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

Programme de recherche et développement en méthodologie : réalisations en 2011-2012

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.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Programme de recherche et développement en méthodologie : réalisations en 2011-2012

Ce rapport fait la synthèse des réalisations de 2011-2012 du Programme de recherche et développement en méthodologie (PRDM) parrainé par la Direction de la méthodologie de Statistique Canada. Ce programme englobe les travaux 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 travaux qui, autrement, ne seraient pas entrepris 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 (Programme de recherche et développement en méthodologie) comprend des activités de soutien aux clients pour la mise en application de travaux de développement antérieurs fructueux. Des renseignements supplémentaires sur tout projet décrit peuvent être obtenus auprès de la personne-ressource mentionnée. Pour en savoir davantage sur le PRDM (Programme de recherche et développement en méthodologie) dans son ensemble, communiquez avec :

Mike Hidiroglou
(613-951-0251, mike.hidiroglou@statcan.gc.ca).

Projets de recherche

Recherche, développement et consultation à la DRIS (Division de la recherche et de l'innovation en statistique)

La Division de la recherche et de l'innovation en statistique (DRIS) a été créée au sein de la Direction de la méthodologie le 21 juin 2006. La DRIS (Division de la recherche et de l'innovation en statistique) est responsable de la recherche, de l'élaboration et de la promotion de techniques novatrices en méthodologie statistique, ainsi que de la surveillance et de l'encadrement de l'adoption de ces techniques en vue d'appuyer les programmes statistiques de Statistique Canada. Son mandat comprend aussi l'offre d'un leadership technique, de conseils et d'un encadrement aux employés des autres secteurs du Programme de recherche et développement en méthodologie. Ce soutien prend la forme de conseils sur les problèmes méthodologiques que posent les projets en cours ou l'élaboration de nouveaux projets.

La DRIS (Division de la recherche et de l'innovation en statistique) travaille aussi de concert avec d'autres employés à des projets de recherche parrainés par le Programme de recherche et développement en méthodologie qui portent sur des sujets particuliers, comme les méthodes d'estimation, les méthodes d'imputation, l'estimation sur petits domaines, l'utilisation des données administratives, la collecte des données et les méthodes applicables aux séries chronologiques.

En 2011-2012, la DRIS (Division de la recherche et de l'innovation en statistique) a participé à de nombreux projets de recherche, de développement et de consultation. La contribution de ses employés a été importante notamment en ce qui concerne l'estimation sur petits domaines, l'imputation et l'estimation robuste, la production de données synthétiques et les techniques applicables aux séries chronologiques. Des renseignements détaillés sur les progrès accomplis sont présentés dans l'examen des sujets de recherche, plus loin dans le rapport.

En plus de participer aux activités de recherche du Programme de recherche et développement en méthodologie (PRDM) à titre de chefs de projet et de chercheurs, les employés de la DRIS (Division de la recherche et de l'innovation en statistique) ont pris part aux activités suivantes :

- Le personnel a donné des conseils aux membres d'autres divisions sur des questions techniques de façon ponctuelle et de façon officielle. Les conseils ponctuels portaient, par exemple, sur l'estimation de la variance dans le cas d'enquêtes complexes, la coordination des échantillons, l'estimation sur petits domaines, l'estimation de la variance en présence de données imputées et les méthodes applicables aux séries chronologiques.
- La DRIS (Division de la recherche et de l'innovation en statistique) a consulté les membres du Comité des méthodes et des normes, ainsi qu'un certain nombre d'autres gestionnaires de Statistique Canada pour établir les priorités du programme de recherche.

- La DRIS (Division de la recherche et de l'innovation en statistique) a continué de soutenir activement la revue *Techniques d'enquête*. Mike Hidiroglou est rédacteur en chef de la revue depuis janvier 2010. Cinq employés de la DRIS (Division de la recherche et de l'innovation en statistique) contribuent aussi à la revue : un à titre de rédacteur associé et trois autres à titre de rédacteurs adjoints; enfin, Céline Ethier,

gestionnaire adjointe de la production, est responsable de la composition de la revue dans les deux langues officielles.

- Les employés ont poursuivi leurs activités au sein de divers comités de la Direction de la méthodologie, dont le Comité de l'apprentissage et du perfectionnement, le Groupe de la réflexion stratégique et le Comité de l'informatique. En particulier, ils ont participé activement à la recherche et à la discussion des articles du mois.
- La DRIS (Division de la recherche et de l'innovation en statistique) a présenté un article au Comité consultatif des méthodes statistiques (Choudhry, Hidiroglou et Laflamme, 2011).
- Durant le Symposium 2011, des membres de la DRIS (Division de la recherche et de l'innovation en statistique) ont donné, en collaboration avec le professeur J.N.K. Rao, un cours d'un jour sur les travaux de développement courants concernant l'estimation sur petits domaines et leur mise en œuvre informatique.
- Deux cours distincts ont été donnés par des employés de la DRIS (Division de la recherche et de l'innovation en statistique), en collaboration avec les membres d'autres divisions de la méthodologie, au 58e congrès de l'Institut international de statistique, à Dublin, en Irlande. L'un, donné par Mike Hidiroglou et Wesley Yung, avait pour thème les méthodes d'enquête auprès des entreprises et l'autre, donné par John Kovar, Eric Rancourt et Jean-François Beaumont, était un atelier sur la vérification et l'imputation des données d'enquête.
- Jean-François Beaumont a donné un exposé sur le SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) et a commenté deux exposés sur l'estimation de la variance des mesures de variation durant l'atelier sur l'estimation de la variance pour le projet EU-SILC (European Union-Statistics on Income and Living Conditions)2 qui s'est tenu en mars 2012 à Luxembourg.
- Mike Hidiroglou a participé au groupe consultatif de la National Science Foundation – Census Research Network. Il a donné des conseils sur le choix de quelque 30 propositions de recherche présentées au U.S. Bureau of the Census.
- Mike Hidiroglou a présenté deux communications sollicitées au deuxième atelier international sur la recherche pour l'évaluation des politiques publiques qui était un événement satellite de la troisième école sur l'échantillonnage et la recherche. Ces réunions ont eu lieu à l'Institut des sciences exactes de l'Université fédérale de Juiz de Fora, au Brésil, à la fin de novembre 2011.
- Les employés de la DRIS (Division de la recherche et de l'innovation en statistique) ont évalué à titre d'examinateurs plusieurs articles soumis pour publication dans des revues statistiques.
- Les employés de la DRIS (Division de la recherche et de l'innovation en statistique) ont rédigé ou corédigé 51 articles (qui ont été publiés, présentés à des conférences ou diffusés sous forme de rapports). Nombre de ces articles ont été présentés à des conférences telles que le congrès de la Société statistique du Canada et les Joint Statistical Meetings. Ces exposés ont été publiés dans les *actes du congrès de la Société statistique du Canada* et dans les *Proceedings of the Joint Statistical Meetings,* ou publiés dans des revues statistiques savantes, dont *Techniques d'enquête*, *Biometrika, Revue Internationale de Statistique, International Journal of Forecasting* et *Pakistan Journal of Statistics.*

Pour obtenir plus de renseignements, veuillez communiquer avec :
Mike Hidiroglou (613 951-0251, mike.hidiroglou@statcan.gc.ca).

Vérification et imputation

Un volume considérable de travaux de recherche portant sur plusieurs sujets liés à la vérification et à l'imputation ont été entrepris ces dernières années. Toutefois, un très petit nombre seulement des idées les plus novatrices sont mises en œuvre dans les programmes statistiques. Cette situation tient notamment au fait que i) souvent, les méthodologistes responsables des programmes des méthodes statistiques n'ont pas le temps de lire et de comprendre les découvertes les plus récentes, ii) ces nouvelles idées ne sont intégrées dans aucun progiciel statistique sur le marché et iii) dans un contexte de production, le temps manque pour développer les logiciels appropriés.

L'objectif principal du projet est donc de soutenir le développement et la maintenance d'outils informatiques accompagnés de documentation qui mettent en application des idées de recherche susceptibles de profiter aux programmes statistiques. Ce soutien aidera à faire en sorte que les programmes de Statistique Canada développent et appliquent des théories et des méthodes statistiques de pointe, ce qui correspond en fait au mandat de la Direction de la méthodologie.

Il convient de mentionner que l'objectif du projet n'est pas de développer des systèmes généralisés. Une fois qu'un outil informatique arrive à maturité et est utilisé par de nombreux programmes statistiques, son intégration dans la famille de systèmes généralisés devrait être envisagée.

SEVANI – Recherche et développement sur l'estimation de la variance en présence d'imputation

Le SEVANI est le Système d'estimation de la variance due à la non-réponse et à l'imputation. Durant l'année courante, nous avons fait progresser le développement du système et la recherche connexe. Du côté du développement :

1. Nous avons achevé le développement d'un prototype qui peut être utilisé pour estimer la variance lorsqu'on estime un ratio. La version actuelle de SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) ne traite que les totaux et les moyennes. Ce prototype a été développé pour le programme de l'Enquête sur les voyages des résidents du Canada. Il repose sur l'hypothèse que, si elles manquaient, les variables du numérateur et du dénominateur ont été imputées simultanément par la méthode d'imputation par donneur en se servant du même donneur pour les deux variables.
2. PISE : Des rencontres ont eu lieu pour discuter de l'utilisation de SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) dans PISE. Deux requêtes exprimées par PISE demandent que SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) fournisse des informations qui ne sont pas disponibles dans la version actuelle. Une première version modifiée a été mise au point pour satisfaire à la première requête. L'implémentation de la deuxième requête devrait se faire dans les prochaines semaines.
3. Intégration dans G-EST : Des rencontres ont eu lieu pour planifier la fusion du Système généralisé d'estimation (SGE) et de SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation). Une notation commune pour l'écriture de spécifications pour la programmation a été suggérée. Le gros du travail reste à faire.
4. Nouvelle version en bêta : au cours de la dernière année, plusieurs développements ont été entrepris.
    1. Plusieurs études ont été faites pour améliorer les approches non paramétriques des calculs des espérances et variances par rapport au modèle d'imputation. En effet, la procédure TPSPLINE donnait parfois des résultats erratiques ou encore exigeait trop de temps-machine.
    2. De l'optimisation a été apportée pour améliorer la performance. Ce travail a été déclenché par des discussions avec un utilisateur qui

prévoyait une durée d'exécution de plus d'une semaine avec la version actuelle. Les améliorations ont ramené la durée d'exécution à moins d'une demi-heure.

3. Une façon plus flexible de préciser les domaines d'intérêt a été implémentée.

Il convient de souligner que les travaux de développement susmentionnés ont été entrepris à la demande des utilisateurs. Nous avons également continué d'offrir un soutien à nos utilisateurs en ce qui concerne la méthodologie et l'utilisation du système.

Du côté de la recherche :

- Nous avons préparé un exposé sollicité sur la méthodologie et l'utilisation du SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) qui a été donné à la fin de mars 2012 à l'atelier sur l'estimation de l'erreur-type et d'autres questions liées à l'échantillonnage dans le cadre des statistiques européennes sur le revenu et les conditions de vie (EU-SILC (European Union-Statistics on Income and Living Conditions)), qui était organisé dans le contexte du projet « Net-SILC (Statistics on Income and Living Conditions)2 » financé par l'Union européenne.
- Notre article sur la méthodologie du SEVANI (Système d'estimation de la variance due à la non-réponse et à l'imputation) pour l'imputation composite (Beaumont et Bissonnette, 2011) a été publié dans *Techniques d'enquête* et notre article sur l'imputation par valeur auxiliaire (Beaumont, Haziza et Bocci, 2011a) a été publié dans *Statistica Sinica.*

Outil de détection des valeurs aberrantes

Un outil graphique pour la détection de valeurs aberrantes a été développé en SAS au cours des cinq dernières années. Cet outil contient huit méthodes qui sont les plus usuelles à Statistique Canada. Son utilité est d'aider les méthodologistes à identifier la méthode de détection de valeurs aberrantes qui est la plus appropriée pour leurs enquêtes, à optimiser les paramètres d'une méthode spécifique de détection et à comparer les méthodes entre elles.

Au cours de cette année, une variante de la méthode de l'écart-sigma a été développée pour la comparer avec les résultats de la méthode de l'écart-sigma déjà en place. Cette variante, basée sur la différence entre deux unités, est celle qui est utilisée par l'enquête du commerce de gros. Il fallait vérifier si les deux versions donnaient des résultats similaires. De plus, on agit toujours comme consultant pour l'implantation de la méthode de l'écart-sigma dans le système généralisé BANFF.

On a fait une présentation de cet outil dans le cadre du symposium international de 2011 sur les questions de méthodologie. À la suite de cette présentation, il y a eu beaucoup d'intérêt pour cet outil tant à l'intérieur de Statistique Canada qu'à l'extérieur. Une présentation a été donnée à la division de la démographie qui évalue la possibilité d'utiliser une/des méthode(s) univariée(s). Une consultation a eu lieu avec l'équipe de l'EUE pour savoir si on pouvait améliorer leur méthodologie de détection des valeurs aberrantes en employant cet outil. Les méthodologistes de l'enquête mensuelle des manufactures sont intéressés à la méthode Hidiroglou-Berthelot pour des analyses. Des présentations sont à venir pour d'autres sections en méthodologie.

Il y a une étude en cours pour évaluer la possibilité d'utiliser les capacités graphiques de JMP au lieu de SAS Graph. Cela permettrait d'améliorer l'interaction graphique avec l'utilisateur de façon significative. Malgré les avantages évidents, il n'est pas encore certain que l'utilisation de JMP soit la meilleure alternative puisqu'il faudrait refaire une partie de la programmation, avoir deux environnements (SAS et JMP), s'assurer que la licence de JMP soit renouvelée, *etc. (et ainsi de suite)*

Un guide de référence est amorcé. Ce guide expliquera la marche à suivre pour utiliser l'outil et la théorie sous-jacente des méthodes disponibles.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Jean-François Beaumont (613 951-1479, jean-francois.beaumont@statcan.gc.ca).

Échantillonnage et estimation

Ce sujet regroupe les projets de recherche suivants :

- Le sondage indirect appliqué aux populations asymétriques
- Le bootstrap moyen
- Calage pour l'estimation de variance
- Coordination des échantillons
- Méthode des microstrates pour la coordination des échantillons
- Ajustement de la non-réponse par le calage aux marges
- Generalized bootstrap

Le sondage indirect appliqué aux populations asymétriques

Il est possible de produire des estimations à l'échelle des entreprises à partir d'un échantillon à l'échelle des établissements en utilisant la Méthode généralisée du partage des poids (MGPP) (Lavallée, 2002, 2007). On sait toutefois que les établissements constituent une population asymétrique. En appliquant la MGPP (Méthode généralisée du partage des poids) classique à une telle population, on peut obtenir des variances très élevées. Le but du projet était de proposer une alternative à la façon de partager les poids de façon à réduire la variance. On a proposé huit méthodes différentes de procéder, qui sont toutefois largement basées sur une pondération économique des liens entre la base de sondage (population d'établissements) et la population cible (population d'entreprises constituées d'établissements).

Le projet de recherche en soi est terminé. La comparaison des huit méthodes alternatives pour réduire la variance de la MGPP (Méthode généralisée du partage des poids) classique est terminée. L'analyse a été concluante : une série de méthodes ressortent du lot, en particulier celles avec l'utilisation des liens faiblement optimaux proposés par Deville et Lavallée (2006). Ces dernières donnent des variances de beaucoup inférieures à la MGPP (Méthode généralisée du partage des poids) originale. Le projet et les résultats ont été présentés au congrès de la Société statistique du Canada en juin

2011 de même qu'aux Journées de méthodologie statistique de l'INSEE (Institut National de la Statistique et des Études Économiques) (France) en janvier 2012. Un article résumant le projet a été rédigé et soumis à *Techniques d'enquête*.

Le bootstrap moyen

Le bootstrap moyen est une méthode de rééchantillonnage similaire au bootstrap standard de Rao-Wu. Au cours de l'année précédente, nous avions observé que le bootstrap moyen et le bootstrap de Rao-Wu ne fonctionnaient pas bien pour la médiane d'une variable discrète, comme le nombre de nuits passées à l'extérieur du pays, même pour une grande taille d'échantillon. Nous avons donc mieux étudié le problème de l'estimation de la médiane d'une variable discrète. Premièrement, nous avons confirmé que le problème disparaissait pour une distribution continue telle que la gamma. Ensuite, nous avons réalisé que la médiane devait être définie à partir d'une fonction de répartition empirique qui converge asymptotiquement vers une fonction continue (et lisse). Pour une variable discrète, ce n'est pas le cas. Le problème peut être réglé en faisant une interpolation linéaire entre les points de la fonction de répartition empirique. Des simulations indiquent que l'interpolation linéaire donne des résultats satisfaisants bien que des méthodes plus sophistiquées d'interpolation ou de lissage pourraient donner de meilleurs résultats. Nous sommes présentement en train de compléter les simulations et l'écriture d'un article.

Calage pour l'estimation de variance

Dans ce projet, la variance est considérée comme un total pondéré. Pour une population de taille N et une variable d'intérêt y_i i = 1, 2, …, N, la variance est une somme pondérée de z_k k = 1, 2, …, N^2 avec $z_k = y_i\, y_j$ si $k = ri + ji$, j =1, 2, …, N. Le calage pourrait être utilisé pour estimer la somme pondérée des z_k. Sous calage, la distance entre le vecteur des poids calés des unités dans un échantillon s, $\mathbf{w}_s$, et le vecteur des poids d'Horvitz-Thompson, $\mathbf{a}_s$, est minimisée (sous une contrainte faisant intervenir des données auxiliaires). Cette distance peut être exprimée par $(\mathbf{w}_s - \mathbf{a}_s)'\, \mathbf{U}_s\, (\mathbf{w}_s - \mathbf{a}_s)$, où $\mathbf{U}_s$ est une matrice diagonale positive. Comme les z_k sont corrélés, nous devrions généraliser le calage de façon que la matrice $\mathbf{U}_s$ soit semi-définie positive, mais pas forcément diagonale. Dans les conditions générales de régression, nous dirions que nous voulons généraliser l'estimateur au cas où la matrice de covariance du vecteur d'intérêt, sous le modèle, n'est pas nécessairement diagonale.

On a défini un cadre asymptotique utilisant un modèle avec une matrice de variance-covariance non nécessairement diagonale. Ce cadre asymptotique est défini de façon à ne plus avoir besoin d'une superpopulation. On a généralisé l'estimateur par calage au cas de $\mathbf{U}_s$ définie positive. On a établi des conditions sous lesquelles cet estimateur généralisé est optimal, en ce sens qu'une borne inférieure est atteinte. Cette borne est une généralisation de celle de Godambe et Joshi (1965).

Il faut généraliser l'estimateur et la borne inférieure au cas de $\mathbf{U}_s$ semi-définie positive. Il faudrait démontrer que la borne inférieure est valable pour la famille des estimateurs non biaisés (pas seulement les estimateurs linéaires non biaisés).

Coordination des échantillons

Nous avons élaboré des méthodes de coordination des échantillons applicables à une enquête répétée réalisée selon un plan de sondage stratifié avec échantillonnage aléatoire simple sans remise (EASSR) dans chaque strate, quand la composition ou la définition des strates évolue. Étant donné le premier échantillon sélectionné avant les mises à jour de la base de sondage, notre objectif est de sélectionner dans le deuxième échantillon un nombre minimal de nouvelles unités, tout en obtenant les probabilités d'inclusion de premier ordre des unités dans le deuxième plan de sondage. Nos méthodes s'appuient sur la programmation linéaire (PL) pour obtenir des solutions optimales. Voir Mach, Reiss, Schiopu-Kratina (2006), ainsi que Matei et Skinner (2009) pour d'autres exemples de méthodes fondées sur la PL (programmation linéaire).

Nous avons élaboré une méthode qui maximise non seulement le chevauchement attendu, mais aussi le chevauchement de n'importe quel échantillon sélectionné au départ avec le deuxième échantillon sélectionné après les mises à jour de la base de sondage. Nous avons défini le concept d'*erreur de ligne* et montré que la minimisation des erreurs de ligne mène à la maximisation du chevauchement conditionnel. Nous avons rédigé le rapport technique intitulé « Maximizing the conditional overlap in business surveys » (Maximiser le chevauchement conditionnel dans les enquêtes-entreprises), qui comprend la théorie que nous avons élaborée, de nombreux exemples numériques et des simulations. À l'heure actuelle, le rapport est évalué par des examinateurs en tant que document de travail de Statistique Canada. Nous avons organisé la séance de communication sollicitée ayant pour thème « *Coordination of samples selected from overlapping populations* » (coordination d'échantillons tirés de populations chevauchantes) pour la quatrième conférence internationale sur les enquêtes auprès des établissements. Nous avons présenté nos résultats au Comité technique des enquêtes auprès des entreprises le 24 février 2012.

Nous soumettrons l'article intitulé « Maximizing the conditional overlap in business surveys » pour publication dans la revue *Journal of Statistical Planning and Inference*. Nous examinerons également la méthode de coordination des échantillons appelée « méthode de microstrates » qui a été utilisée par divers organismes statistiques, dont l'INSEE (Institut National de la Statistique et des Études Économiques). Nous proposons d'étudier, de formaliser et éventuellement d'améliorer les propriétés d'optimalité de la méthode des microstrates en la comparant à nos méthodes de coordination des échantillons et en important certaines techniques sur lesquelles s'appuient ces méthodes.

Méthode des microstrates pour la coordination des échantillons

La méthode des microstrates élaborée par Pascal Rivière (Rivière, 2001 et Rubin-Bleuer, 2002) offre une approche générale de la coordination des échantillons issus de plans EASSR (échantillonnage aléatoire simple sans remise) stratifiés. La technique est fondée sur la permutation des numéros aléatoires à l'intérieur des microstrates (définies comme étant les intersections de toutes les strates et tous les plans d'échantillonnage précédents). Le tri est effectué à l'intérieur des microstrates en fonction du fardeau de réponse cumulé tout en respectant le rang des numéros aléatoires attribués aux unités. Non seulement l'approche tient compte du fardeau de réponse cumulé de chaque unité, mais elle permet aussi à l'utilisateur de définir ses propres strates.

L'objectif du projet est de décrire et de discuter en détail les divers types de coordination que peut offrir la méthode des microstrates et de donner des preuves formelles de sa fiabilité dans chaque cas (coordination négative et positive, coordination avec un échantillon ou avec plusieurs échantillons

provenant d une série d échantillons sélectionnés en se fondant sur le fardeau de réponse, *etc. (et ainsi de suite)*). En collaboration avec Rivière, nous. avons rédigé un rapport d'avancement des travaux (Rivière et Rubin-Bleuer, 2011), qui comprend des preuves complètes, des exemples instructifs et un aperçu des méthodes de coordination fondées sur l'attribution de numéros aléatoires permanents (NAP). En outre, un exposé sur la méthode a été présenté au comité technique de la DMEE (Division des méthodes d'enquêtes auprès des entreprises) (Mach, Rubin-Bleuer et Schiopu-Kratina, 2012).

Ajustement de la non-réponse par le calage aux marges

Ce projet est le prolongement d'un autre projet (Verret et Kevins, 2010) qui visait à appliquer les méthodes de calage pour corriger la non-réponse de Särndal et Lundström (2008) dans un contexte d'un fichier maître à apparier. Dans l'approche de Särndal et Lundström, la fonction de lien entre l'inverse de la probabilité de réponse et des variables auxiliaires est implicitement supposée être linéaire. Ce projet tente d'intégrer les ajustements de non-réponse et de calage courants dans les enquêtes ménages de manière plus complète et intégrée, par exemple en permettant un lien logistique entre la probabilité de répondre et les variables explicatives.

On s'est familiarisé avec les articles de Chang et Kott (2008) et de Kott et Chang (2010). Ces auteurs y présentent une méthode de calage pour ajuster la non-réponse où le lien peut être non linéaire et où les variables de calage peuvent différer des variables modélisant la probabilité de réponse. On a ciblé deux enquêtes où les méthodes avaient un potentiel d'application : l'Enquête nationale sur les ménages et l'Enquête sur les dépenses des ménages (EDM).

On a programmé la méthode de calage aux marges de Kott et Chang (2008) en SAS/IML en utilisant la routine d'optimisation non linéaire NLPCG (*nonlinear optimization by conjugate gradient method*). On a identifié des problèmes computationnels relatifs à l'exécution de la méthode dans le cadre de l'application à l'estimation de l'EDM (Enquête sur les dépenses des ménages). On a corrigé ces problèmes en utilisant autant que possible les leçons apprises par le passé (Verret et Kevins (2010)) lors de l'application des méthodes de Särndal et Lundström (2008).

Bootstrap généralisé

La méthode du bootstrap généralisé peut être utilisée pour estimer la variance des estimateurs sous des plans d'échantillonnage généraux. Dans le contexte de cette méthode, les poids bootstrap sont définis de manière qu'il y ait concordance entre les deux premiers (ou plus) moments de l'erreur d'échantillonnage sous le plan et le moment bootstrap correspondant. La plupart des méthodes bootstrap décrites dans la littérature peuvent être considérées comme des cas particuliers, en particulier le bootstrap de Rao-Wu.

Nous avons accordé une attention particulière au cas de l'échantillonnage de Poisson, qui est souvent utilisé pour sélectionner les échantillons des enquêtes menées pour établir les indices des prix et avons illustré les propriétés de notre méthode bootstrap dans une étude par simulation. Nous avons achevé la révision d'un article (Beaumont et Patak, 2012) qui a été accepté pour publication dans la *Revue Internationale de Statistique*. Un exposé (Beaumont, 2011) a également été donné à la conférence de Statistique Canada qui s'est tenue à Montréal.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Pierre Lavallée (613 951-2892, pierre.lavallee@statcan.gc.ca).

Bibliographie

Chang, T.D., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 3, 555-571.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol (volume). 32, Nº (numéro). 2, 185-196.

Godambe, V.P., et Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

Kott, P.S., et Chang, T.D. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association,* 105, 491, 1265-1275.

Lavallée, P. (2002). *Le Sondage Indirect*, *ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.

Mach, L., Reiss, P.T. et Schiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the Northwest corner rule. *Journal of the American Statistical Association,* 101, 1671-1679.

Matei, A., et Skinner, C. (2009). Optimal sample coordination using controlled selection. *Journal of Statistical Planning and Inference*, 139, 3112-3121.

Rivière, P. (2001). Random permutations of random vectors as a way to coordinate samples. Rapport interne, juillet 2001, University of Southampton, Royaume-Uni.

Rubin-Bleuer, S. (2002). Report on Rivière's random permutations method of sampling coordination. Rapport interne (2002-02-26), Statistique Canada.

Särndal, C.-E., et Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, Vol (volume). 24, Nº (numéro). 2, 167-191.

Verret, F., et Kevins, C. (2010). Calage aux marges des poids d'enquêtes à poids complexe pour le refus à l'appariement. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada*.

Estimation sur petits domaines

L'estimation sur petits domaines devient plus pertinente aujourd'hui en raison de la hausse des coûts de la collecte de données, de la croissance de la

demande de statistiques fiables pour les petits domaines et de la nécessité de réduire le fardeau de réponse.

Les estimateurs directs s'appliquant à un domaine utilisent seulement les données provenant de l'échantillon dans ce domaine et leur précision n'est pas suffisante pour les petits domaines, parce que la taille de l'échantillon est petite. Par ailleurs, les estimateurs indirects applicables à un domaine (estimateurs sur petits domaines) empruntent des données à des domaines apparentés afin d'accroître la taille effective de l'échantillon. Des données sont empruntées aux autres domaines en s'appuyant sur une série d'hypothèses ou « modèle », puis les estimations sont produites en se basant sur ce modèle.

Nos travaux de recherche ont pour objectif de tenter de répondre aux questions suivantes :

- Existe-t-il une méthode d'estimation sur petits domaines qui fournit des estimations de suffisamment bonne qualité pour être publiées ? Pouvons-nous fournir une mesure de la qualité fondée sur le plan de sondage (fondée sur les propriétés d'échantillonnage) ?
- Les résultats statistiques sont-ils suffisamment généraux pour pouvoir être utilisés pour d'autres enquêtes ?
- Pouvons-nous dégager des thèmes de recherche de ces études ?
- Pouvons-nous mettre cette méthode en œuvre dans un environnement de production (coûts de développement et coûts opérationnels, actualité, réputation de l'organisme versus demande des clients, fardeau de réponse, coûts de collecte) ?
- Comment les résultats du projet et les questions qu'il soulève aideront-ils Statistique Canada à élaborer une stratégie de production d'estimations fondées sur un modèle ?

Durant la période visée par le présent rapport, nous avons fait progresser les travaux dans le cadre des 12 projets suivants grâce à des applications aux enquêtes sociales et aux enquêtes-entreprises.

Système SAS d'estimation sur petits domaines

Ce projet comportait l'examen et la documentation des méthodes d'estimation sur petits domaines et leur mise en œuvre dans un programme prototype développé en utilisant SAS. Nous avons passé en revue la méthodologie publiée pour l'estimation sur petits domaines et mis en œuvre quatre méthodes sous le modèle de Fay-Herriot au niveau du domaine, à savoir la maximisation de la densité corrigée (ADM pour *Adjusted Density Maximization*), le maximum de vraisemblance restreint (REML pour *Restricted Maximum Likelihood*), la méthode de Fay-Herriot (FH) et la méthode de Wang-Fuller (WF). Pour le modèle au niveau de l'unité, nous avons examiné et mis en œuvre les méthodes d'estimation de l'EBLUP (Empirical Best Linear Unbiased Predictor) et du pseudo-EBLUP (Empirical Best Linear Unbiased Predictor).

Le prototype que nous avons développé comprend un groupe de macros SAS modulaires. Ces macros valident les diverses entrées de données, produisent les estimations requises et affichent des critères diagnostiques pour aider les utilisateurs à évaluer la validité des modèles sous-jacents. Nous avons accru l'efficacité de ces travaux de développement en utilisant bon nombre de macros disponibles dans StatMx. La suite logique consiste à transférer le prototype dans StatMx ou dans un produit intégré bénéficiant d'un soutien complet dans le cadre du nouveau remaniement des systèmes généralisés.

Nous avons développé et mis à l'essai un programme en vue de produire des estimations sur petits domaines par les méthodes EBLUP (Empirical Best Linear Unbiased Predictor) et pseudo-EBLUP (Empirical Best Linear Unbiased Predictor). Nous l'avons groupé avec les programmes antérieurs d'estimation fondée sur la modélisation au niveau du domaine pour produire un progiciel offrant la modélisation au niveau du domaine ainsi qu'au niveau de l'unité.

Nous avons achevé la production de la documentation sur la méthodologie utilisée pour coder le programme d'estimation au niveau de l'unité. Nous l'avons ajoutée à la documentation antérieure concernant le modèle au niveau du domaine.

Nous avons présenté ces travaux de développement durant l'un des ateliers organisés dans le cadre du Symposium international sur les questions de méthodologie qui s'est tenu en novembre 2011. Nous avons décrit la méthodologie qui a été mise en œuvre et illustré l'utilisation des programmes d'estimation sur petits domaines en les appliquant à quatre exemples de données d'enquête.

Répartition de la taille de l'échantillon pour une estimation par domaine efficace

Les problèmes de répartition de l'échantillon sont étudiés dans le contexte de l'estimation des moyennes de sous-population (strate ou domaine), ainsi que de la moyenne de population agrégée sous échantillonnage aléatoire simple stratifié. Un article intitulé « On sample size allocation for efficient domain estimation » (À propos de la répartition de l'échantillon pour une estimation sur domaine efficace) a été révisé en vue de sa publication éventuelle dans la revue *Techniques d'enquête*. Cet article décrit l'utilisation d'une méthode de programmation non linéaire (NLP) pour obtenir la répartition « optimale » de l'échantillon entre les strates; cette optimisation minimise la taille totale de l'échantillon sous la contrainte des tolérances spécifiées pour le coefficient de variation des estimateurs des moyennes de strate et de la moyenne de population. La taille totale d'échantillon résultante est alors utilisée pour déterminer les répartitions de la taille de l'échantillon par la méthode de Costa, Satorra et Ventura (2004) qui s'appuie sur une répartition intermédiaire ou de compromis (*compromise allocation*) et par celle de Longford (2006) fondée sur des « priorités inférencielles » spécifiées. En outre, nous avons étudié la répartition de l'échantillon entre les strates quand sont également spécifiées des exigences de fiabilité pour des domaines qui recoupent les strates. Les propriétés des trois méthodes sont étudiées au moyen de données provenant de l'Enquête mensuelle sur le commerce de détail (EMCD) auprès des entreprises à établissement unique menée par Statistique Canada.

Questions concernant l'estimation de la variance pour le modèle de Fay-Herriot

L'estimateur fondé sur le meilleur prédicteur linéaire empirique (EBLUP pour *Empirical Best Linear Unbiased Predictor*) d'une moyenne de petit domaine, obtenu par ajustement du modèle de Fay-Herriot (1979), est une moyenne pondérée de l'estimateur direct d'après les données d'enquête et de l'estimateur hybride régression-synthétique. Les poids dépendent de la variance des effets de domaine aléatoires. De nombreuses méthodes d'estimation de la variance produisent souvent une valeur négative, auquel cas nous donnons au poids de l'estimateur direct la valeur zéro et l'estimateur EBLUP (Empirical Best Linear Unbiased Predictor) devient un estimateur synthétique. La plupart des praticiens hésitent à utiliser des estimateurs synthétiques pour les moyennes de petit domaine, parce qu'ils sont souvent biaisés et que leur erreur quadratique moyenne (EQM) fondée sur le plan de sondage est plus grande que celle de l'estimateur direct. Ce problème a donné le jour à une série de méthodes d'estimation de la

variance qui produisent toujours des estimations positives. Cependant, le prix en est souvent un accroissement de l'EQM (Erreurs quadratiques moyennes) sous le modèle des EBLUP (Empirical Best Linear Unbiased Predictor). Dans le cadre de la présente étude, nous avons proposé une autre méthode, celle de l'estimateur de variance MIX, qui est une combinaison des estimateurs de variance par le maximum de vraisemblance restreint (REML) et par la vraisemblance maximale ajustée (ADM). Nous avons démontré que les propriétés asymptotiques de l'estimateur MIX sont supérieures à celles de l'estimateur ADM (Adjusted Density Maximization). En outre, nous avons comparé les méthodes REML (Restricted Maximum Likelihood), ADM (Adjusted Density Maximization) et MIX d'estimation de la variance en exécutant une simulation Monte Carlo et en comparant les EQM (Erreurs quadratiques moyennes) sous le modèle. Nos résultats de simulation préliminaires montrent que l'estimateur MIX a de meilleures propriétés que l'estimateur REML (Restricted Maximum Likelihood) ou ADM (Adjusted Density Maximization) quand les variances d'échantillonnage ne suivent pas une loi normale.

Modèles chronologiques et transversaux d'estimation sur petits domaines de type Yu-Rao pour l'Enquête sur l'emploi, la rémunération et les heures de travail (EERH)

L'EERH (Enquête sur l'emploi, la rémunération et les heures de travail) est une enquête mensuelle conçue pour produire des estimations des niveaux et des tendances mensuelles de la rémunération, de l'emploi, des heures rémunérées et des gains. À l'heure actuelle, les paramètres de l'EERH (Enquête sur l'emploi, la rémunération et les heures de travail) sont estimés en se servant de l'estimateur par la régression généralisée (GREG). Afin d'obtenir des estimateurs d'une plus grande précision, le programme de l'EERH (Enquête sur l'emploi, la rémunération et les heures de travail) prévoit utiliser un estimateur composite RC qui est un estimateur par la régression généralisée (GREG) renforcé par l'emprunt d'information au cours du temps en utilisant comme données auxiliaires des renseignements provenant de cycles antérieurs de l'enquête. À long terme, l'objectif du projet est d'examiner divers modèles pour l'estimation sur petits domaines (SCIAN4 par province) pour l'EERH (Enquête sur l'emploi, la rémunération et les heures de travail). Nous avons commencé par utiliser un modèle transversal et chronologique de type Rao-Yu (Rao, 2003) et étudié diverses méthodes d'estimation de la variance, en particulier les propriétés sous le modèle de l'estimateur de variance par la méthode du maximum de vraisemblance ajusté (ADM) proposé par Li et Lahiri (2010) pour le modèle de Fay-Herriot (1979) et élaboré par Rubin-Bleuer, Yung et Landry (2010) ainsi que Yung, Rubin-Bleuer et Landry (2010) pour le modèle de Rao-Yu. Durant la période visée par le présent rapport, nous avons conçu une simulation fondée sur un modèle pour étudier les propriétés finies de la méthode ADM (Adjusted Density Maximization) d'estimation de la variance pour le modèle pris en considération. Nous avons constaté que la combinaison des estimateurs REML (Restricted Maximum Likelihood) et ADM (Adjusted Density Maximization), c'est-à-dire la procédure MIX, produit l'estimateur EBLUP (Empirical Best Linear Unbiased Predictor) le plus efficace si on la compare aux procédures strictement fondées sur le REML (Restricted Maximum Likelihood) ou sur l'ADM (Adjusted Density Maximization). Nous avons montré que l'estimateur MIX a les mêmes propriétés asymptotiques que l'estimateur REML (Restricted Maximum Likelihood), en ayant l'avantage supplémentaire de donner lieu à des estimations de variance toujours positives. Nous avons présenté ces résultats à une conférence consacrée à l'estimation sur petits domaines (SAE 2011) qui était une conférence satellite du congrès de 2011 de l'Institut international statistique. Nous continuons d'explorer les propriétés des divers estimateurs de variance pour des populations ayant des structures de variance d'échantillonnage plus complexes. Nous donnerons un exposé résumant les résultats des travaux que nous avons mené depuis 2010 au Symposium on the Analysis of Survey Data and Small Area Estimation du Fields Institute qui se déroulera du 30 mai au 1er juin 2012.

Extensions de l'estimateur Pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) avec application à l'EERH (Enquête sur l'emploi, la rémunération et les heures de travail)

En 2007, nous avons étudié divers estimateurs sur petits domaines transversaux et présenté une comparaison de ces derniers en vue d'une application éventuelle à l'estimation sur petits domaines pour l'EERH (Enquête sur l'emploi, la rémunération et les heures de travail). Durant cet exercice, nous avons étendu la théorie des estimateurs fondés sur le pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) développée au départ par You et Rao (2002) et élaboré des estimateurs sur petits domaines fondés sur le pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) de moyennes pondérées où les poids correspondaient aux poids économiques des modèles pour les erreurs au niveau de l'unité avec variances non constantes. En

outre, nous avons prouvé que la propriété d'auto-étalonnage des pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) originaux était vérifiée également pour les moyennes pondérées et nous avons développé la formule des erreurs quadratiques moyennes (EQM) sous le modèle correspondantes. Nous avons constaté que l'EQM (Erreurs quadratiques moyennes) sous le modèle ne correspondait pas bien à l'EQM (Erreurs quadratiques moyennes) Monte Carlo sous le plan de sondage. Tous ces résultats ont été publiés dans le Recueil de 2007 du Groupe des méthodes d'enquête de la Société statistique du Canada (SSC), dans les actes de la conférence satellite sur l'estimation pour petits domaines du congrès bisannuel de 2007 de l'Institut international de statistique, dans un exposé interne donné à Statistique Canada (Rubin-Bleuer, 2008), dans un exposé à l'intention du Comité consultatif des méthodes statistiques et dans le cadre de diverses autres communications. Durant la présente période, nous prévoyons rédiger un article qui sera soumis pour publication dans une revue à comité de lecture. Nous avons révisé le code du programme afin d'inclure l'estimateur EBLUP (Empirical Best Linear Unbiased Predictor) simple dans la comparaison, et nous prévoyons exécuter le programme durant le nouvel exercice. Nous avons rédigé une ébauche d'article qui comprend les développements théoriques qui sous-tendent ces travaux.

Estimation de l'erreur quadratique moyenne pour des modèles à spline pénalisée au niveau du domaine pour les données sur les entreprises

Dans le cadre de ce projet, nous avons étudié les propriétés statistiques d'un meilleur prédicteur linéaire sans biais (PS-EBLUP (Empirical Best Linear Unbiased Predictor)) des moyennes de petit domaine sous un modèle au niveau du domaine avec un modèle de lien à spline pénalisée (PS). Nos études antérieures portaient sur des modèles PS au niveau de l'unité et comprenaient un nombre élevé de petits domaines dans lesquels il était possible d'emprunter de l'information. Dans nos enquêtes-entreprises, le nombre de domaines est habituellement compris entre 20 et 50. Par conséquent, au cours de la période de recherche antérieure, nous avons exécuté des simulations en utilisant des populations de 50 domaines, et en travaillant avec un volume considérablement plus faible de données que pour les cas étudiés par d'autres auteurs. Nous n'avions pas obtenu des résultats concluants permettant une recommandation quant aux meilleures pratiques. Durant la présente période de recherche, nous avons réécrit le code en R, afin de rendre le programme plus flexible, et nous avons considéré un modèle plus simple. Nous avons étudié les propriétés du PS-EBLUP (Empirical Best Linear Unbiased Predictor) sous un modèle au niveau du domaine contenant des covariables qui suivent une loi uniforme. Nous nous sommes concentrés sur le comportement des estimateurs de la variance et de l'EQM (Erreurs quadratiques moyennes) du PS-EBLUP (Empirical Best

Linear Unbiased Predictor) pour 50, 100 et 200 domaines. Nous avons étudié les propriétés en échantillon fini de l'estimateur de variance par la méthode des moments ainsi que de l'estimateur de variance par la méthode du maximum de vraisemblance restreint (REML). Pour l'estimateur de variance REML (Restricted Maximum Likelihood), nous avons élaboré un autre algorithme que celui utilisé les années précédentes. Nous avons calculé la moyenne et la variance Monte Carlo des estimateurs de variance, ainsi que l'EQM (Erreurs quadratiques moyennes) du PS-EBLUP (Empirical Best Linear Unbiased Predictor). Selon les résultats préliminaires, la méthodologie pourrait être fiable si le nombre de domaines est grand, mais les estimateurs pourraient être instables si le nombre de domaines desquels on peut emprunter de l'information est inférieur à 50.

Estimation sur petits domaines avec des modèles au niveau des unités en présence d'un plan de sondage informatif

Les modèles au niveau des unités de population sont souvent utilisés en estimation pour petits domaines reposant sur des modèles pour des totaux et des moyennes. Ces modèles peuvent ne pas être applicables à l'échantillon si le plan d'échantillonnage est informatif. Un plan est informatif si les probabilités de sélection des domaines sont liées aux moyennes dans les domaines ou si les probabilités de sélection des unités sont liées aux valeurs de la variable étudiée des unités. Les méthodes habituelles, supposant que le modèle est approprié pour l'échantillon, peuvent mener à des estimateurs biaisés. Dans Verret, Hidiroglou et Rao (2010), on a étudié par simulation des méthodes alternatives utilisant les poids de sondage comme variable auxiliaire supplémentaire dans le modèle ajusté à l'échantillon et/ou dans l'estimation des moyennes et des EQM (Erreurs quadratiques moyennes) en utilisant l'approche pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) proposée par You et Rao (2002). Ces simulations ont montré que l'ajustement de ce modèle augmenté permettait des gains importants de précision dans les estimations ponctuelles sous des plans informatifs, tant du point de vue du biais que de l'EQM (Erreurs quadratiques moyennes). Cette inclusion donnait aussi des gains intéressants sur la précision des estimateurs d'EQM (Erreurs quadratiques moyennes) des estimateurs ponctuels. Le projet de recherche tentera d'expliquer théoriquement les résultats empiriques observés. Ceci permettra potentiellement d'apporter des améliorations aux estimateurs considérés et de déterminer si l'ajout du poids est suffisant pour compenser l'informativité du plan. De plus, les modèles considérés dans l'article sont des modèles d'analyse de la variance avec un facteur aléatoire et le projet de recherche étendra les résultats à des modèles linéaires. Finalement, on étendra la recherche d'un plan stratifié où les domaines sont les strates à un plan de sondage à deux degrés où les domaines sont les grappes. Les travaux dans Verret *et al*. (2010) et ceux accomplis par la suite ont été présentés par le Professeur Jon K. Rao à Trier en Allemagne en août 2011 à la *SAE 2011 Workshop*. De plus, les simulations des années fiscales passées ont été étendues de modèles à moyenne commune à des modèles linéaires. Les plans de sondage étaient stratifiés à un degré PPT, où l'on contrôlait l'informativité de la sélection à la manière d'Asparouhov (2006) ou à celle de Pfeffermann et Sverchkov (2007). Finalement, un article résumant les recherches a été écrit et soumis à la revue *Techniques d'enquête*.

Étalonnage des estimateurs sur petits domaines fondé sur un modèle en se basant sur le modèle de Fay-Herriot au niveau du domaine

Dans l'estimation sur petits domaines, il est important que les utilisateurs de données d'enquête étalonnent les estimations sur petits domaines fondées sur un modèle de manière que la somme des estimations sous modèle étalonnées soit égale au total des estimations directes au même niveau d'agrégation ou à un niveau plus élevé. You et Rao (2002) ont proposé un estimateur pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) du modèle au niveau de l'unité ayant une propriété d'auto-étalonnage. Pour les modèles au niveau du domaine, Wang, Fuller et Qu (2008) ont proposé une méthode en vue d'ajouter les variances d'échantillonnage pondérées en tant que variable auxiliaire supplémentaire dans le modèle de régression pour obtenir la propriété d'auto-étalonnage des estimations EBLUP (Empirical Best Linear Unbiased Predictor). Dans le présent projet, nous avons appliqué la procédure de You-Rao (2002) au modèle au niveau du domaine et avons obtenu l'estimateur de l'EQM (Erreurs quadratiques moyennes) pour l'estimateur à auto-étalonnage de You-Rao (YR) sous le modèle de Fay-Herriot. Nous avons comparé les deux estimateurs à auto-étalonnage, c'est-à-dire l'estimateur EBLUP (Empirical Best Linear Unbiased Predictor) augmenté de Wang-Fuller-Qu (WFQ) et l'estimateur YR, au moyen d'une étude par simulation. Nous avons achevé une étude par simulation pour comparer les estimateurs EBLUP (Empirical Best Linear Unbiased Predictor), YR et WFQ en ce qui concerne l'estimation de l'EQM (Erreurs quadratiques moyennes). Nous avons également étudié les deux estimateurs étalonnés sous une spécification incorrecte du modèle. Un document de travail de la Direction de la méthodologie a été rédigé (You, Rao et Hidiroglou, 2011). L'article a également été soumis à une revue pour publication et est révisé à l'heure actuelle.

Estimation sur petits domaines pseudo hiérarchique bayésienne (HB) fondée sur des modèles de régression à erreurs emboîtées généraux

Dans ce projet, nous considérons un modèle de régression à erreurs emboîtées général avec variances d'erreur inégales. You et Rao (2003) ont proposé des estimateurs pseudo-HB fondés sur un modèle élémentaire au niveau de l'unité. Dans le présent projet, nous avons étendu la méthode de You et Rao (2003) au modèle général au niveau de l'unité et obtenu des estimateurs fondés sur un modèle convergents sous le plan ayant la propriété d'auto-étalonnage. Nous avons proposé différents estimateurs pseudo-HB et parachevé le développement théorique de ces estimateurs. Les estimateurs HB proposés ont également été étendus à l'estimation d'une moyenne de population pondérée. Nous avons mis la touche finale à la méthode proposée et rédigé un document de travail de la Direction de la méthodologie, et nous rédigerons un document décrivant la méthode (You et Hidiroglou, 2012). Nous prévoyons aussi soumettre l'article à une revue pour publication éventuelle.

Comparaison et évaluation des modèles au niveau de l'unité et au niveau du domaine fondés sur l'échantillonnage PPT sous spécification incorrecte du modèle

Des modèles au niveau du domaine et des domaines au niveau de l'unité sont généralement utilisés dans l'estimation sur petits domaines pour obtenir des estimations efficaces pour les petits domaines. Dans le cadre de ce projet, nous examinons l'estimation des intervalles de confiance fondée sur les estimations d'après des modèles au niveau du domaine et au niveau de l'unité. En particulier, nous nous intéressons aux propriétés des estimateurs au niveau domaine et au niveau de l'unité sous un modèle correctement spécifié. Nous avons développé les estimateurs directs sous échantillonnage PPT pour le modèle au niveau du domaine. Nous considérons les estimateurs EBLUP (Empirical Best Linear Unbiased Predictor) et pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) pour les modèles au niveau de l'unité. Le résultat montre que le modèle au niveau de l'unité est plus efficace que le modèle au niveau du domaine, et que les estimateurs EBLUP (Empirical Best Linear Unbiased Predictor) et pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) donnent tous deux de bons résultats sous modélisation correcte. Cependant, en cas de spécification incorrecte du modèle, les estimateurs pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) avec poids d'échantillonnage donnent de meilleurs résultats que les

estimateurs EBLUP (Empirical Best Linear Unbiased Predictor) sans utilisation de ces poids, particulièrement sous un plan d'échantillonnage informatif. Les résultats seront présentés au congrès annuel 2012 de la SSC (Hidiroglou et You, 2012).

Méthodes de lissage de la variance d'échantillonnage pour des estimations de proportions dans l'estimation sur petits domaines

Le lissage de la variance d'échantillonnage est un sujet important pour ce qui est de l'estimation sur petits domaines. Dans le cadre de ce projet, nous étudions la méthode de lissage de la variance d'échantillonnage pour les estimateurs de proportions sur petits domaines. Nous avons proposé deux méthodes pour les estimateurs directs de la variance d'échantillonnage pour les proportions, à savoir la méthode des effets de plan (EP) et la méthode de la fonction de variance généralisée (FVG). En particulier, nous avons montré que les méthodes de lissage EP (effets de plan) et FVG (fonction de variance généralisée) proposées sont équivalentes. Nous avons évalué et comparé les estimations de variance lissées sous les méthodes proposées en procédant à l'analyse de données provenant de diverses enquêtes menées par Statistique Canada, dont l'EPA, l'ESCC et l'EPLA. Les méthodes de lissage de la variance d'échantillonnage EP (effets de plan) et FVG (fonction de variance généralisée) proposées peuvent également être appliquées et étendues à des problèmes d'estimation plus généraux, y compris l'estimation de proportions et de dénombrements. Un article (You et Hidiroglou, 2012) a été rédigé et les résultats seront présentés au Symposium on the Analysis of Survey Data and Small Area Estimation, qui se tiendra à Ottawa en 2012.

Estimation sur petits domaines au niveau de l'unité pour les données d'enquêtes-entreprises

Dans les enquêtes-entreprises, la stratification est habituellement fondée sur la géographie, l'industrie et la taille de l'entreprise. Les estimateurs les plus simples de divers domaines et paramètres d'intérêt (il pourrait s'agir d'une moyenne de population pondérée) s'appuient sur les poids de sondage originaux. Les données auxiliaires existantes peuvent être utilisées pour caler ces poids de sondage originaux. Si les données auxiliaires sont bien corrélées avec la variable d'intérêt, l'estimateur calé résultant (estimateur direct) sera plus efficace que celui obtenu avec les poids originaux. Cependant, à mesure que les domaines d'intérêt s'écartent de la stratification originale, des méthodes d'estimation sur petits domaines peuvent être utilisées pour améliorer l'estimateur direct. Dans le présent projet, nous avons étudié l'utilisation d'un modèle de régression à erreurs emboîtées général avec variance d'erreur inégale pour illustrer la façon dont les données auxiliaires disponibles au niveau de l'unité peuvent être utilisées. En particulier, nous avons étendu la théorie de Stukel et Rao (1999) et de You et Rao (2002) pour traiter ce cas. Nous avons développé des estimateurs EBLUP (Empirical Best Linear Unbiased Predictor) et pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) sous le modèle de régression à erreurs emboîtées général et obtenu les estimateurs de l'EQM (Erreurs quadratiques moyennes) correspondants. Pour les enquêtes-entreprises, le paramètre d'intérêt est habituellement une simple moyenne de population; cependant, il pourrait aussi s'agir d'une moyenne de population pondérée, où la pondération détermine l'importance de chaque unité. Rubin-Bleuer, Godbout et Morin (2007) ont élaboré un estimateur sur petits domaines convenant pour estimer la moyenne de population pondérée et ont exécuté une simulation en vue de comparer les estimateurs pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) à un certain nombre d'estimateurs directs. Rubin-Bleuer, Godbout et Hidiroglou (2012) donnent une version technique plus détaillée de cet article. Nous avons étendu les estimateurs EBLUP (Empirical Best Linear Unbiased Predictor) et pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) à l'estimation de la moyenne de population pondérée (Estevao, Hidiroglou et You 2011) en tenant compte de l'hétérogénéité de la variance. L'un de nos estimateurs pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) ponctuel correspond à celui obtenu par Rubin-Bleuer et coll. (2007). Nous illustrons l'application des méthodes et modèles proposés à des données d'enquête réelles sur les entreprises. Les méthodes et les résultats seront présentés au Symposium on the Analysis of Survey Data and Small Area Estimation, qui se tiendra à Ottawa en 2012.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Susana Rubin-Bleuer (613 951-6941, susana.rubin-bleuer@statcan.gc.ca).

Bibliographie

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, 35, 439-460.

Costa, A., Satorra, A. et Ventura, E. (2004). Using Composite estimators to improve both domain and total area estimation. *SORT*,28, 69-86.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Li, H., et Lahiri, P. (2010). *An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis*, 101, 882-892.

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, Nº (numéro). 1, 97-106.

Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, Vol (volume). 102, Nº (numéro). 480, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey : John Wiley & Sons, Inc. (incorporated)

Rubin-Bleuer, S. (2008). Evaluation of Small Area Estimators for the Canadian Survey of Employment Payrolls and Hours. Présentation interne.

Rubin-Bleuer, S., Godbout, S. et Morin, Y. (2007). Evaluation of small domain estimators for the survey of employment payroll and hours. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada*, juin 2007.

Rubin-Bleuer, S., Yung, W. et Landry, S. (2010). Adjusted maximum likelihood method for a small area model accounting for time and area effects. SRID-2010-005E.

Stukel, D.M., et Rao, J.N.K. (1999). Small area estimation under two-fold nested errors regression models. *Journal of statistical planning and inference*, 78, 131-147.

Verret, F., Hidiroglou, M.A. et Rao, J.N.K. (2010). Small area estimation under informative sampling. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada* 2010.

Wang, J., Fuller, W.A. et Qu, Y. (2008). Estimation pour petits domaines sous une contrainte. *Techniques d'enquête*, 34, Nº (numéro). 1, 33-40.

You, Y., et Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics,* 30, 431-439.

You, Y., et Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111, 197-208.

Yung, W., Rubin-Bleuer, S. et Landry, S. (2010). Small area estimation for business surveys. *Proceedings of the Survey Research Section*,American Statistical Association.

Recherche sur l'analyse des données (RAD)

Les ressources affectées à la recherche sur l'analyse des données sont utilisées pour mener des travaux de recherche concernant des problèmes de méthodologie liés à l'analyse courante qui ont été relevés par les analystes et les méthodologistes; elles sont également consacrées à des travaux de recherche sur des problèmes qui devraient avoir une importance stratégique dans un avenir prévisible.

Méta-analyse des données d'enquête

Les travaux de recherche accomplis jusqu'à présent portaient sur diverses questions statistiques ayant trait à la méta-analyse des données d'enquête. Ces questions comprennent la création d'un cadre méthodologique pour combiner les données d'enquête, la comparaison de ce cadre au cadre classique proposé par Cochran (1954) pour la combinaison des expériences, la proposition d'une nouvelle méthode de pondération qui tient compte des différences de variabilité dues au plan d'échantillonnage, l'examen de la convergence des estimateurs méta-analytiques, et la discussion des nombreuses hypothèses implicites que font les chercheurs qu'ils appliquent des méthodes de méta-analyse aux données d'enquête. En outre, en vue d'aider les chercheurs, nous avons étendu les lignes directrices sur l'exécution d'examens et la présentation des résultats proposées par des groupes tels la Collaboration Cochran et le Meta-analysis of Observational Studies in Epidemiology (MOOSE) (Stroup, Berlin, Morton, Olkin, Williamson, Rennie, Mohar, Becker, Sipe et Thacker, 2000). Certains résultats de la recherche faisaient partie d'une thèse de doctorat et ont également été présentés au Comité consultatif de Statistique Canada à l'automne 2011 [(Fox (2011a) et Fox (2011b)].

L'inférence par équations d'estimation à l'aide de modèles multi-niveaux et de données d'enquêtes

Des estimateurs résolvant les équations d'estimation résultant de la maximisation de la vraisemblance composite (MVC) (Lele & Taper, 2002; Rao, Verret & Hidiroglou, 2010) avaient été développés en 2010-2011 pour des modèles linéaires mixtes appliqués à des données d'enquêtes ayant un plan de sondage informatif. En 2011-2012, les estimateurs correspondants du maximum de vraisemblance (Korn & Graubard, 2003) et suivant les développements d'Asparouhov (2006) ont été explicités et tous les estimateurs ont été intégrés aux programmes de simulations.

À la lumière de biais observés dans les simulations dans l'estimation de la variance entre les grappes, les estimateurs du MVC ont été révisés pour suivre plus étroitement la théorie de l'article de Lele et Taper (2002). Ceci a eu pour effet de réduite les biais, mais l'estimation de la variance de ces nouveaux estimateurs serait plus complexe. D'autre part, il faudra raffiner l'algorithme utilisé pour obtenir les estimateurs complexes du maximum de vraisemblance, car il présente des problèmes de convergence. Les estimateurs de variance des estimateurs ponctuels du MVC, inspirés de Binder (1983), ont été évalués dans une autre simulation. Ils ont un biais négatif, mais faible. Finalement, l'écriture d'un article a débuté.

Analyse spatiale des données géocodées

Une dépendance spatiale peut être découverte dans des observations fortement regroupées dans l'espace, comme celles sur les habitudes de vote régionales, le cancer du poumon, les récoltes et la criminalité. Statistique Canada possède des données administratives ainsi que des données d'enquêtes présentant ce genre de dépendance. Lorsque l'on ne tient pas compte de cette dépendance spatiale, les analyses de régression peuvent produire des estimations inefficaces des paramètres et donner des tests de signification inexacts. Le modèle de décalage spatial (*spatial lag model*) et le modèle d'erreur spatiale (*spatial error model*) sont tous deux utilisés fréquemment pour l'analyse des données spatiales. Ces deux modèles intègrent la dépendance spatiale au moyen d'une matrice d'association spatiale spécifiée par l'utilisateur, appelée matrice de pondération spatiale, qui indique quelles sont les régions qui s'influencent l'une l'autre. Le but du projet est de déterminer l'effet de l'utilisation, pour ajuster un modèle spatial autorégressif, d'une matrice de pondération différente de celle employée pour générer les données spatiales. La probabilité d'obtenir le modèle réel, ainsi que l'erreur quadratique moyenne et le biais relatif sont examinés sous diverses conditions. Des programmes SAS ont été écrits pour simuler des ensembles de données possédant des structures connues de dépendance spatiale, puis pour ajuster un modèle en employant différentes matrices de pondération. Un grand nombre de simulations sont exécutées à l'heure actuelle et des programmes sont créés pour résumer les résultats. Ceux-ci aideront à déterminer quels types de modèles spatiaux seraient les plus utiles pour les données de Statistique Canada.

Inférence sous échantillonnage informatif

Lorsque l'on souhaite procéder à l'analyse de données d'enquête dans un cadre de superpopulation, il est important de savoir si le plan de sondage est ignorable ou informatif (Binder et Roberts, 2001). L'objectif de la présente étude est double : a) arriver à comprendre pleinement ces deux concepts et b) créer un cadre mathématique approprié dans lequel nous pouvons quantifier le rôle de ces concepts dans l'analyse. L'étape (a) peut aboutir à l'introduction, au moyen de cours donnés à l'interne, de nouveaux outils techniques pour analyser les données d'enquête. Pour ce qui est de l'étape (b), nous avons l'intention de définir un cadre très général, qui étendra ou étoffera la structure de cadres similaires décrits dans la littérature, par exemple Hartley et Sielken (1975), Hajek (1981), Rubin-Bleuer et Schiopu-Kratina (2005).

Nous avons discuté des résultats que nous devons à Sugden et Smith (1984) et pris note des idées exposées dans D.F. Heitjan (1997). Selon ce dernier

article, l'approche adoptée pour les problèmes de données « grossières » (*problems of coarse data*) est suffisamment générale pour être utilisée dans l'analyse des données d'enquête avec observations manquantes ou information incomplète sur le plan de sondage. Dans un article fondamental, Dawid (1979) fait valoir l'utilisation du concept d'indépendance conditionnelle en statistiques en général, et dans l'analyse des données en particulier. Il exprime également formellement diverses conditions d'ignorabilité et les liens entre elles. L'approche unifiante de Dawid, illustrée au moyen d'exemples simples, a pu être adaptée à des fins de consultation et d'enseignement. Nous avons ensuite appliqué le concept de l'indépendance conditionnelle pour interpréter et pour dériver formellement d'autres conditions d'ignorabilité. Nous avons l'intention de poursuivre les travaux concernant les objectifs a) et b).

Modèle à risques proportionnels pour données d'enquête

Un article a été publié sur le sujet (Rubin-Bleuer, 2011). Les travaux de recherche ont été exécutés durant l'exercice précédent.

Certains sujets relatifs aux méthodes fondées sur le plan de sondage

Ce projet comprend l'étude de divers problèmes de recherche qui ont été cernés durant des contacts avec les analystes ou les méthodologistes. Souvent, ces sujets sont choisis à la suite de consultations pour lesquelles aucune réponse satisfaisante ne s'est dégagée immédiatement.

1. Inférence de la causalité d'après des données d'enquête

Puisque les enquêtes à l'échelle de la population sont des études observationnelles, il semble naturel de se demander quel pourrait être l'effet du plan de sondage sur les inférences causales. Nous avons procédé à une revue de la littérature sur l'inférence causale à partir d'études observationnelles. En nous appuyant sur la délimitation des divers niveaux de causalité de Cox et Wermuth (2004), nous avons étudié quelles hypothèses sont appropriées pour considérer que le plan de sondage est ignorable.

Une communication sollicitée sur le sujet a été faite au congrès annuel de 2011 de la SSC. Un article (Binder, 2011a) a été soumis pour publication dans le Recueil du Groupe des méthodes d'enquête de la Société statistique du Canada.

2. Un cadre de randomisation sous un modèle et sous le plan de sondage pour les méthodes analytiques appliquées aux données d'enquêtes complexes

Lorsqu'un analyste doit estimer les paramètres d'un modèle ajusté à des données provenant d'une enquête complexe, l'une des premières questions qu'il se pose souvent est celle de savoir s'il doit ou non utiliser les poids de sondage. Cependant, la bonne question qu'il faut poser est celle de savoir si l'information proprement dite sur le plan de sondage est pertinente et, si elle l'est, de quelle façon elle doit être intégrée dans l'analyse. Le débat entre les tenants de l'approche fondée sur le plan de sondage et ceux de l'approche fondée sur un modèle pour faire des inférences au sujet des paramètres d'un modèle peut être expliqué et clarifié en utilisant un cadre de randomisation sous un modèle et sous le plan pour décrire comment ont été obtenues les observations sur les unités échantillonnées. Un article sur le sujet a été publié dans la revue *Pakistan Journal of Statistics*. (Binder, 2011b), dans un numéro spécial dédié à Ken Brewer.

3. Nombre de degrés de liberté pour une enquête dont le nombre d'UPE échantillonnées est limité

Supposons que l'on mène une enquête selon un plan de sondage stratifié avec sélection d'UPE (de manière entièrement ou approximativement indépendante) dans chaque strate. Si l'on utilise l'approximation de Satterthwaite pour déterminer le nombre de degrés de liberté d'une estimation de variance d'après une telle enquête, la règle empirique ddl=(n^{bre} d'UPE échantillonnées – n^{bre} de strates) fréquemment utilisée pourrait donner lieu à une surestimation importante lorsque l'on estime les intervalles de confiance et que l'on teste les inférences si le nombre d'UPE dans le plan de sondage est faible. Cela pourrait être le cas de l'Enquête canadienne sur les mesures de la santé (ECMS), où le nombre d'UPE échantillonnées est égal à 15 et le nombre de strates est égal à 4. Nous avons procédé à une revue de la littérature sur le sujet et préparé un résumé d'une partie de l'information (Roberts et Binder, 2012), en formulant certains commentaires s'appliquant spécifiquement à l'ECMS.

4. Bootstrap pour les paramètres d'un modèle

Le bootstrap de Rao-Wu est souvent utilisé pour estimer la variance sous le plan des estimateurs des paramètres de population finie. Lorsqu'on estime les paramètres d'un modèle, deux sources de variabilité devraient normalement être prises en compte pour estimer la variance : le plan d'échantillonnage et le modèle hypothétique à l'origine de la population finie. Si la fraction d'échantillonnage est négligeable, on peut en principe ignorer la variabilité du modèle, ce qui est souvent le cas en pratique, et le bootstrap de Rao-Wu demeure valide. Toutefois, cette simplification n'est pas toujours appropriée. Le fait de supposer que l'échantillonnage se fait avec remise ne suffit pas pour refléter la variance sous le modèle. La méthode du bootstrap généralisé peut être utilisée pour tenir compte comme il convient des deux sources de variabilité. Cette procédure générale peut s'appliquer à tout paramètre défini au moyen d'une équation d'estimation, à condition d'émettre l'hypothèse que les observations sont indépendantes sous le modèle. Elle est facile à mettre en œuvre une fois que l'on a obtenu les poids bootstrap qui reflètent les deux premiers moments sous le plan (par exemple, en utilisant la méthode du bootstrap de Rao-Wu) ou le bootstrap généralisé susmentionné. Un article révisé (Beaumont et Charest, 2012) vient d'être accepté pour publication dans la revue *Computational Statistics and Data Analysis* et paraîtra en 2012.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Georgia Roberts (613 951-1471, georgia.roberts@statcan.gc.ca).

Bibliographie

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, 35, 439-460.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D., et Roberts, G. (2001). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section*, American Statistical Association, Issue 12.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, Nº (numéro). 1, 101-129.

Cox, D.R., et Wermuth, N. (2004). Causality: A statistical view. *Int. Statist. Rev.*, 72, 285-305.

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, Séries B, 41, Nº (numéro). 1, 1-31.

Hajek, J. (1981). (assemblé après sa mort par Vaclav Dupac). *Sampling from a finite population*. New York : Dekker.

Hartley, H.O., et Sielken, R.L. (1975). A "superpopulation viewpoint" for finite population sampling. *Biometrics*, 31, 411-422.

Heitjan, D.F. (1997). Ignorability, sufficiency and ancillarity. *Journal of the Royal Statistical Society*, Séries B, 59, 375-381.

Korn, E.L., et Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society*, Séries B, 65, 175-190.

Lele, S., et Taper, M.L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 109, 117-135.

Rao, J.N.K., Verret, F. et Hidiroglou, M.A. (2010). A weighted estimating equations approach to inference for two-level models from survey data. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada 2010*.

Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005). On the two- phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, Nº (numéro). 6, 2789-2810.

Stroup, D., Berlin, J., Morton, S., Olkin, I., Williamson, D., Rennie, D., Mohar, D., Becker, B., Sipe, T.A. et Thacker, S. (2000). Meta-analysis in observational studies a proposal for reporting. *Journal of the American Medical Association*, 283, Nº (numéro). 15.

Sugden, R.A., et Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*,71, 495-506.

Collecte des données

Les projets de recherche en collecte visent essentiellement l'avancement de nos connaissances afin de pouvoir mettre en place des processus de collecte plus efficaces en termes de coût et de qualité et qui permettront de répondre aux besoins émergents.

Huit types de projets sont présentés : 1) de nombreux projets en recherche opérationnelle permettant la revue et l'évaluation des procédures de collecte, 2) l'établissement de lignes directrices pour le développement des questionnaires électroniques, 3) une approche innovatrice au moyen de la théorie des graphes pour analyser la complexité des questionnaires afin de potentiellement réduire le fardeau de réponse et faciliter le traitement post-collecte, 4) le développement d'un cadre de travail théorique pour la priorisation des suivis lors de la collecte, 5) le développement et l'analyse de différentes méthodes afin de prioriser le suivi des non-répondants d'enquêtes auprès des entreprises, 6) une simulation des stratégies de collecte afin d'améliorer l'efficacité et ce, sans recourir à des tests onéreux sur le terrain (microsimulation), 7) le développement d'un nouveau modèle permettant d'optimiser la répartition des tâches des interviewers (macrosimulation), et finalement 8) le développement d'un cadre de travail pour évaluer les effets des modes de collecte sur la qualité.

Recherche opérationnelle

Les divers projets du volet de la recherche opérationnelle en collecte des données ont pour but d'examiner et d'évaluer les processus et les pratiques de collecte des données, d'évaluer l'effet des nouvelles initiatives, d'élaborer de meilleurs moyens de fournir rapidement de la rétro-information et des données sur les progrès de l'enquête, et de cerner les possibilités stratégiques en vue d'améliorer la façon dont les enquêtes sont exécutées et gérées à Statistique Canada. Voici les principaux projets entrepris durant la période visée par le présent rapport.

Nous avons décrit les apprentissages qui ont découlé de la mise en œuvre réussie de plans de collecte adaptatifs (plans de collecte dynamiques) pour deux enquêtes par ITAO (Laflamme et St-Jean, 2011). Nous avons commencé à étudier la faisabilité de la mise en œuvre de plans de collecte adaptatifs pour les enquêtes par IPAO.

Nous avons développé et documenté de nouveaux indicateurs pour mesurer le rendement des intervieweurs travaillant par ITAO qui tiennent compte de la complexité du processus de collecte des données d'enquête et de la charge de travail de l'intervieweur (Laflamme, F. et St-Jean, H., 2011).

Étant donné qu'un nombre croissant d'enquêtes utiliseront un questionnaire électronique (QE) dans le contexte d'une collecte multi-mode, nous avons commencé à établir un programme de recherche pour le développement et l'étude des paradonnées sur la collecte en ligne.

Lignes directrices et normes pour les questionnaires électroniques

L'utilisation de modes de collecte multiples, en particulier l'emploi de questionnaires électroniques auto-administrés ou de la collecte en ligne, dans les enquêtes de Statistique Canada pose de nouveaux défis. Pour être certains que l'information recueillie réponde à nos exigences de qualité, des précautions doivent être prises durant l'élaboration du cadre conceptuel des questionnaires électroniques.

L'un des objectifs du projet de recherche est de résumer et de consigner les connaissances acquises jusqu'à présent en s'appuyant sur la recherche internationale et l'expérience de Statistique Canada en matière de collecte en ligne. Concrètement, il s'agira d'un guide pour la conception des questionnaires électroniques pour les enquêtes de Statistique Canada. Un autre objectif est de cerner les problèmes particuliers de conception des questionnaires électroniques qui doivent être évalués plus en profondeur. Par exemple, différents principes peuvent s'appliquer à nos enquêtes-entreprises et à nos enquêtes sociales; le choix peut dépendre du contenu de l'enquête, de la population cible et d'autres contraintes, comme les normes d'unité de présentation.

Nous avons continué de participer aux travaux du Comité des normes relatives aux questionnaires électroniques. Ce groupe de travail interdisciplinaire

et interdivisionnaire se penche sur les aspects conceptuels et techniques, tant existants que nouveaux, des enquêtes avec questionnaire électronique courantes ainsi que nouvelles de Statistique Canada qu'il est prévu de mettre à l'essai ou en production. La première version du manuel sur les *Lignes directrices et normes de conception de questionnaires électroniques* a été diffusée en 2011. Ce document continue d'évoluer à mesure que divers éléments de la conception des questionnaires électroniques sont améliorés ou introduits dans le processus.

En plus de ces travaux, le Centre de ressources en conception de questionnaires (CRCQ) continue de participer directement à l'élaboration et à la mise en œuvre des stratégies d'essais de convivialité pour les utilisateurs finaux pour plusieurs enquêtes-entreprises qui passent à l'environnement des questionnaires électroniques. Il s'agit d'un effort collectif comptant des représentants des principaux secteurs de Statistique Canada responsables de l'élaboration et de la conception des questionnaires électroniques.

Il s'agit d'un projet permanent. Nous prévoyons continuer à résumer et à consigner les connaissances acquises en nous appuyant sur la recherche internationale et sur l'expérience de Statistique Canada en matière de collecte en ligne.

Théorie des graphes et conception des questionnaires

L'objectif du projet est d'établir des règles structurelles pour la conception des questionnaires d'ITAO, ce qui faciliterait leur programmation et leur mise à l'essai, ainsi que l'analyse de la collecte des données. Nous utilisons des graphes pour appuyer cette approche structurelle, et les concepts de la théorie des graphes pour définir des outils diagnostiques pour évaluer la structure et la complexité des questionnaires déjà développés, ainsi que le fardeau de réponse qu'ils imposent. Un questionnaire pourrait être considéré comme un graphe orienté (Picard 1980), où les questions et conditions correspondent aux nœuds et où les flux représentent les chemins entre les nœuds.

Nos travaux de recherche se sont concentrés sur le module de l'emploi (EM) de l'Enquête sur l'accès et le soutien à l'éducation et à la formation (EASEF) menée par Statistique Canada en 2008. Nous avons élaboré une méthode systématique d'évaluation de la structure d'un questionnaire et défini les étapes en vue de l'améliorer. L'application de ces étapes a produit un module de questionnaire comprenant moins de questions, exempt de

répétitions, dont les principaux chemins étaient clairement définis et dont la couverture était facile à vérifier. Malgré la simplification, la quantité d'information qu'il permettait de recueillir demeurait la même, tandis que le fardeau de réponse éventuel était réduit grâce à la diminution du nombre de questions. Nous avons présenté nos résultats aux analystes de Statistique Canada, et une version abrégée de la communication intitulée « A structural approach to questionnaire design » (une approche structurée de la conception des questionnaires) a été présentée à la session annuelle des Joint Statistical Meetings, qui s'est tenue à Miami Beach, en Floride, en août 2011. Les deux exposés ont été bien accueillis. Depuis, nous avons examiné les logiciels disponibles pour faciliter la production de graphes pour les questionnaires existants et nous sommes en train de rédiger un article pour publication dans une revue. Les travaux ont été exécutés en collaboration avec le professeur Christina M.D. Zamfirescu, du Hunter College de la City University of New York.

Dans le cadre de futurs travaux, nous procéderons à la description théorique des transformations élémentaires qui mènent à la simplification du module sur l'emploi (ME) et donnerons la preuve que leur application peut réduire le nombre prévu de questions dans le questionnaire, tout en préservant le contenu analytique. Nous décrirons ces résultats dans un article de recherche. Les travaux entrepris avec la collaboratrice externe et deux de ses étudiants se poursuivront. Les travaux des étudiants porteront sur la simplification du module EC (Éducation formelle) du questionnaire de l'ASEF et l'établissement des sous-graphes d'information. Ces sous-graphes traduisent le transfert de l'information des nœuds parents (questions) à leurs successeurs. En collaboration avec un autre méthodologiste, nous dégagerons d'autres points faibles du questionnaire de l'ASEF qui peuvent mener à la découverte de nouvelles règles en vue d'améliorer sa structure et de faciliter l'imputation des données recueillies.

Plans de collecte adaptatifs

Nous avons établi un cadre théorique pour les plans de collecte adaptatifs dans le contexte des enquêtes avec interview téléphonique assistée par ordinateur. Par plan de collecte adaptatif, nous entendons toute procédure dynamique de détermination de l'ordre de priorité des appels et/ou de répartition des ressources qui tient compte de l'évolution de la collecte des données; autrement dit, la procédure s'appuie sur des paradonnées ou d'autres renseignements pour s'adapter à ce qui est observé durant la collecte des données. Nous nous sommes concentrés sur la détermination de l'ordre de priorité des appels. L'objectif d'un plan de collecte adaptatif est d'augmenter la qualité des données pour un coût donné ou de réduire le coût pour un niveau de qualité donné.

La littérature porte essentiellement sur la recherche de plans de collecte qui aboutissent à une réduction du biais de non-réponse d'un estimateur qui n'est pas corrigé pour la non-réponse. Donc, l'amélioration de la qualité est associée à la réduction du biais de non-réponse. Selon nous, ce critère n'est pas le meilleur, car le biais que l'on peut éliminer à l'étape de la collecte des données d'une enquête grâce à un plan de collecte adaptatif peut également être éliminé à l'étape de l'estimation grâce à une méthode appropriée de repondération pour tenir compte de la non-réponse. Nous proposons plutôt de minimiser la variance due à la non-réponse d'un estimateur qui est corrigé pour la non-réponse. Nous avons élaboré une procédure de détermination de l'ordre de priorité des appels qui vise à atteindre cet objectif. Notre théorie est décrite dans Beaumont, Haziza et Bocci (2011b) et nous rédigeons à l'heure actuelle un article qui sera présenté à la prochaine réunion du Comité consultatif des méthodes statistiques.

Répartition optimale/dynamique des opérations de suivi durant la collecte des données des enquêtes-entreprises

L'objectif de ce projet est d'élaborer une approche méthodologique efficace sur le plan des coûts et de la qualité pour sélectionner les unités qui feront l'objet d'un suivi durant la collecte des données des enquêtes-entreprises. L'approche méthodologique comprend deux grandes étapes : la répartition des unités nécessitant un suivi au niveau de la strate, suivie par la sélection des unités nécessitant un suivi à partir d'une méthode de sélection donnée. Il convient de souligner que le processus de répartition est exécuté de manière dynamique. Une nouvelle répartition a lieu à différentes périodes durant le processus de collecte en tenant compte des données sur les répondants reçues le plus récemment.

Les travaux effectués dans le cadre de ce projet ont mené au développement d'une nouvelle méthode de répartition des unités nécessitant un suivi fondées sur une méthode modifiée de répartition avec puissance (*Power Allocation*). Essentiellement, un nombre déterminé d'unités nécessitant un suivi (contrainte du processus de collecte) est affecté à chaque strate afin de minimiser la distance entre le CV cible de la strate et le CV réalisé. L'efficacité relative de cinq méthodes de sélection des unités nécessitant un suivi (quatre méthodes de sous-échantillonnage : EAS, PPTAS, PPTSYS, Bernoulli, ainsi qu'une méthode avec score de priorité dans laquelle la sélection est fondée sur un poids économique classé par ordre décroissant) est évaluée en

ce qui a trait au biais relatif et à l'erreur quadratique moyenne relative, ainsi que le coût associé à chacune d'elle.

Les programmes ont été développés/améliorés et les simulations ont été exécutées pour deux industries, à savoir la fabrication de produits métalliques et le commerce de détail. Les résultats sont les suivants :

- Les résultats préliminaires montrent qu'il est possible d'améliorer la qualité en procédant à une nouvelle répartition des unités sélectionnées pour le suivi entre les strates quand la taille de l'échantillon d'unités de suivi attribuée au départ est supérieure au nombre total d'unités admissibles dans une strate.
- L'imputation par donneur par la méthode du plus proche voisin ainsi que des méthodes de repondération (en utilisant des chiffres de population) ont été utilisées pour tenir compte de la non-réponse dans les simulations. Comme prévu, l'imputation par donneur par la méthode du plus proche voisin est celle qui a donné les meilleurs résultats.
- L'approche de répartition dynamique permet d'utiliser l'information auxiliaire disponible ainsi que les renseignements les plus récents sur la collecte pour répartir les unités sélectionnées pour le suivi pour chaque exécution de la collecte. Durant les opérations de suivi, la répartition dynamique peut être ajustée facilement d'après la différence entre le CV cible et le CV réalisé en fonction des exigences.
- L'accroissement du nombre d'unités devant faire l'objet d'un suivi (ou du coût) ne permettait pas toujours d'améliorer la qualité de l'estimation. Par conséquent, il est important d'appliquer de bonnes méthodes de suivi pour obtenir des estimations d'un certain niveau de qualité à un coût relativement faible.

En plus des travaux achevés susmentionnés, les quatre méthodes de sous-échantillonnage avec répartition dynamique ont été comparées dans un environnement simulé afin d'évaluer l'efficacité d'une approche méthodologique simple de sélection des unités devant faire l'objet d'un suivi. Dans cette approche, les unités sélectionnées pour le suivi sont réparties uniquement au début de l'opération de suivi, lorsque seule l'information initiale de la base de sondage est disponible. Ces travaux se poursuivront durant le prochain exercice.

Microsimulation de la collecte téléphonique

Ce projet de recherche consiste à construire un modèle de microsimulation qui représente le plus fidèlement possible le processus de collecte téléphonique des données d'une enquête. Le but du projet est d'étudier l'effet de la modification des paramètres de collecte, tels que la définition des tranches de temps et la répartition des intervieweurs, dans un environnement contrôlé. Le projet devrait permettre d'obtenir des résultats de manière plus efficace au moyen d'essais réels sur le terrain.

Le projet comprend deux grands volets. Premièrement, avant l'exécution de toute simulation, les paramètres du modèle sont calculés en utilisant les paradonnées d'enquête réelles. Le deuxième volet consiste à simuler la collecte. Les paramètres calculés durant le premier volet sont utilisés durant le deuxième pour attribuer un résultat et une durée à chaque tentative d'appel.

En 2009-2010, nous avons construit un prototype élémentaire en SAS Simulation Studio pour représenter la collecte de l'Enquête canadienne sur le don, le bénévolat et la participation. Les fonctions suivantes ont été mises en œuvre : tranches de temps, limite du nombre d'appel, horaires des intervieweurs et priorité des cas.

Durant la période visée par le présent rapport, un exposé a été fait au SAS Global Forum, à Las Vegas, sur le prototype en SAS Simulation Studio. Un article sur le sujet a également été rédigé (Bélanger, Couture et Neusy, 2011).

Le prototype SAS Simulation Studio a été adapté pour représenter la collecte d'une autre enquête, l'Enquête sur les ménages et l'environnement (EME). Un paramètre a été ajouté afin d'accroître la capacité du prototype de refléter la perte de productivité à mesure que le nombre de tentatives d'appel augmente.

En raison de certaines limites de SAS Simulation Studio, une équipe de la Division des systèmes d'information statistiques (DSIS) a entrepris de construire un modèle de simulation en utilisant Modgen. Le soutien pour ce projet a été fourni par la DSIS.

L'analyse des paradonnées d'enquête a mené à l'élaboration d'une méthode qui, selon nous, peut aider à répartir les intervieweurs entre les différentes enquêtes, sur la période de collecte de celles-ci. Nous avons rédigé un document décrivant l'analyse et la méthode proposée.

Macrosimulation

L'un des principaux défis que doit relever Statistique Canada est de soutenir des stratégies de collecte rentables tout en maintenant un haut niveau de qualité. La recherche sur les paradonnées a permis d'améliorer les processus et pratiques de collecte des données. Elle donne à penser qu'à l'heure actuelle, les ressources de collecte ne sont pas toujours réparties de manière optimale pour ce qui est de la charge de travail attribuée et de la productivité correspondante attendue. Des modèles ont été développés pour prédire la probabilité qu'un appel téléphonique aboutisse à un questionnaire rempli en fonction du moment de la journée et des ressources cumulatives dépensées durant une période de collecte particulière. Les paramètres estimés ont été entrés dans une fonction de perte qui optimise l'ordonnancement des appels sous des contraintes.

Durant la période visée par le rapport, plusieurs exposés sur la procédure proposée ont été donnés à l'atelier sur les modèles de microsimulation du MISS (Présentateurs : Hidiroglou et Laflamme 2011), aux Joint Statistical Meetings (Choudhry) et au Comité consultatif des méthodes statistiques (Présentateur : Hidiroglou 2011).

La procédure comprenait deux étapes. Dans la première, un modèle de régression logistique a été utilisé pour relier la probabilité d'une réponse dans une tranche de temps donnée au nombre cumulé d'appels jusqu'à cette tranche de temps inclusivement et aux variables indicatrices précisant la période de la journée de la tranche de temps en question. À la deuxième étape, nous avons utilisé les probabilités lissées produites d'après les valeurs prédites pour optimiser une fonction de perte qui représente le coût total sous les contraintes opérationnelles. Au départ, la procédure permettait d'optimiser la répartition de l'ITAO pour une seule enquête. Elle a maintenant été étendue afin de traiter plusieurs enquêtes simultanément. Des commentaires utiles ont été faits par les membres du Comité consultatif des méthodes statistiques. Ils nous ont encouragés à utiliser la microsimulation pour valider les résultats obtenus par cette procédure.

Études des effets de mode

La littérature démontrant que la modification de certains aspects administratifs d'une enquête peut altérer les résultats de l'enquête (par exemple estimations, variances et non-réponse) est abondante. L'un des aspects de l'administration des enquêtes auxquelles beaucoup d'attention est accordée à l'heure actuelle est le mode d'exécution de l'enquête (Dillman). Deux projets de recherche ont été exécutés afin de mieux comprendre l'incidence du mode d'exécution de l'enquête sur la qualité des données.

Le premier projet, l'élaboration d'un cadre pour les études des effets de mode, consistera d'abord à créer un cadre généralisé pour les effets de mode qui permettra de commencer à étudier les meilleures pratiques. Le cadre comprend la définition de l'effet de mode, un examen des méthodes permettant de l'évaluer et de déterminer s'il est applicable, et la détermination de mesures correctives éventuelles.

Des travaux collectifs ont été entrepris grâce à la création d'un groupe de travail sur les effets de mode chargé de définir l'effet de mode et d'élaborer des lignes directrices pour évaluer et mesurer cet effet sur les estimations des variables et sur l'erreur d'enquête totale (EET).

Le groupe a commencé à examiner l'abondante littérature dans ce domaine et a obtenu un consensus sur un cadre mathématique pour évaluer l'incidence de l'effet de mode sur l'erreur de mesure. Cette dernière est définie comme la différence entre la réponse à l'enquête et la réponse réelle. On parle d'un biais de mesure quand cette différence a lieu systématiquement dans une seule direction.

Le lien entre le mode et l'erreur de mesure et/ou le biais de mesure est défini en s'appuyant sur deux concepts. Quand tous les autres facteurs qui jouent un rôle sont maintenus constants, l'effet du mode sur la réponse à l'enquête est égal à

Concept 1 : la différence entre la réponse observée à une question en utilisant un mode de collecte donné et la réponse « réelle » inconnue (erreur de mesure) (Groves, 1989).

Concept 2 : la différence entre les résultats observés de l'enquête si l'enquête est administrée en utilisant différents modes de collecte des données (Couper, 2011).

Cependant, afin d'appliquer les deux concepts, le groupe a convenu que la définition formelle du « mode » n'était pas claire. De nouveau, le groupe est arrivé à un consensus et a déterminé qu'un « mode » est constitué de deux facteurs. Le premier est la présence ou l'absence de l'intervieweur (le messager) et le deuxième est le moyen grâce auquel la personne interrogée répond à l'instrument ou au questionnaire (le support). De cette façon, l'effet de mode peut être évalué suivant une approche de plan d'expérience permettant de le quantifier. Nous avons rédigé un document interne définissant explicitement ces concepts. L'année prochaine, nous continuerons de définir l'incidence du mode sur les erreurs dues à la sélection et à la non-réponse, et à consigner les méthodes d'analyse pour évaluer les effets de mode.

Notre deuxième projet de recherche a été entrepris afin de mieux comprendre les raisons d'un effet de mode. Certains faits saillants de la littérature sur le sujet ont été consignés dans un document de travail (Binder, 2012).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Hélène Bérard (613 951-1458, helene.berard@statcan.gc.ca).

Bibliographie

Couper, M.P. (2011). The future of modes of data collection. *Public Opinion Quarterly*, Vol (volume). 75, Nº (numéro). 5, 889-908.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York : John Wiley & Sons, Inc. (incorporated)

Picard, C.F. (1980). Graphs and Questionnaires, North-Holland Mathematics Studies. Vol (volume). 32.

Centre de ressources sur le contrôle de la divulgation

Dans le cadre de son mandat, le Centre de ressources sur le contrôle de la divulgation (CRCD) a fourni aux programmes de Statistique Canada des conseils et un soutien concernant les méthodes d'évaluation et de contrôle du risque de divulgation. Il a également partagé de l'information et des conseils sur les pratiques de contrôle de la divulgation avec d'autres ministères et organismes, dont le Secrétariat du Conseil du Trésor du Canada, l'Agence canadienne d'inspection des aliments, l'Office of Statistics and Information du gouvernement de l'Alberta, le Centre for Disease Control de la Colombie-Britannique, Istat (Italie) et l'Office for National Statistics (Royaume-Uni). Statistique Canada a également participé, avec cinq autres pays, à un projet de collaboration internationale sur le contrôle de la divulgation.

Un soutien continu concernant les méthodes de contrôle de la divulgation est également offert au programme des centres de données de recherche (CDR) de Statistique Canada. La majorité du soutien prend la forme d'une aide pour l'application et l'interprétation des règles de contrôle de la divulgation ayant trait aux fonds de données des CDR (centres de données de recherche).

Méthodes de contrôle de la divulgation dans un environnement d'accès à distance en temps réel (ADTR)

Une version améliorée de l'algorithme d'arrondissement contrôlé maintenant l'additivité (ACROUND) a été testée et mise en œuvre dans le système d'ADTR. L'algorithme d'arrondissement trouve de meilleures solutions plus rapidement et inclut un taux d'additivité dans les tableaux de sortie. Des règles de contrôle de la divulgation ont été élaborées pour le prochain ensemble de statistiques qui seront offertes, à savoir les centiles, les moyennes, les proportions, les ratios, les coefficients Gini et les moyennes géométriques. Les méthodes de contrôle de la divulgation sont fondées sur une combinaison de deux méthodes : une règle découlant d'un nombre minimal de répondants pour la statistique en question dans la cellule et l'arrondissement. Des indicateurs de qualité ont également été développés.

Des communications ont été présentées au 58e congrès de l'Institut international de statistique (Simard 2011a) et à la réunion de travail CEE-

ONU/Eurostat sur la confidentialité des données statistiques (Simard, 2011b).

Utilisation d'une méthode alternative avec G-Confid

Certaines enquêtes rencontrent des problèmes avec la macro Suppress de G-Confid dans des situations extrêmes (tableaux très volumineux). Par exemple, (i) la macro arrête après plusieurs heures de traitement avec un message d'erreur fatale, (ii) la macro semble être prise dans une boucle infinie, (iii) la limite de mémoire de l'ordinateur est atteinte. Afin de satisfaire les besoins des clients, des alternatives doivent être envisagées. Deux approches pour la suppression de cellules sont en développement. Une fonction de score est une méthode heuristique souvent utilisée pour déterminer les cellules à supprimer dans un tableau. Par exemple, un score plus élevé (plus de chance à être supprimé) est assigné à une cellule qui peut servir à protéger plusieurs cellules confidentielles à la fois. L'approche séquentielle, quant à elle, permet de diminuer la complexité d'optimisation (problème LP) en divisant le problème en plusieurs parties. Cette approche diminue le nombre de dimensions du problème tout en tentant de protéger adéquatement les données entre eux.

Un document préliminaire décrivant la stratégie développée pour l'approche séquentielle a été écrit. Plusieurs ensembles de données ont été utilisés pour évaluer la performance de cette approche. Les résultats montrent qu'il est possible de produire un patron de suppression plus rapidement tout en assurant une protection adéquate. Cependant, un ensemble de données montre quelques cellules non protégées. Une investigation plus détaillée est nécessaire pour en connaître la cause.

Suppression séquentielle de cellules dans les tableaux appariés

Ces travaux de recherche ont également porté sur des méthodes permettant d'aborder le processus séquentiel de suppression de cellules dans le contexte de plusieurs tableaux appariés. Ce processus s'appuie sur le profil de suppression provenant des anciens tableaux pour « forcer » les profils de suppression appropriés dans une exécution séquentielle de tableaux en utilisant G-Confid. La performance de la méthode a été évaluée en se servant de données simulées.

Nous avons simulé des données pour tester plusieurs hypothèses que nous avons faites quant aux difficultés que pose l'exécution d'une suppression séquentielle. Nous voulions déterminer si le nombre total de cellules supprimées dépend de l'ordre des suppressions. L'une de nos hypothèses initiales était que la suppression primaire était indépendante de l'ordre. De plus, nous voulions voir quels facteurs, tels que la distribution, étaient associés à ce problème. Nous avons également examiné les liens avec l'analyse des données et la théorie des graphes qui pourraient éventuellement fournir de nouveaux éclaircissements sur ce problème. Les premières simulations ont montré que la dépendance à l'égard de l'ordre semblait être liée à la suppression secondaire (comme nous l'avions d'abord supposé). En outre, la dépendance à l'égard de l'ordre est liée à l'asymétrie (et éventuellement au kurtosis) de la distribution et aux valeurs extrêmes présentes dans les données. Nous avons rédigé un rapport provisoire et écrit des programmes SAS généralisables pour ce projet (Fox et Post, 2012).

Autre méthode d'estimation de la variance pour les fichiers de microdonnées à grande diffusion

Les poids de sondage peuvent comprendre de l'information discernable sur le plan d'échantillonnage que les utilisateurs pourraient employer pour accroître leur capacité à identifier des personnes dont les réponses figurent dans un fichier de microdonnées à grande diffusion (FMGD). L'étude avait pour objectif de comparer l'estimation de la variance en utilisant les poids sous un plan d'échantillonnage stratifié avec celles utilisant des poids fondés sur un échantillonnage de Poisson, comme il est décrit dans Beaumont et Patak (2010). Les résultats des simulations ont été consignés dans un rapport interne (Wright et Beaumont, 2012).

Création de données synthétiques

Il existe une documentation naissante sur les méthodes de création de données synthétiques ou simulées. Ces méthodes visent à préserver autant que possible les relations entre les données originales, tout en maintenant à un faible niveau le risque de divulgation de renseignements confidentiels. Les méthodes courantes comportent la modélisation des relations multivariées dans les données recueillies, de façon à reproduire ces relations observées dans les données synthétiques.

Nous avons achevé de synthétiser des données pour le projet du Cross National Equivalent File (CNEF), c'est-à-dire que nous avons synthétisé les données pour les quatre dernières années du panel de six années, et nous avons rédigé la documentation connexe décrivant les modèles en détail. Six pays participent au projet. Le volet canadien du CNEF est un sous-ensemble d'environ 40 variables tirées de l'Enquête sur la dynamique du travail et du revenu. Notre méthodologie et certaines validations empiriques transversales sont décrites dans un rapport provisoire (Beaumont et Bocci, 2012). Les données synthétiques mènent à des conclusions similaires à celles obtenues avec les données réelles, mais il peut exister des divergences. Celles-ci pourraient être éliminées en perfectionnant les modèles utilisés pour produire les données synthétiques. Nous prévoyons mettre à jour notre article et le soumettre pour publication dans une revue à comité de lecture. Cette année, nous avons procédé à certaines validations empiriques longitudinales supplémentaires (Mach et Moussa, 2012). Malheureusement, ces nouveaux résultats sont moins prometteurs et des travaux de modélisation plus approfondis seraient nécessaires en vue d'améliorer nos données synthétiques pour ce genre d'analyse.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Jean-Louis Tambay (613 951-6959, jean-louis.tambay@statcan.gc.ca).

Bibliographie

Beaumont, J.-F., et Patak, Z. (2010). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. Document de travail de la méthodologie.

Recherche sur le couplage d'enregistrements

Les ressources affectées au couplage d'enregistrements ont été consacrées à la résolution des problèmes auxquels faisaient face les chercheurs, les

utilisateurs des données et les analystes, à Statistique Canada ainsi qu'à l'extérieur de l'organisme. Notre financement tient compte des répercussions sur l'accessibilité et la compréhension des sources de données couplées, sur la qualité et la mesure de l'erreur des sources de données couplées et sur la pertinence découlant de l'amélioration des méthodes analytiques. Les personnes qui accomplissent cette recherche acquièrent aussi de l'expérience en transfert et en échange de connaissances en publiant des articles techniques et en donnant des séminaires, des exposés et des cours.

Estimation de l'erreur dans le couplage d'enregistrements : algorithme espérance-maximisation (EM)

Une nouvelle méthodologie a été développée pour le couplage probabiliste d'enregistrements à l'aide de G-Coup (G-LINK). Ce système produit moins d'erreurs de couplage, donne des estimations de l'erreur plus exactes et exécute les couplages plus rapidement dans les projets à grande échelle. Il est fondé sur la généralisation de l'algorithme espérance-maximisation (EM) qui comprend un modèle de critères de groupement, de nouveaux moyens de mesurer les erreurs de couplage et de nouvelles règles matricielles avec concordance partielle. Nous avons également comparé les propriétés des algorithmes EM (espérance-maximisation) sur un ensemble de données synthétiques en utilisant Octave (un logiciel libre semblable à Matlab). Les résultats montrent que l'algorithme amélioré produit des estimations plus exactes et des régions d'examen manuel plus petites que les algorithmes s'appuyant sur l'hypothèse d'indépendance conditionnelle. Des règles personnalisées pour les comparaisons matricielles ont été élaborées afin de contourner les limites des comparaisons matricielles intégrées dans G-Coup, y compris des lignes directrices concernant la pondération par les fréquences d'observation (voir Dasylva, 2011). Ces résultats de recherche ont été décrits dans trois rapports internes (Dasylva 2010a, Dasylva 2010b et Dasylva 2011).

Estimation des erreurs de couplage probabiliste

L'algorithme EM (espérance-maximisation) sous indépendance conditionnelle a été développé conjointement avec Abel Dasylva comme un module indépendant de G-Coup. Une interface entre G-Link et SAS a été créée. Cela permet de contourner le problème de boîte noire que constitue G-Coup. L'algorithme EM (espérance-maximisation) a été implémenté à l'aide de macros SAS. Il a été testé sur un fichier de taille modeste (70.000 x 70.000) et comparé à l'algorithme itératif ad-hoc de G-Coup. Le temps d'exécution de l'algorithme EM (espérance-maximisation) est tout à fait acceptable. Le calcul des poids de base de niveaux sous EM (espérance-maximisation) permet aussi l'estimation directe des taux d'erreurs empiriques pour les faux positifs et les liens manqués. Des macros SAS ont été écrites pour les calculer de même qu'une procédure graphique sous SAS a été implémentée de manière à déterminer graphiquement les seuils pour des taux d'erreurs fixés. Les macros SAS et l'amélioration de la méthode probabiliste ont été appliquées au projet de couplage entre les données de santé du registre Ontario et celles du recensement de la population de l'Ontario (voir Hortop, Saïdi and Reicker, 2011).

Comprendre l'erreur de mesure dans les ensembles de données d'enquête couplées : extension des méthodes de régression applicables aux données d'enquête et évaluation d'une approche d'imputations multiples pour l'estimation de la variance – Partie 1

Nous avons étudié les propriétés théoriques des estimateurs de Scheuren et Winkler (1993), Lahiri et Larsen (2005) ainsi que celui de Chambers (2009). Aussi, nous comparons ces estimateurs par le biais d'une simulation. Dans un premier temps, nous reprendrons les simulations effectuées par Chambers. La présence d'erreurs de couplage dans le fichier de données couplées distord le lien qui existe entre les variables X et Y provenant de fichiers différents, ce qui aura une répercussion sur les estimateurs calculés. L'atténuation de la relation sera d'autant plus prononcée que le lien entre X et Y sera fort, ce lien sera observable à travers le coefficient de corrélation entre ces deux variables. Dans un second temps, nous observerons comment se comportent les estimateurs en fonction de la force du coefficient entre X et Y. Le biais relatif de la constante pour les estimateurs de Scheuren et Winkler ainsi que ceux de Lahiri et Larsen s'écartent plus du biais relatif de la constante du vrai beta que celui de l'estimateur de Chamber. On peut dire que l'estimateur de Chambers est sensiblement moins biaisé que ceux de Scheuren et Winkler ainsi que ceux de Lahiri et Larsen.

En outre, les travaux ont débuté en vue d'appliquer au couplage d'enregistrements le cadre proposé par Schouten, van den Brakel et Klausch (2011) pour décomposer les effets de mode en leurs composantes en se basant sur l'erreur de mesure et les erreurs de sélection. Cette application a pour objectif de déterminer s'il est possible de décomposer la différence entre les rapports de cotes calculés pour les ensembles de données couplées et de données non couplées pour obtenir les composantes provenant des erreurs de mesure (p. ex. (par exemple) erreurs typographiques) et du pouvoir discriminatoire d'une variable. Certains résultats de la recherche font partie d'une thèse de maîtrise en cours de préparation (F. Bene Tchaleu sous la supervision de D. Haziza, 2011).

Pour obtenir plus de renseignements, veuillez communiquer avec :
Karla Fox (613 951-4624, karla.fox@statcan.gc.ca).

Bibliographie

Chambers, R. (2009). Regression analysis of probability-linked data. *Statisphere*, 4, Official Statistics Research Series, Statistics New Zealand.

Dasylva, A. (2010a). COS Step 2: Estimating events odds ratios with an EM (Expectation Maximization) algorithm for probabilistic record linkage. Rapport interne, DMES (Division des méthodes d'enquêtes sociales), décembre.

Dasylva, A. (2010b). COS Step 2: Calibration of error rates and adjustment of weight thresholds. Rapport interne, DMES (Division des méthodes d'enquêtes sociales), décembre.

Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*,100, 222-230.

Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*,19, Nº (numéro). 1, 45-65.

Schouten, B., van den Brakel, J. et Klausch, T. (2011). Mode Effects in Social Surveys. Document de travail de Statistics Netherland, février.

Activités de soutien

Séries chronologiques

Les projets se répartissent en sept sous-catégories :

- Consultation concernant les séries chronologiques (y compris l'élaboration de cours)
- Traitement et désaisonnalisation des séries chronologiques
- Étalonnage (y compris le développement et le soutien du logiciel G-Series)
- Réconciliation (y compris le développement et le soutien du logiciel G-Series)
- Calendarisation
- Estimation de la tendance
- Autres projets de R-D sur les séries chronologiques

Consultation concernant les séries chronologiques (y compris l'élaboration de cours)

Dans le cas de son mandat, le Centre de recherche et d'analyse en séries chronologiques a offert des consultations sur demande. Les sujets les plus fréquemment abordés durant la période visée par le rapport comprennent le traitement des discontinuités historiques dans les séries, diverses techniques de projection et l'analyse des variations possibles de la tendance saisonnière.

Le personnel du Centre a donné divers cours et ateliers d'introduction. Le cours intitulé Introduction aux composantes des séries chronologiques et à la désaisonnalisation (H-0431) a été entièrement révisé et donné pour la première fois depuis 2004.

Les membres du Centre ont révisé divers articles pour des publications officielles de Statistique Canada ou à titre d'examinateurs pour des revues externes à comité de lecture.

Traitement et désaisonnalisation des séries chronologiques

L'objectif de ce projet est de surveiller les activités de haut niveau reliées au soutien et au développement d'un système de traitement des séries chronologiques. La désaisonnalisation est effectuée en utilisant X-12-ARIMA (pour l'analyse et le développement ou la production) ou Proc X12 de SAS (pour la production).

Le module d'étalonnage (qui utilise le logiciel G-series/Forillon) est maintenant intégré dans le système de traitement des séries chronologiques. Ses fonctions ont été étendues pour permettre l'étalonnage « sur demande » afin de pouvoir appliquer correctement divers ajustements (pour un point particulier dans le temps ou une période de couverture plus longue). Nous avons également intégré un module qui permet d'ajuster correctement les niveaux annuels bruts des séries en dollars constants désaisonnalisées.

Un examen horizontal de la désaisonnalisation à Statistique Canada a été effectué (Julien et Fortier, 2011). Les résultats provisoires et les bonnes pratiques ont été présentés au Congrès mondial de la statistique de 2011 de l'IIS (Fortier, Julien et Quenneville, 2011). Susie Fortier a participé à une table ronde qui s'est tenue à l'Office of National Statistics, à Newport, au Royaume-Uni, en août 2011. Les thèmes de discussion comprenaient les programmes de recherche courants, le traitement des valeurs aberrantes et les événements inhabituels, ainsi que X-13.

Nous avons procédé à une analyse empirique en vue d'évaluer l'effet de l'utilisation de prévisions pour produire des données désaisonnalisées forcées de concorder avec les totaux annuels (tableau D 11.A) et sommes en train de la documenter pour produire un document de travail ou éventuellement faire une communication à une conférence (Wu et Ferland, 2011).

Étalonnage (y compris le développement et le soutien du logiciel G-Series)

L'étalonnage s'entend des techniques utilisées pour s'assurer de la cohérence entre les données chronologiques ayant trait à une variable cible mesurée à diverses fréquences, par exemple, infra-annuellement et annuellement.

La version 1.04 de PROC BENCHMARKING a été développée et certifiée. Cette nouvelle version comprend l'utilisation d'un coefficient d'altérabilité pour la série indicatrice et/ou la série repère.

Un article sur l'utilisation de l'étalonnage dans le contexte de la désaisonnalisation a été révisé (Quenneville et Fortier, 2011b).

L'élaboration de tests de compatibilité s'est poursuivie et a été documentée. La technique a d'abord été présentée au Comité consultatif des méthodes statistiques, puis mise à jour et soumise pour publication (Quenneville et Gagné, 2011).

L'exploration et la description des techniques d'étalonnage du point de vue du modèle de l'erreur se poursuivent (Chen et Wu, 2011).

Réconciliation (y compris le développement et le soutien du logiciel G-Series/Forillon)

La réconciliation est une méthode utilisée pour imposer des contraintes d'agrégation simultanées à des tableaux de séries chronologiques, de sorte que les sommes des « cellules » des séries soient égales à la série appropriée de « totaux de marge ».

La version 1.04 de PROC TSRAKING a été certifiée. La nouvelle version comprend des modifications mineures qui donnent plus de souplesse à la vérification des données de sortie.

La Division de l'ingénierie des systèmes a mis en œuvre pour PROC TSRAKING de nouvelles spécifications s'appuyant sur la terminologie de l'optimisation sous contrainte dans un prototype de validation de concept qui utilise PROC OPTMDL. Cette configuration, qui est une option de rechange pour les formules matricielles courantes, devrait raccourcir le temps de traitement et pourrait résoudre les problèmes liés aux solutions négatives inacceptables.

Le plus souvent, la réconciliation des séries chronologiques est appliquée à des données désaisonnalisées pour résoudre les incohérences entre un ajustement direct et un ajustement indirect. Cette utilisation est présentée dans une version révisée de l'article de Mazzi, Fortier et Quenneville (2011).

Un article de synthèse rédigé pour résumer les travaux de développement des méthodes et des outils d'étalonnage et de réconciliation effectués ces dernières années (Quenneville et Fortier, 2011) a été révisé et accepté pour publication dans un recueil en l'honneur de David Findley.

Calendarisation

Cette sous-catégorie englobe les travaux de soutien et de développement de méthodes de calendarisation tant sous l'angle de l'étalonnage que des techniques plus récentes et très prometteuses d'interpolation

Les techniques d'interpolation par spline pour la calendarisation ont été présentées dans Quenneville, Picard et Fortier, 2010. Les auteurs sont en train de réviser et d'étendre les travaux afin d'inclure une comparaison avec l'approche des espaces d'états (Quenneville, Picard et Fortier, 2012).

Estimation de la tendance

Une autre façon de présenter les tendances globales et, fait plus important, les tendances locales a été élaborées et décrite dans un article rédigé par Bemrose, Quenneville et Meszaros (2011). La tendance locale est estimée en utilisant une approche fondée sur un filtre de lissage robuste inspiré de l'algorityme X-11 pour la désaisonnalisation.

Autres projets de R-D sur les séries chronologiques

La modélisation est un autre domaine de la recherche sur les séries chronologiques, et les tests de la racine unitaire sont un élément central de l'identification du modèle ARIMA. En collaboration avec des professeurs de l'Université chinoise de Honk Kong et de l'Université normale de Hangzhou, nous avons accompli de grands progrès en ce qui concerne l'utilisation du variogramme pour obtenir un test de la racine unitaire (Chen, Wu and Wang, 2012).

En vue d'appuyer la publication de Wyman (2010), nous avons rédigé un article sur l'amplification (*gain shift*) et le déphasage (Quenneville et Findley, 2011).

Nous avons également procédé à un examen des méthodes de modélisation avancées recourant à la co-intégration et à des fonctions de transfert en vue d'une application éventuelle aux données environnementales.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Susie Fortier (613 951-4751, susie.fortier@statcan.gc.ca).

Bibliographie

Quenneville, B., Picard, F. et Fortier, S. (2010). Interpolation, Benchmarking and Temporal Distribution with Natural Splines. *Proceedings of the Business and Economic Section*, American Statistical Association.

Wyman, D. (2010). Seasonal adjustment and identifying economic trends. Canadian Economic Observer, Statistique Canada, mars 2010, № (numéro). 11-010-X au catalogue.

Centre de ressources en couplage d'enregistrements (CRCE)

Les objectifs du Centre de ressources en couplage d'enregistrements (CRCE) sont les suivants :

- Offrir des services de consultation aux utilisateurs internes et externes des méthodes de couplage d'enregistrements, y compris des recommandations concernant les logiciels et les méthodes à utiliser et des travaux concertés sur les applications de couplage d'enregistrements.
- Évaluer d'autres méthodes de couplage des enregistrements et améliorer les méthodes existantes.
- Évaluer les progiciels de couplage d'enregistrements et, au besoin, développer des prototypes de logiciel intégrant des méthodes non offertes dans les progiciels existants.
- Faciliter la diffusion de l'information concernant les méthodes, les logiciels et les applications de couplage d'enregistrements aux personnes intéressées à l'intérieur et à l'extérieur de Statistique Canada.

Voici une liste de nos activités en 2011-2012

- Continuation de l'offre d'un soutien à l'équipe de développement du système G-coup, y compris la participation aux réunions du groupe d'utilisateurs du couplage d'enregistrements. Recueil d'information sur l'indexation et les interrogations SQL efficaces pour l'étape de présélection et collaboration avec la Division de la statistique de la santé et l'équipe de G-Coup en vue d'améliorer la performance de G-Coup et de résoudre les questions de traitement par lot.
- Soutien pour déceler les problèmes informatiques et méthodologiques et améliorer la vitesse d'exécution de G-Link sur de gros projets utilisant des fichiers comme le recensement de la population ou le registre de l'assurance maladie de l'Ontario.
- Travail sur le couplage des données DLAS et utilisation de ce projet comme une occasion de tester sur le terrain les nouvelles fonctions de G-Coup, et élaboration d'approches plus systématiques et ayant un meilleur fondement théorique pour définir et ajuster les couplages d'enregistrements.
- Offre de services de consultation (étapes pour avoir accès aux données pour le couplage d'enregistrements, fourniture de documentation et d'information sur les politiques de couplage d'enregistrements) et de services associés aux probabilités calculées en utilisant G-coup pour choisir les meilleures règles et seuils.
- Organisation d'un atelier sur le couplage d'enregistrements pour le symposium international sur les questions de méthodologie de 2011. Le Centre de ressources en couplage d'enregistrements (CRCE) et les plus importants projets de couplage d'enregistrements mis en œuvre à Statistique Canada ont été présentés.
- Mise à jour du cours français H419 sur le couplage d'enregistrements en accord avec la nouvelle version 2.3 du logiciel G-Coup.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Abdelnasser Saidi (613 951-0328, abdelnasser.saidi@statcan.gc.ca).

Activités de soutien du centre de ressources en analyse de données (CRAD)

Le Centre de ressources en analyse de données (CRAD) est le centre interne de ressources en analyse de données de Statistique Canada dont le mandat est de proposer et de fournir de bonnes méthodes et de bons outils pour l'analyse des données de Statistique Canada et de promouvoir leur application. Bien que les activités de soutien du CRAD (Centre de ressources en analyse de données) soient financées par un certain nombre de sources, le présent rapport porte principalement sur les travaux imputés au fonds global de financement de la Direction de la méthodologie. Le CRAD (Centre de ressources en analyse de données) (Centre de ressources en analyse de données) consacre aussi une partie des ressources provenant de ce fonds à des travaux de recherche. Ces activités de recherche sont aussi financées partiellement par le Projet de recherche en analyse de données (1919) et sont décrites dans le rapport ayant trait à ce projet.

Consultations

En 2011-2012, le CRAD (Centre de ressources en analyse de données) a donné un grand nombre de consultations sur les méthodes statistiques aux analystes de nombreux programmes de Statistique Canada. En plus de conseils sur une approche méthodologique, bon nombre de consultations comprenaient aussi un soutien pour le choix et l'utilisation d'un logiciel approprié pour une analyse. Bon nombre de problèmes avaient trait à des approches assez classiques d'analyse des données d'enquête, telles que l'estimation et l'inférence liées à des statistiques descriptives, ainsi que l'ajustement de modèles linéaires et logistiques à des données d'enquête. Cependant, d'autres consultations avaient trait à des extensions des

méthodes classiques et à de nouvelles questions plus complexes, comme l'intégration des données provenant de plus d'une enquête dans une seule analyse et les tests de signification des différences entre les médianes pour différents groupes de population. Certaines consultations ne portaient pas sur les données d'enquête, comme celles sur les méthodes pour comparer des groupes non expérimentaux en vue de fournir des données probantes pour les évaluations de programme.

De nouveau cette année, nous avons donné un grand nombre de consultations à d'autres méthodologistes qui effectuaient eux-mêmes des analyses ou qui offraient un soutien aux analystes des données d'enquête des divisions spécialisées dont ils avaient la responsabilité. En est un exemple le soutien offert aux méthodologistes de l'ECMS pour étudier comment modifier les méthodes d'analyse afin de tenir compte du nombre limité de degrés de liberté pour l'estimation de la variance (qui est décrit plus en détail dans le rapport sur la recherche en analyse de données). Un autre exemple est l'évaluation de l'utilité analytique des données longitudinales synthétiques créée pour le volet canadien du projet du Cross National Equivalent File (voir Mach et Moussa, 2012). Un dernier exemple est celui des consultations offertes aux méthodologistes rattachés à l'EPA, l'ESG et certaines enquêtes-entreprises

en vue de concevoir des expériences qui pouvaient être intégrées dans leurs enquêtes pour évaluer l'effet des changements de mode d'exécution des enquêtes.

Plusieurs consultations se sont aussi adressées à des personnes travaillant à l'extérieur de Statistique Canada et au réseau des CDR (centres de données de recherche). L'une d'elles a été offerte à des chercheurs de l'Institut de recherche en services de santé qui souhaitaient obtenir de l'aide pour utiliser les données de l'ESCC couplées à celles des bases de données administratives de Santé Ontario pour développer des modèles de prédiction du décès et des maladies chroniques.

Nos activités de consultation comportaient aussi la révision technique et l'examen d'articles destinés à des revues internes ainsi qu'externes, dont Tendances sociales canadiennes, le Recueil du Groupe des méthodes d'enquête de la SSC, les documents de travail de la Direction de la méthodologie et le Bulletin technique et d'information des CDR (centres de données de recherche).

Activités de formation

Durant l'exercice 2011-2012, le CRAD (Centre de ressources en analyse de données) a donné le cours 0438, intitulé Analyse statistique des données d'enquête. Les deux parties (0438A et 0438B) ont été données en anglais ainsi qu'en français. Du matériel de cours supplémentaire a été élaboré. Certains méthodologistes n'appartenant pas au CRAD (Centre de ressources en analyse de données) ont donné le cours.

Les membres du CRAD (Centre de ressources en analyse de données) ont participé à des séances de remue-méninges pour chaque séance de l'atelier sur l'interprétation des données, à l'examen de projets et à la présentation d'un séminaire sur l'analyse des données d'enquête. Un suivi a eu lieu auprès de certains participants qui souhaitaient une aide supplémentaire pour leurs projets d'analyse.

Dans la série de séminaires s'adressant aux nouveaux employés, le CRAD (Centre de ressources en analyse de données) a donné un séminaire sur l'analyse des données d'enquête.

Plusieurs séminaires téléphoniques à l'heure du dîner ont été offerts aux analystes des CDR (centres de données de recherche) sur des sujets analytiques qui les intéressaient. En outre, deux ateliers d'une demi-journée, ayant pour thème « enquête complexe : utilisation des poids d'échantillonnage et des poids bootstrap » et « enquête complexe : combiner des données similaires provenant de plus d'une enquête ou de plus d'un cycle », ont été donnés à la conférence annuelle des chercheurs des CDR (centres de données de recherche), à Edmonton, en octobre 2011. La plupart du temps de préparation de ces séminaires était couvert par le budget des CDR (centres de données de recherche), mais un petit nombre d'heures a été imputé au CRDM puisque le matériel produit est utilisé à l'extérieur des CDR (centres de données de recherche).

Un article sur la pondération appropriée pour l'analyse des fichiers de données du questionnaire complet du recensement sera bientôt publié (Roberts, 2012).

Le CRAD (Centre de ressources en analyse de données) examine à l'heure actuelle les possibilités d'« apprentissage à distance » qui pourraient être utilisées afin d'offrir plus efficacement une formation aux employés de Statistique Canada, aux chercheurs des CDR (centres de données de recherche) et à d'autres.

Évaluation, développement et promotion de logiciels

Durant l'exercice 2011-2012, le CRAD (Centre de ressources en analyse de données) a poursuivi l'évaluation de divers progiciels du commerce en vue de déterminer s'ils étaient adaptés à l'analyse des données d'enquête de Statistique Canada. Une des activités a consisté à se tenir au courant des nouvelles fonctions des logiciels qui permettent d'utiliser les poids de sondage bootstrap dans l'analyse fondée sur le plan de sondage (p. ex. (par exemple) SUDAAN 10, Stata 11 et SAS 9.2). Un séminaire sur certaines fonctions de SUDAAN a été présenté à la Division de l'analyse de la santé.

Nous avons achevé la première version d'un document axé sur la pratique à l'intention des nouveaux chercheurs en guise d'introduction à la mise en œuvre de l'estimation pondérée et de l'estimation bootstrap de la variance au moyen de divers progiciels (voir Gagné, Roberts et Keown 2012). Ce document avait été demandé par les analystes des CDR (centres de données de recherche) (et sa préparation a été financée principalement par les ressources de soutien des CDR (centres de données de recherche)), mais il sera utile à d'autres chercheurs qui analysent les données de Statistique Canada.

Autres activités

Wei Lin, un étudiant qui a travaillé au CRAD (Centre de ressources en analyse de données) à l'automne 2010 dans le cadre du programme de stage de MITACS, a maintenu certains liens avec nous en révisant et en ajoutant des discussions à des articles sur les expériences intégrées dans des enquêtes ici à Statistique Canada, activité qui était le sujet de son projet MITACS.

Karla Fox a co-organisé une étude de cas intitulée « Mark-recapture on Atlantic cod (*Gadus morhua*) off Eastern and Southern Newfoundland : Estimation of Exploitation Rate and Measuring Growth » pour le congrès annuel de 2012 de la SSC. Les groupes d'étudiants inscrits pour l'étude de cas seront en compétition pour un prix.

Georgia Roberts a siégé aux comités de rédaction de Rapports sur la santé, de Tendances sociales canadiennes et du Bulletin technique et d'information.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Georgia Roberts (613 951-1471, georgia.roberts@statcan.gc.ca).

Systèmes généralisés

Contrôle de la divulgation : algorithme Shuttle

L'algorithme Shuttle est une méthode qui permet de déterminer l'intervalle des valeurs possibles que des variables données peuvent prendre sous

certaines contraintes. Il s'agit d'une méthode heuristique très rapide qui produit d'excellents résultats dans la majorité des cas. La recherche a pour objectif d'adapter l'algorithme en vue de produire des valeurs maximale et minimale pour les cellules supprimées dans un tableau de données. À l'heure actuelle, G-Confid utilise une fonction d'optimisation pour déterminer ces valeurs, mais le temps de calcul peut être long. L'algorithme Shuttle pourrait être envisagé comme méthode de rechange (approximative) dans la macro Audit, ou comme méthode en vue de réduire le nombre d'appels de la fonction d'optimisation.

Nous avons développé un prototype de l'algorithme Shuttle. Nous avons rédigé un rapport (Xi, 2011) qui décrit la recherche et la méthodologie relatives à cet algorithme. Le rapport comprend aussi un résumé des résultats obtenus à l'aide du prototype.

Une version modifiée de la macro Audit de G-Confid a été développée par la Division de l'ingénierie des systèmes. Cette version permet à l'utilisateur d'entrer les bornes à partir de l'algorithme Shuttle. Les essais ont montré que la macro Audit s'exécute plus rapidement en utilisant ces bornes qu'en ne les utilisant pas, dans le cas où les bornes initiales sont asymétriques.

Contrôle de la divulgation : Arrêt quand il n'y a pas de solution

À l'occasion, G-Confid n'arrive pas à trouver une solution. Dans ce cas, le système arrête immédiatement le traitement et envoie un message d'erreur à l'utilisateur. À l'heure actuelle, l'utilisateur n'a aucun moyen de déterminer la cause du problème. L'objectif de ce projet de recherche est de trouver une approche pour déterminer non seulement l'origine du problème, mais aussi un moyen de le résoudre. Cette approche pourrait être utilisée pour éviter de perdre des heures, même des jours, de temps de traitement dans la macro Suppress de G-Confid.

Les essais ont montré que la version 9.3 de SAS est nettement supérieure à la version 9.2. Cependant, le problème de la capacité du système informatique persiste. Une solution éventuelle consisterait à diviser le problème en plus petits morceaux.

Contrôle de la divulgation : renonciations

Certains répondants aux enquêtes-entreprises ont signé des ententes, appelées renonciations, qui permettent à Statistique Canada de publier leurs données même si elles sont confidentielles. Dans le cas de tableaux de données, les renonciations permettent la publication de cellules qui, autrement, seraient supprimées afin de respecter leur caractère confidentiel. Cependant, il faut s'assurer que les données confidentielles de tous les autres répondants sont protégées dans les cellules publiées. Ce projet de recherche consiste à élaborer une méthode qui tient compte des cellules contenant les répondants qui ont signé une renonciation dans le calcul de la sensibilité. Nous aimerions faire en sorte que les cellules contenant des données sous renonciation ne soient pas supprimées non plus durant la suppression résiduelle, dans la mesure du possible.

Nous avons rédigé un rapport décrivant les situations où les renonciations permettent la publication d'une cellule dans un tableau de données. L'un des plus grands défis tient au fait que, souvent, lorsqu'on force certaines cellules à ne pas être supprimées, il devient impossible d'obtenir une solution pour le problème de suppression tout entier. Les travaux sont en cours en vue de résoudre ce problème.

Développement de nouveaux prototypes

StatMx est un ensemble de macros SAS qui fournit des fonctions et des capacités en plus de celles disponibles à l'heure actuelle dans le Système généralisé d'échantillonnage (SGECH) ou le Système généralisé d'estimation (SGE). D'autres prototypes sont ajoutés à mesure que la méthodologie arrive à maturité.

Nous avons mis à jour la description des poids de calage et de la méthodologie d'estimation. Nous avons également mis à jour la description de la méthodologie de stratification de la Lavallée-Hidiroglou. Nous avons créé un prototype pour calculer les poids bootstrap, mais celui-ci n'a pas été achevé.

Nouveau système généralisé d'échantillonnage G-Ech

Le système généralisé d'échantillonnage SGECH est arrivé à la fin de son cycle de vie. Le développement de son remplaçant, le G-Ech, a débuté. Ce projet de recherche a été consacré à la rédaction des spécifications de la fonction de répartition de l'échantillon de Bernoulli stratifié à deux phases. Les spécifications ont été livrées aux programmeurs de la Division de l'ingénierie des systèmes dans les délais prévus.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Laurie Reedman (613 951-7301, laurie.reedman@statcan.gc.ca).

Assurance de la qualité

Reconnaissance intelligente de caractères (RIC) générique

La RIC (Reconnaissance intelligente de caractères) est une technologie de saisie des données qui s'appuie sur une combinaison de saisie machine automatisée (en utilisant la reconnaissance optique de caractères, de marques et d'images) et de saisie manuelle avec visualisation « tête haute » des données par les opérateurs. Le logiciel utilisé est appelé « AnyDoc ». La DMEE (Division des méthodes d'enquêtes auprès des entreprises) a développé et mis en œuvre un contrôle de la qualité générique pour ce système de RIC (Reconnaissance intelligente de caractères) pour les étapes de préparation du document, de balayage, de saisie automatique des données (SAD) et de saisie clavier à partir d'images (SCI) du traitement des données.

Une étude par simulation a été conçue et exécutée afin d'utiliser le logiciel non seulement pour déceler la présence de données sur un questionnaire, mais aussi pour saisir ces données. Le résultat inattendu de l'étude par simulation a été de se rendre compte qu'en fait le paramètre du logiciel qui permet à l'utilisateur de déterminer la valeur du seuil de confiance pour la saisie des données avait été désactivé, ce qui a été confirmé lors d'une visite sur place d'un technicien de AnyDoc. Le paramètre a été réactivé et une nouvelle version a été livrée à Statistique Canada. L'étude par simulation

reprendra dans les mois à venir.

Méthodologie d'analyse syntaxique et d'attribution de score pour le codage automatique (G-Code)

G-Code est utilisé pour attribuer automatiquement des codes prédéfinis aux réponses à des questions ouvertes. Cela se fait en deux étapes. À la première étape, le texte en entrée ainsi que le texte de référence sont soumis à une analyse syntaxique selon une stratégie définie par l'utilisateur, afin de réduire le texte à une forme standard. L'analyse syntaxique résout des problèmes tels que les variantes orthographiques fréquentes et les abréviations. La stratégie d'analyse syntaxique joue un rôle important dans la détermination du taux de succès du processus de codage. La deuxième étape consiste à apparier le texte en entrée analysé à une liste de descriptions analysées dans un fichier de référence et à attribuer le code associé quand un appariement a lieu. On peut procéder à l'appariement direct ou indirect. Pour l'appariement indirect, un poids est attribué à chaque mot donnant un appariement dans la phrase en entrée et un score est calculé pour cette phrase d'après les poids et le nombre de mots en commun entre le texte en entrée et les descriptions de référence.

Des spécifications ont été rédigées afin d'inclure l'algorithme de Levenshtein dans G-Code. La programmation et les essais sont presque achevés. Des plans ont été faits en vue de développer un cours pour G-Code. Ce cours s'adressera aux clients qui utilisent le logiciel pour procéder au codage automatisé, ainsi qu'aux méthodologistes qui offrent un soutien à ces clients. D'autres méthodes possibles de codage automatisé, comme celles utilisées dans les correcteurs orthographiques courants et les méthodes fondées sur l'appariement phonétique ont été étudiées. Des travaux ont été entrepris en vue d'étudier les méthodes de prétraitement fondées sur deux algorithmes phonétiques (SOUNDEX et NYSIIS) pour corriger l'orthographe des descriptions afin de les normaliser avant d'essayer de les coder.

Formation statistique en contrôle de la qualité et consultation

La formation en méthodologie de contrôle de la qualité et ses applications est offerte aux méthodologistes ainsi qu'aux non-méthodologistes. Une formation et des conseils personnalisés sur le contrôle et l'assurance de la qualité sont fournis individuellement à des employés de Statistique Canada ainsi qu'à l'externe.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Laurie Reedman (613 951-7301, laurie.reedman@statcan.gc.ca).

Formation statistique

Le Comité de la formation statistique (CFS) coordonne l'élaboration et l'exécution de 25 cours offerts à intervalles réguliers sur les méthodes d'enquête, la théorie et la pratique de l'échantillonnage, la conception de questionnaires, les méthodes applicables aux séries chronologiques et les méthodes statistiques d'analyse des données. Durant l'année, 29 sessions de cours réguliers (80 jours de formation) ont été données, en français ou en anglais, à un total de 309 participants.

La série de cours continue de s'allonger, trois nouveaux cours étant élaborés à l'heure actuelle :

- 0494 : Introduction à l'analyse des données d'enquêtes longitudinales.
- 0460 : Questions de méthodologie concernant les nouvelles méthodes de collecte des données.
- 0492 : Codage automatisé.

Pour obtenir plus de renseignements, veuillez communiquer avec :
François Gagnon (613 951-1463, francois.gagnon@statcan.gc.ca).

Conférences

Symposium 2011

Le 27^{e} Symposium international sur les questions de méthodologie, qui avait pour thème « Stratégies de normalisation des méthodes et des outils - Comment y arriver », a eu lieu du 1er au 4 novembre 2011 au Centre des congrès d'Ottawa. Environ 440 personnes ont participé à l'événement.

Le discours principal a été donné par Susan Linacre et le prix Waksberg a été décerné à Danny Pfeffermann. Trois ateliers ont eu lieu le premier jour du symposium :

- Abdelnasser Saidi et Eric Hortop – « Méthodes de couplage d'enregistrements : théorie et application sous G-Coup ».
- Mike Hidiroglou, Victor Estevao et J.N.K. Rao – « Développements dans l'estimation pour petits domaines : méthodes, applications et développement de logiciel ».
- Natalie Shlomo – « Contrôle de la divulgation statistique : un cadre de travail fondé sur les risques de divulgation et l'utilité des données ».

Le comité organisateur a mis sur pied un programme de plus de 60 communications données par des participants provenant de plus d'une douzaine de pays. Le comité s'est également occupé de la logistique des inscriptions, des opérations et de la gestion des installations. Les actes du symposium devraient être diffusés à la fin de 2012.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Colin Babyak (613 951-2045, colin.babyak@statcan.gc.ca).

Revue Techniques d'enquête

Techniques d'enquête (TE) est une revue internationale dans laquelle sont publiés dans les deux langues officielles des articles portant sur divers aspects des faits statistiques nouveaux pertinents pour un organisme statistique. Le comité de rédaction compte des chefs de file de renommée mondiale du domaine des méthodes d'enquête issus des secteurs public universitaire et privé Il s'agit de l'une des deux seules grandes revues au

monde traitant de la méthodologie liée aux statistiques officielles.

Le numéro de juin 2011 (TF (Techniques d'enquête) 37-1), a été diffusé le 29 juin 2011. Il contient neuf articles.

Le numéro de décembre 2011 (TF (Techniques d'enquête) 37-2), diffusé le 21 décembre 2011, renferme 9 articles, dont le onzième de la série d'articles annuels sollicités en l'honneur de Joseph Waksberg. Le lauréat du Prix Waksberg 2011 est Danny Pfeffermann. Le numéro comprend aussi une section spéciale du U.S. Census Bureau sur les plans de sondage pour les enquêtes démographiques comprenant une introduction, trois articles ordinaires et un document de discussion.

En 2011, la revue a reçu 60 propositions d'articles de divers auteurs.

Au cours de la période de référence, les membres du comité de rédaction et les rédacteurs adjoints ont normalisé, simplifié et décrit le processus de rédaction, y compris les divers contacts avec les auteurs et les rédacteurs associés.

Pour obtenir plus de renseignements, veuillez communiquer avec :
Susie Fortier (613 951-4751, susie.fortier@statcan.gc.ca).

Documents de recherche parrainés par le PRDM (Programme de recherche et développement en méthodologie)

Beaumont, J.-F. (2011). A Generalized Bootstrap Procedure Applied to Finite Population Sampling. Exposé donné à la conférence de Statistique Canada, Montréal, juillet 2011.

Beaumont, J.-F., et Bissonnette, J. (2011). Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI. *Techniques d'enquête*, 37, Nº (numéro). 2, 183-192.

Beaumont, J.-F., et Bissonnette, J. (2012). An overview of SEVANI. Présentation qui sera faite lors de l'atelier sur le Standard Error Estimation and Other Related Sampling Issues in EU-SILC (European Union-Statistics on Income and Living Conditions), Luxembourg, mars 2012.

Beaumont, J.-F., et Bocci, C. (2012). Some Theory on Synthetic Data Generation and its Application to the Canadian Portion of the Cross National Equivalent File. Rapport interne (ébauche).

Beaumont, J.-F., et Charest, A.-S. (2012). Bootstrap variance estimation with survey data when estimating model parameters. *Computational Statistics and Data Analysis* (à paraître).

Beaumont, J.-F., Haziza, D. et Bocci, C. (2011b). A Theoretical Framework for Adaptive Collection Designs. Documentation présentée à l'International Total Survey Error Workshop, ville de Québec, juin 2011.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2011a). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2012). A unified approach to robust estimation in finite population sampling. Article en cours de révision pour publication dans *Biometrika*.

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

Beaumont, J.-F., et Rancourt, E. (2011). Workshop on Edit and Imputation. Cours de deux jours donné au 58e congrès de l'Institut international de statistique qui s'est tenu à Dublin, en Irlande, du 21 au 26 août 2011.

Bélanger, Y., Couture, K. et Neusy, E. (2011). An application of SAS® Simulation Studio: The microsimulation of a computer assisted telephone interviewing system. *Proceedings of the 2011 SAS® Global Forum Conference*, SAS Institute Inc. (incorporated), Cary, Caroline du Nord, État-Unis.

Bemrose, R., Quenneville, B. et Meszaros, P. (2011). Estimating and presenting global and local trends in environmental data. Soumis au *Journal of Agricultural*, *Biological and Environmental Sciences* (JABES).

Bene Tchaleu, F. (sous la supervision de D. Haziza) (2011). Thèse de maîtrise (ébauche), "Record Linkage".

Binder, D. (2011a). Causal Inference for observational data obtained from a complex survey. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada*. (accessible le 26 mars 2012).

Binder, D. (2011b). Estimating model parameters from a complex survey under a model-design randomization framework. *Pakistan Journal of Statistics*, 27, 371-390. (accessible le 26 mars 2012).

Binder, D. (2012). Mode Effects: Why would using different modes for similar surveys give different survey results? Statistique Canada, rapport interne.

Chen, Z.G., et Wu, K.H. (2011). Benchmark forecast and error modeling. Soumis au *Journal of Forecasting*.

Chen, Z.G., Wu, K.H. et Wang, W. (2012). Unit Root Tests and Poly-variogram Methodology. Document de travail qui sera soumis à une publication avec comité de lecture. Sera présenté à l'Australian Statistics Conference, qui se tiendra du 9 au 12 juillet 2012, par le co-auteur Ka Ho Wu, et dans une série de séminaires du CRASC (Centre de recherche et d'analyse en séries chronologiques) par Zhao-Guo Chen.

Choudhry, G.H., Hidiroglou, M.A. et Laflamme, F. (2011). Optimizing CATI (Computer-Assisted Telephone Interview) workload to minimize data collection cost. Communication libre présentée pour publication dans les actes des *JSM (Joint Statistical Meeting) de 2011*.

Choudhry, G.H., Hidiroglou, M.A. et Laflamme, F. (2011). Optimizing CATI (Computer-Assisted Telephone Interview) Workload to Simultaneously Minimize Data Collection Cost for Several Surveys. Exposé donné au Comité consultatif des méthodes de statistiques à l'automne 2011.

Choudhry, G.H., Hidiroglou, M.A., Laflamme, F., Bélanger, Y., Neusy, E. et Couture, K. (2011). Microsimulating and Optimizing CATI (Computer-Assisted Telephone Interview) Call Scheduling. Workshop on microsimulation model NISS (National Institute of Statistical Sciences) (Washington, DC (District of Columbia)). Federal Way (Washington) du 7 au 8 avril 2011.

Choudhry, G.H., Rao, J.N.K. et Hidiroglou, M.A. (2012). À propos de la répartition de l'échantillon pour une estimation sur domaine efficace. Accepté pour publication dans *Techniques d'enquête*.

Dasylva, A. (2011). COS Step 2: Implementation of an EM (espérance-maximisation) algorithm for the estimation of events odds ratios. Statistique Canada, rapport interne, DMES (Division des méthodes d'enquêtes sociales), janvier.

Estevao, V., Hidiroglou, M.A. et You, Y. (2011). Methodology Software Library Small-Area Estimation Fay-Herriot Area Level Model with EBLUP (Empirical Best Linear Unbiased Predictor) Estimation Methodology Specifications. Février 2011. Division de la recherche et de l'innovation en statistique, Statistique Canada.

Estevao, V., Hidiroglou, M.A. et You, Y. (2012). Methodology Software Library Small-Area Estimation Unit Level Model with EBLUP (Empirical Best Linear Unbiased Predictor) and Pseudo EBLUP (Empirical Best Linear Unbiased Predictor) Estimation Methodology Specifications. Février 2012. Division de la recherche et de l'innovation en statistique, Statistique Canada.

Fillion, J.-M. (2011). Renunciation (waivers). Statistique Canada, document interne.

Fortier, S., Julien, J. et Quenneville, B. (2011). How to monitor and Improve Seasonal Adjustment Quality - Statistics Canada's Practical Experience. Présenté à l'International Statistical Institute (ISI) 2011 à Dublin par Susie Fortier.

Fox, K. (2011a). Meta-analysis of Survey Data. Comité consultatif des méthodes statistiques, réunion 53, Document de présentation, Ottawa, Ontario.

Fox, K. (2011b). A Framework for the Meta-Analysis of Survey Data. Thèse de doctorat, Queen's University, Kingston, Ontario.

Fox, K., et Post, S. (2012). Sequential Cell Suppression with Linked Tables, Rapport provisoire.

Gagné, C., Roberts, G. et Keown, L.A. (2012). Weighted Estimation and Bootstrap Variance Estimation for Analysing Survey Data: How to Implement in Selected Software. Paraîtra dans le *Information and Technical Bulletin* in 2012.

Haziza, D., Hidiroglou, M.A. et Rao, J.N.K. (2011). Comparisons of variance estimators in two-phase sampling: An empirical investigation. *Pakistan Journal of Statistics*, 27, 477-492 (Article sollicité pour un numéro spécial en l'honneur de Ken Brewer).

Hidiroglou, M., Estevao, V. et You, Y. (2012). Unit level small area estimation for business survey data. Sera présenté au Symposium on the Analysis of Survey Data and Small Area Estimation du Fields Institute qui aura lieu du 30 mai au 1er juin 2012.

Hidiroglou, M.A. (2011). Current Methodology and Systems Developments in Small Area Estimation at Statistics Canada. Allocution sollicitée donnée le 25 novembre 2011 à l'Université fédérale de Juiz de Fora, au Brésil.

Hidiroglou, M.A. (2011). Methodology Research. Division de la recherche et de l'innovation en statistique. Séance d'information à l'intention du Comité des politiques, 1er octobre 2011.

Hidiroglou, M.A. (2011). On sample allocation for domains. Invited talk given at the Department of Statistics and Actuarial Science, Waterloo, University, 20 octobre 2011.

Hidiroglou, M.A. (2011). On sample allocation for domains. Invited talk given at the Instituto Brasileiro de Geografia e Estatística (IBGE) - Escola Nacional de Ciências Estatísticas, Rio de Janeiro Brazil, 20 novembre 2011.

Hidiroglou, M.A. (2011). Optimizing CATI (Computer-Assisted Telephone Interview) call scheduling. Conférencier invité à l'atelier International Total Survey Error qui s'est tenue à Québec: du 20 au 22 juin 2011.

Hidiroglou, M.A. (2011). Progrès récents en matière d'estimation sur petits domaines à Statistique Canada. Recueil des présentations faites à Tanger, Maroc, Pratiques et Méthodes de Sondage. Éditeurs, Mari-Ève Tremblay, Pierre Lavallée, et Mohamed El haj Tirari. 349-357. (Allocution de clôture sollicitée).

Hidiroglou, M.A., Estevao, V., You, Y. et Rao, J.N.K. (2011). Small Area Estimation Methods, Applications and Practical Demonstration. Cours d'une journée donné au Symposium 2011, à Ottawa, Canada, le 1er novembre 2011.

Hidiroglou, M.A., et You, Y. (2012). Calibrating to incomplete constraints. Article préliminaire en cours de rédaction.

Hidiroglou, M.A., et Yung, W. (2011). Workshop on Business Survey Methods. Cours de deux jours donné au 58e congrès de l'Institut international de statistique qui s'est tenu à Dublin, en Irlande, du 21 au 26 août 2011.

Hortop, E., Saïdi, A. et Reicker, A. (2011). One-to-one Probabilistic External Linkage Assisted by a Deterministic Pre-linkage Outside of G-Link. Statistique Canada, rapport interne.

Jones, J., et Hidiroglou, M.A. (2012). Capturing, Coding, and Cleaning the Data. Chapter XX (vingt): à paraître dans un livre publié par Wiley intitulé" How to design and conduct Business Surveys"; Éditeurs, Gustav Haraldsen (Statistics Norway), Jacqui Jones (United Kingdom (UK) Office of National

Statistics), Ger Snijkers (Statistics Netherlands), et Diane Willimack (United States (US) Bureau of Census).

Julien, C., et Fortier, S. (2011). Seasonal Adjustment Horizontal Review. Statistique Canada, rapport interne (Secrétariat de la qualité).

Laflamme, F., et St-Jean, H. (2011). Highlights and Lessons from the First Two Pilots of Responsive Collection Design for CATI (Computer-Assisted Telephone Interview) Surveys. Sera publié dans les actes de 2011 des Joint Statistical Meeting.

Laflamme, F., et St-Jean, H. (2011). Proposed Indicators to Assess Interviewer Performance in CATI (Computer-Assisted Telephone Interview) Surveys. Pour être publié dans le 2011 Proceedings of the Joint Statistical Meeting.

Lavallée, P., et Labelle-Blanchet, S. (2011). Indirect sampling applied to skewed populations. *Recueils de la section des méthodes d'enquête de l'assemblée annuelle de la Société Statistique du Canada*, juin 2011.

Lavallée, P., et Labelle-Blanchet, S. (2012). Le sondage indirect appliqué aux populations asymétriques. Journées de Méthodologie Statistique, Institut National de la Statistique et des Études Économiques (INSEE), Paris, janvier.

Mach, L., et Moussa, S. (2012). Evaluation of the Synthetic Longitudinal Data Created for the Canadian Portion of the Cross National Equivalent File. Statistique Canada, rapport interne.

Mach, L., Rubin-Bleuer, S. et Schiopu-Kratina, I. (2012). Coordination of samples in business surveys. Diapositives de l'exposé donné au comité technique de la Division des méthodes d'enquêtes auprès des entreprises (DMEE).

Mantel, H., et Hidiroglou, M.A. (2011). Non-response Follow-up Allocation for Domains. Communication sollicitée présentée pour publication dans les actes des *JSM (Joint Statistical Meeting)* 2011.

Mazzi, G.L., Fortier, S. et Quenneville, B. (2011). Multivariate Benchmarking - The Direct Versus Indirect Problem. Soumis pour publication éventuelle dans le Handbook on Seasonal Adjustment d'Eurostat.

Mode Effect Working Group (2012). Mode Effects: The Medium and the Messenger. Statistique Canada, document interne.

Pelletier, C., Yeung, A., et Akpoué, B.P. (2011). Optimization rules for the Levenshtein's algorithm. Statistique Canada, document interne.

Quenneville, B., et Findley, D. (2011). Gain and phase shift for the annual difference, the monthly difference and the annual sum operators. *Taiwan Economic Forecast and Policy*. À paraître.

Quenneville, B., et Fortier, S. (2011). Restoring accounting constraints in time series – Methods and software for a statistical agency. Economic Time Series: Modeling and Seasonality, Chapman & Hall/CRC.

Quenneville, B., et Fortier, S. (2011b). Benchmarking and temporal consistency. Soumis pour publication éventuelle dans le *Eurostat Handbookon Seasonal Adjustment*.

Quenneville, B., et Gagné, C. (2011). Testing time series data compatibility for benchmarking. *International Journal of Forecasting*. À paraître.

Quenneville, B., Picard, F. et Fortier, S. (2012). Calendarisation with interpolating splines and state space models. À soumettre pour publication dans Journal of the Royal Statistical Society, Séries C (*JRSS-C).*

Rivière, P., et Rubin-Bleuer, S. (2011). Random permutation method of sampling coordination. Rapport sur l'avancement des travaux SRID (Statistical Research and Innovation Division)-2011-001E.

Roberts, G. (2012). Analyzing Census microdata in an RDC (Research Data Centre): What weight to use? À paraître dans le numéro du printemps 2011 du Bulletin technique et de l'information (12-002-X, vol. (volume). 5).

Roberts, G., et Binder, D. (2012). Satterthwaite Adjustments and Degrees of Freedom when Analysing Survey Data: What could this mean for the CHMS (Canadian Health Measures Survey)? Statistique Canada, document interne.

Rubin-Bleuer, S. (2011). The proportional hazards model for survey data from independent and clustered super-populations. *Journal of Multivariate Analysis*, 102, Issue 5, mai 2011, 884-895.

Rubin-Bleuer, S., Godbout, S. et Hidiroglou, M.A., (2012). Extensions of the Pseudo-EBLUP (Empirical Best Linear Unbiased Predictor) estimator with application to the Survey of Employment, Payrolls and Hours (SEPH). Article préliminaire en cours de rédaction.

Rubin-Bleuer, S., Jamroz, E. et Rao, J.N.K., (2012). Variance estimation for a small area model with penalized splines. Manuscrit interne en cours de documentation.

Rubin-Bleuer, S., et Yong, Y. (2012). ADM variance estimation for the Fay-Herriot model. Ébauche en cours.

Rubin-Bleuer, S., Yung, W. et Landry, S., (2011). Adjusted maximum likelihood method for time series and cross-sectional small area models. Diapositives de l'exposé donné à la conférence SAE 2011, compte rendu et article préliminaire en cours de rédaction.

Schiopu-Kratina, I., Fillion, J.-M., Mach, L. et Reiss, P.T., (2012). Maximizing the conditional overlap in business surveys. Document technique en cours de révision.

Schiopu-Kratina, I., et Zamfirescu, C.M.D. (2011). A structural approach to questionnaire design. Communication libre donnée à la réunion des Joint Statistical Meetings de l'American Statistical Association (ASA), Miami Beach, 2011.

Simard, M. (2011a). Real Time Remote Access at Statistics Canada: Development Challenges and Issues. Communication présentée au 58^{e} congrès de l'Institut international de statistique qui s'est tenu à Dublin, en Irlande, du 21 au 26 août 2011.

Simard, M. (2011b). Progress with Real Time Remote Access. Documentation présentée à la réunion de travail de la Commission économique des Nations Unies pour l'Europe (CEE-ONU)/Eurostat sur la confidentialité des données statistiques qui a eu lieu à Tarragona, en Espagne, du 26 au 28 octobre 2011.

Tambay, J.L., et Fillion, J.-M. (2011). New business survey confidentiality software G-Confid. Document présenté à la Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality in Tarragona, Espagne, du 26 au 28 octobre 2011.

Verret, F., Hidiroglou, M.A. et Rao, J.N.K. (2012). Model-based small area estimation under informative sampling. Soumis à *Techniques d'enquête* pour publication éventuelle.

Wright, P., et Beaumont, J.-F. (2012). Bootstrap Variance Estimation using a Poisson Sampling Approach for a Without Replacement Design. Statistique Canada, rapport interne.

Wu, M., et Ferland, M. (2011). Forecasts and Seasonally Adjusted Series with Forced Annual Totals: An Empirical Study. Document de travail en cours de rédaction.

Xi, D.Z. (Dexen) (2011). Methodology Research and G-Confid Development: Research on the Shuttle Algorithm in the macro AUDIT of G-Confid. Statistique Canada, document interne.

Xie, H. (2011). Discussion: Dynamic Allocation of Follow-Up Units in Data Collection for Business Surveys. Communication interne de Statistique Canada.

Xie, H. (2012). Sub-Sampling Methods for Selecting Follow-Up Units in Data Collection for Business Surveys. Communication présentée dans le cadre de la série de séminaires de la Division des méthodes d'enquête auprès des entreprises de Statistique Canada.

You, Y. (2012). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models. Manuscript interne de la Division de la recherche et de l'innovation en statistique (DRIS).

You, Y., et Hidiroglou, M.A. (2012). Confidence interval comparison of small area estimators using unit level and area level models under model mis-specification. Article préliminaire en cours de rédaction.

You, Y., et Hidiroglou, M.A. (2012). Pseudo Hierarchical Bayes approaches to small area estimation based on general nested error regression models. Statistique Canada, Manuscript interne de la Division de la recherche et de l'innovation en statistique (DRIS). (En cours de révision et de documentation).

You, Y., et Hidiroglou, M.A. (2012). Sampling variance smoothing methods for small area proportion estimators with applications. Manuscrit interne de la Division de la recherche et de l'innovation en statistique (DRIS), qui sera présenté au Symposium on the Analysis of Survey Data and Small Area Estimation du Fields Institute qui aura lieu du 30 mai au 1er juin 2012.

You, Y., Rao, J.N.K. et Hidiroglou, M.A. (2011). Benchmarking small area estimators under the Fay-Herriot model and MSPE (Mean Squared Prediction Error) estimation. Document de travail de la Direction de la méthodologie, SRID (Statistical Research and Innovation Division)-2011-004E, Statistique Canada, Ottawa, Canada. (Le document a également été soumis à une revue et est en cours de révision pour publication).

You, Y., et Zhou, M.Q. (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, Nº (numéro). 1, 31-44.

Yung, W., Rubin-Bleuer, S. et Landry, S. (2012). Variance estimation for a time series and cross-sectional model. Exposé qui sera donné au Symposium on the Analysis of Survey Data and Small Area Estimation du Fields Institute qui aura lieu du 30 mai au 1er juin 2012.