Centre de collaboration internationale et d'innovation en méthodologie (CCIIM).