Stratégies de normalisation des méthodes et des outils - Comment y arriver

Recueil

Symposium international de Statistique Canada sur les questions de méthodologie

1 au 4 novembre 2011

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :	
Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

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

Poste :
Statistique Canada
100, promenade du Pré Tunney
Ottawa (Ontario) K1A 0T6

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. Le normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

Symposium 2011 - Catalogue no. 11-522-XCB

Pour obtenir d'autres renseignements
Service d'octroi de licences
Division des services à la clientèle, Statistique Canada
Immeuble R.-H. Coats, 9e étage, section A
Ottawa (Ontario) K1A 0T6
Courriel : licences@statcan.gc.ca
Téléphone : 613-951-1122
Télécopieur : 613-951-1134

STRATÉGIES DE NORMALISATION DES MÉTHODES ET DES OUTILS – COMMENT Y ARRIVER

TABLE DES MATIÈRES

PRÉFACE..... 6

DISCOURS PRINCIPAL

01A-1 Normalisation des méthodes et des outils : un examen rétrospectif de plus de 30 années d'expérience..... 8
S. Linacre, Australian Bureau of Statistics, Australie

SÉANCE 2A : ARCHITECTURE OPÉRATIONNELLE DU BUREAU : BASES DE SONDAGE DES ENQUÊTES-MÉNAGES

02A-1 Remaniement à Statistics Netherland 25
F. Hofman, Statistics Netherlands, Pays-Bas

02A-2 Vers la normalisation à Statistics New Zealand 34
J. Lopdell et G. Dunnet, Statistics New Zealand, Nouvelle-Zélande

02A-3 Développement d'une base de sondage commune pour les enquêtes-ménages à Statistique Canada 35
L. MacNabb, M. St-Pierre et M. Grenier, Statistique Canada

SÉANCE 02B : ÉCHANTILLONNAGE ET ESTIMATION

02B-1 Effets de plan différentiels dans les échantillons d'enquêtes en milieu scolaire 44
C. Dahmen et M. Fuchs, Darmstadt University of Technology, Allemagne

02B-2 Enquêtes sur l'effectif des Forces canadiennes : Examen de la pondération et de la non réponse 52
F. Larochelle, T. Gou et I. Goldenberg, Ministère de la defense nationale et les Forces canadiennes, Canada

02B-3 Élaboration d'un cadre d'échantillonnage intégré pour les enquêtes auprès des entreprises : Études de simulation pour évaluer les gains d'efficacité liés à un plan d'échantillonnage à deux phases.....59
Y. Li et F. Picard, Statistique Canada

02B-4 Estimation de la variance par réplication peu dense et efficace pour les enquêtes complexes 66
J. K. Kim, Iowa State University, É.-U. et C. Wu, University of Waterloo, Canada

SÉANCE 03A : HARMONISATION DES MÉTHODES DANS LE CADRE DE PROJETS DE NORMALISATION À GRANDE ÉCHELLE POUR LES ENQUÊTES AUPRÈS DES ENTREPRISES

03A-1 Harmonisation de méthodologies dans le contexte d'un projet d'intégration de Systèmes : défis et leçons apprises 68
J. Andrews, F. Brisebois, I. Delahousse, C. Dochitoiu, M. Lachance, R. Philips et S. Pursey, Statistique Canada

03A-2 Les multiples facettes de la refonte des applications permettant de produire les statistiques conjoncturelles de l'INSEE : le programme Premice 75
F. Guggemos, Institut national de la statistique et des études économiques, France

03A-3 Normalisation des enquêtes-entreprises infra-annuelles au Royaume Uni.....82
S. Merad et P. Brodie, Office for National Statistics, Royaume-Uni

SÉANCE 03B : DIFFUSION ET ACCÈS AUX DONNÉES

03B-1 SICON d'Eurostat : Infrastructure de sécurité pour l'accès aux données confidentielles et le partage de ces données.....91
D. Buono, Eurostat, Luxembourg

03B-2 Aperçu et description technique du système amélioré de production de tableaux : Normalisation de la production de tableaux personnalisés pour accroître la qualité des données et l'efficacité 92
P. Timusk, M. Mansour, É. Pelletier et E. Turgeon, Statistique Canada

SÉANCE 04A : VÉRIFICATION SÉLECTIVE

04A-1 La vérification sélective des données et sa mise en œuvre à Statistics Sweden 101
P. Brundell, Statistics Sweden, Suède

04A-2 SeleMix : Un progiciel R pour la vérification sélective au moyen de modèles de contamination 108
M. Di Zio et U. Guarnera, Istat, Italie

04A-3 Méthodes et outils de vérification sélective – Perspective de l'Australian Bureau of Statistics 116
E. Brinkley, K. Farwell et F. Yu, Australian Bureau of Statistics, Australie

SÉANCE 04B : CONFIDENTIALITÉ

04B-1 G-Confid : le logiciel de confidentialité de Statistique Canada.....126
C. Rondeau et J.-M. Fillion, Statistique Canada

04B-2 Évaluation du risque de divulgation dans le cas de microdonnées perturbées.....132
N Shlomo, University of Southampton, Royaume-Uni

04B-3 Le couplage d'enregistrements probabiliste préservant la confidentialité de A à Z : un exemple d'utilisation du Système généralisé de couplage d'enregistrements à SwissLinkage..... 141
Spoerri, K. Schmidlin, University of Bern, Suisse, R. Schnell, University Duisburg-Essen, Allemagne et K. Clough-Gorr, University of Bern and National Institute of Cancer Epidemiology and Registration, Suisse

04B-4 Produits normalisés et classifications de niveau régional pour les populations minoritaires dans le cadre du Recensement de 2011 de l'Angleterre et du pays de Galles 142
J. Traynor et E. White, Office for National Statistics, Angleterre

SÉANCE 05A : DISCOURS DU GAGNANT DU PRIX WAKSBERG

05A-1 Modélisation des données d'enquêtes complexes : Pourquoi les modéliser? Pourquoi est-ce un problème? Comment pouvons-nous le résoudre? 144
D. Pfeffermann, Hebrew University of Jerusalem, Israël, University of Southampton, Royaume-Uni

SÉANCE 06A : NORMES ET LIGNES DIRECTRICES POUR LA CONCEPTION ET LA MISE À L'ESSAI DE QUESTIONNAIRES INTERNET

06A-1 Lignes directrices pour l'élaboration de questionnaires électroniques : problèmes et défis dans un environnement en évolution 146
A.-M. Côté, D. Lawrence et P. Kelly, Statistique Canada

06A-2 GINO++, un système généralisé pour les enquêtes en ligne 152
R. Torelli, National Institute of Statistics (ISTAT), Italie

06A-3 Expérience intégrée sur des méthodes de suivi des cas de non-réponse visant la collecte de données au moyen d'un questionnaire électronique 161
M. Karaganis, K. Fox, J. Claveau, J. Leung et W. Lin, Statistique Canada

SÉANCE 06B : DONNÉES ABERRANTES ET IMPUTATION

06B-1 Intégration d'une procédure simplifiée de détection des valeurs aberrantes dans un système généralisé complexe 172
L.T Bechtel, U.S. Census Bureau, É.-U.

06B-2 Outil de détection de valeurs aberrantes à Statistique Canada.....179
N. Émond, Statistique Canada

06B-3 Évaluation de méthodes d'imputation de l'exposition au risque dans le profil de risque des acteurs d'un modèle pour la microsimulation.....186
D. Hennessy, Institut de recherche de l'Hôpital d'Ottawa et Statistique Canada, C. Bennett, M. Tuna, Institut de recherche de l'Hôpital d'Ottawa, C. Nadeau, W. Flanagan, Statistique Canada et D. Manuel, Institut de recherche de l'Hôpital d'Ottawa

SÉANCE 07A : MÉTHODES DE PROTECTION DE LA CONFIDENTIALITÉ ET OUTILS POUR L'ACCÈS AUX DONNÉES

07A-1 Méthodes de masquage de l'identité pour les fichiers de données sur la santé à grande diffusion 195
K.E. Emam, Université d'Ottawa, Canada

07A-2 Le système d'analyse des microdonnées du U.S. Census Bureau 196
M. Freiman, U.S. Census Bureau, J. Lucero, Freddie Mac, L. Singh, Georgetown University, J. You, University of California, M. DePersio, University of Delaware et L. Zayatz, U.S. Census Bureau, É.-U.

07A-3 Accorder l'accès aux microdonnées à des fins statistiques – Expérience de l'Australian Bureau of Statistics concernant les serveurs d'analyse à distance 206
J.O. Chipperfield, F. Yu et M. Gare, Australian Bureau of Statistics, Australie

SÉANCE 07B : CONTENUE ET COLLECTE

07B-1 Certaines conséquences de la normalisation des méthodes de surveillance de la qualité des interviews d'enquête 216
D. Currivan, D. Stone, K. Fuller, S. Kinsey et H. Speizer, RTI International, É.-U.

07B-2 Enquête santé européenne par examen : la perspective de l'échantillonnage et du recrutement 224
J. Heldal , S. Jentoft, Statistics Norway, Norvège, K. Kuulasmaa, P. Koponen et S. Ahonen, National Institute for Health and Welfare, Finlande

07B-3 Planification préliminaire de la collecte : Guichet de la collecte..... 232
A. Marcil, Statistique Canada

07B-4 Mise en œuvre de procédures de contrôle de la qualité au centre national des opérations du NASS..... 237
J.M. Boone, J.L. Parsons, S.R. Feld, J.N. Levy et K.L. Flaherty, USDA National Agricultural Statistics Service, É.-U.

SÉANCE 08A : UTILISATION DE MÉTHODES ET D'OUTILS NORMALISÉS POUR LE TRAITEMENT POST-COLLECTE

08A-1 Normalisation du traitement des données après la collecte dans les enquêtes-entreprises à Statistique Canada 245
S. Godbout, Statistique Canada

08A-2 Normalisation des processus 254
F. Hofman, A. Camstra et R.Renssen, Statistics Netherlands, Pays-Bas

08A-3 Codage des réponses dans les enquêtes – Efforts d'assurance de la qualité et outils des TI à Statistics Sweden 263
J. Svensson, Statistics Sweden, Suède

SÉANCE 08B : QUESTIONNAIRES ET EFFETS DU MODE DE COLLECTE

08B-1 Dispositions relatives à la qualité des données d'enquête dans la solution de questionnaire électronique de Statistique Canada : Rétrospective et perspectives 268
Y. Abiza, Statistique Canada

08B-2 Conception d'un questionnaire pour examiner le Programme de sports des Forces canadiennes 269
K.K. Hachey, Department of National Defence and the Canadian Military, Canada

08B-3 Contenu harmonisé : Le nouveau paradigme d'élaboration des enquêtes à Statistique Canada 274
R. Nadwodny et P. Best, Statistique Canada

08B-4 Effets du mode d'étalonnage sur l'enquête hollandaise sur la criminalité 280
B. Buelens et J. van den Brakel, Statistics Netherlands, Pays-Bas

SÉANCE 09A : NORMALISATION DANS LE CADRE D'ÉTUDES COMPARATIVES INTERNATIONALES : AVANTAGES ET DÉFIS

09A-1 Conception, normalisation et suivi des opérations d'enquête dans les études internationales à grande échelle en éducation 282
R. Carstens, IEA Data Processing and Research Center, Allemagne

09A-2 Résumé des réponses d'item dans l'évaluation à grande échelle 293
E. Gonzalez et M. Von Davier, Educational Testing Service, É.-U.

09A-3 Standardisation des plans de sondage et assurance de qualité dans les enquêtes comparatives 294
M. Joncas et S. LaRoche, Statistique Canada

SÉANCE 09B : LOGICIELS NORMALISÉS

09B-1 Harmonisation des pratiques de désaisonnalisation grâce à l'élaboration du logiciel DEMETRA+ 302
J. Palate, National Bank of Belgium et P. Jacques, Eurostat, Luxembourg

09B-2 Comment fonctionne le SCANCIR et peut-il être utile à un plus grand nombre d'utilisateurs? 303
C. W. Liu, S. Crowe et A. Alavi, Statistique Canada

09B-3 Développement de l'environnement de traitement des enquêtes sociales 310
L. MacNabb, Statistique Canada

09B-4 Mise en œuvre de changements ou d'améliorations méthodologiques dans un système de traitement normalisé, ou comment un groupe consultatif peut-il faciliter le changement? 316
K.J. Thompson, U.S. Census Bureau, É.-U.

SÉANCE 10A : CADRES DE TRAVAIL

10A-1 Élaboration d'une méthodologie pour le Cadre canadien pour les statistiques culturelles 325
M.K. Allen, Statistique Canada

10A-2 Combien de Canadiens vivent dans une ville? Conceptualisation, définition et diffusion proposée de normes de rechange 326
R.D. Bollman et P. Murphy, Statistique Canada

10A-3 Rôle des normes de qualité des données dans l'uniformisation des méthodes et des outils d'enquête .334
J.L. Eltinge, Bureau of Labor Statistics, É.-U.

10A-4 Travaux liés à l'architecture intégrée à Statistics Sweden.....335
M. Axelson, J. Engdahl, Y. Fossan, E. Holm, I. Jansson, B. Lorenc et L.G. Lundell, Statistics Sweden, Suède

SÉANCE 10B : EFFETS CALENDRIER ET COHÉRENCE TEMPORELLE

10B-1 Étalonnage et prévision : une approche descendante pour combiner les prévisions faites à plusieurs fréquences 343
M.A. Trovero, E. Blair et M.J. Leonard, SAS Institute, É.-U.

10B-2 Amélioration de la calendarisation en utilisant X-12-ARIMA : application aux données sur la TPS 352
R. Manríquez, Statistique Canada

10B-3 L'erreur dans les estimations du cycle économique obtenues d'après des données désaisonnalisées ...360
T. McElroy, U.S. Census Bureau, É.-U.

SÉANCE 11A : CONCEPTION ET UTILISATION DE SYSTÈMES GÉNÉRALISÉS

11A-1 Systèmes généralisés : l'expérience de Statistique Canada 362
Y. Deguire, L. Reedman et M. Wenzowski, Statistique Canada

11A-2 Triton : un outil général de collecte et de microvérification des données 370
J. Erikson, Statistics Sweden, Suède

11A-3 La boîte à outils méthodologique standard de Statistics New Zealand 377
J. Lopdell et G. Dunnet, Statistics New Zealand, Nouvelle-Zélande

SÉANCE 11B : MODÉLISATION ET ESTIMATION

11B-1 Estimation semi-paramétrique fondée sur un modèle des composantes de séries chronologiques et l'erreur quadratique moyenne des estimateurs 379
M. Sverchkov, R. Tiller, Bureau of Labor Statistics, É.-U. et D. Pfeffermann, Hebrew University of Jerusalem, Israël and University of Southampton, Royaume-Uni

11B-2 Défis et enjeux de la pondération de l'Enquête sur les voyages des résidents du Canada.....380
F. Labrecque-Synnott, Statistique Canada

11B-3 Méthodes statistiques d'évaluation des tendances séculaires en utilisant des estimations provenant d'enquêtes annuelles à échantillons probabilistes complexes transversaux indépendants 385
P.J. Smith et Z. Zhao, Centers for Disease Control and Prevention, É.-U.

Préface

Le Symposium 2011 était le 27e Symposium international sur les questions de méthodologie de Statistique Canada. Chaque année, le Symposium se penche sur un thème précis. En 2011, le thème était : **« Stratégies de normalisation des méthodes et des outils – Comment y arriver ».**

Le Symposium 2011 a eu lieu du 1er au 4 novembre 2011 au Centre des congrès d'Ottawa, Ontario ; plus de 400 personnes provenant de différents pays y ont assisté. En tout, trois ateliers de travail et 61 communications ont été présentés. À l'exclusion de la traduction et de la mise en page, les communications offertes par les auteurs ont été reproduites dans ce compte rendu.

Les organisateurs du Symposium 2011 souhaitent souligner la contribution des différentes personnes — trop nombreuses pour être nommées individuellement — qui ont permis d'assurer le succès de cet événement. Les organisateurs souhaitent également remercier les présentateurs et les auteurs d'avoir pris le soin de transcrire le contenu de leurs présentations en forme écrite. Enfin, les organisateurs remercient tous les participants qui ont suivi les différentes présentations.

Le comité organisateur du Symposium 2011

Colin Babyak, président

Logistique et opérations
Lyne Guertin, présidente
Jack Singleton
Lori Stratychuk

Programme
Sarah Franklin, présidente
José Gaudet
Richard Laroche

DISCOURS PRINCIPAL

Normalisation des méthodes et des outils : un examen rétrospectif de plus de 30 années d'expérience

Susan Linacre[1]

Résumé

Les organismes statistiques nationaux évoluent dans un secteur qui a continuellement pris de l'expansion au cours des dernières décennies. La demande d'information continue d'augmenter dans un marché de plus en plus concurrentiel et soucieux des coûts. Parallèlement, les progrès technologiques se font rapidement et ouvrent un éventail permanent de nouvelles possibilités de révolutionner la façon dont nous travaillons.

L'une d'elles consiste à réaliser des gains d'efficience et de qualité, à en faire davantage et à mener à bien des projets plus complexes avec moins, grâce à des méthodes et des outils uniformisés. La recherche d'uniformisation est en cours depuis que je travaille en statistique et, du fait des innovations technologiques continues et de l'amélioration connexe des ressources d'information, on peut s'attendre à ce que de nouveaux cycles d'uniformisation se poursuivent à l'avenir. Mais qu'apprend-on de chaque cycle sur la façon de l'améliorer la prochaine fois?

Cette présentation expose ma perception de la raison pour laquelle nous travaillons aussi fort pour uniformiser nos méthodes et nos outils, et montre quand et pourquoi la recherche s'est révélée relativement réussie ou moins.

À cette fin, je me fonde sur mon expérience au Australian Bureau of Statistics, en commençant par un système généralisé élaboré pour les enquêtes auprès des entreprises au début des années 1970 et en passant par plusieurs cycles subséquents d'uniformisation et de modernisation. Je m'appuie aussi sur mon expérience personnelle au sein de l'Office for National Statistics du Royaume-Uni.

Je ne suis pas une experte du développement de systèmes. La présentation comprend mon point de vue sur ce qui fonctionne et ne fonctionne pas, tant comme méthodologiste aguerrie, qui tente d'élaborer des méthodes uniformisées et de mettre en œuvre des méthodes et des outils uniformisés, que comme statisticienne spécialisée, qui tente de tirer parti des nombreux avantages d'une approche uniformisée.

1. Introduction

L'infrastructure statistique est au cœur du fonctionnement de tout organisme statistique officiel. Cette infrastructure comprend des composantes conceptuelles réutilisables, telles que des classifications et des normes pour définir les items, des modules de questions, ainsi que des modules de vérification et des ensembles dérivés connexes, de même que des méthodes d'établissement du plan de sondage et d'estimation. Elle englobe aussi les divers outils et systèmes utilisés pour soutenir ces composantes conceptuelles, y compris les systèmes de métadonnées pour les décrire. Cela comporte, entre autres, les codeurs, les modules de sélection et d'estimation, et les outils de vérification.

Au cours de ma carrière, j'ai travaillé en qualité de méthodologiste pendant plus de trois décennies dans deux pays (quoique, dans l'un d'eux pendant une ou deux années seulement) et en qualité de statisticienne spécialisée principale pendant six autres années. Le présent article est un exposé de certaines opinions personnelles fondées sur cette expérience. Y sont présentées mes réflexions sur ce qu'il importe selon moi de prendre en considération pour normaliser l'infrastructure destinée à être utilisée dans l'ensemble d'un organisme statistique.

En premier lieu, je décris brièvement certains aspects de l'édification de l'infrastructure statistique de l'Australian Bureau of Statistics (ABS) qui s'est étalée sur plusieurs décennies. Des contraintes d'espace m'obligeant d'être sélective, j'ai choisi de me concentrer sur la statistique des entreprises, qui est le domaine dans lequel je possède la plus longue expérience méthodologique. Je me concentre aussi sur l'ABS, car mon séjour à l'Office for National Statistics (ONS) n'a pas été suffisamment long pour me donner une perspective appropriée des problèmes et des

[1]Susan Linacre, Australian Bureau of Statistics, susan.linacre@abs.gov.au.

résultats. C'est de ce genre d'antécédents que nous devons tirer les leçons avant de nous embarquer dans de nouveaux projets de normalisation en vue de rendre notre production statistique plus efficace et efficiente.

Après cet historique, je fais part de mes réflexions sur les raisons qui nous poussent à normaliser les méthodes et les outils, ainsi que sur les circonstances dans lesquelles nous ne souhaitons pas le faire. Je discute également de certaines décisions importantes qu'il convient de prendre dans ce contexte. Je conclus par certaines remarques, fondées sur mon expérience passée, quant aux approches qui semblent donner de bons résultats lorsque l'on entreprend d'importants programmes de normalisation, et à certaines embûches qu'il convient d'éviter.

Toutefois, avant tout, je pense qu'il serait utile de réfléchir à quoi pourrait ressembler un organisme statistique qui fonctionne bien.

Un tel organisme aurait, entre autres, les attributs suivants :

- la production fiable d'une trame solide de statistiques robustes, d'une grande intégrité, découlant d'un ensemble de processus bien conçus, bien documentés et bien soutenus ;
- des ressources et une capacité de haut niveau axées sur l'élaboration de nouveaux ensembles de statistiques à intégrer dans cette trame ou à utiliser avec celle-ci, ou sur le renforcement et la modification des composantes de cette trame de manière contrôlée ;
- la capacité de répondre avec souplesse et efficacité aux nouvelles demandes ponctuelles, en recueillant de nouvelles données ou en combinant des données provenant de sources disponibles.

Pour arriver aux résultats susmentionnés, un organisme doit atteindre un très haut niveau d'excellence opérationnelle dans son travail de base, afin d'avoir la capacité d'innover, d'améliorer sa production et de la maintenir pertinente. Pareille excellence opérationnelle aurait l'avantage supplémentaire de procurer aux employés un vif sentiment de satisfaction à l'égard de leur travail et de leur permettre d'utiliser leur créativité pour développer et créer des produits intéressants au lieu de recouvrer des données provenant de dérivations mal spécifiées et de fichiers corrompus.

La principale justification d'une normalisation appropriée est qu'il s'agit d'un déterminant essentiel de l'excellence opérationnelle, donc de la rigueur, de la souplesse et de la pertinence que l'on attend des organismes statistiques nationaux. Elle permet d'atteindre l'excellence opérationnelle parce qu'elle soutient les processus répétitifs au moyen de méthodes et d'outils bien conçus et efficaces. Bien réalisée, elle offre une approche peu coûteuse et de haute qualité à de multiples utilisateurs.

2. Édification de l'infrastructure statistique à l'ABS au fil des décennies

À la fin des années 1960 et au début des années 1970, l'ABS a procédé à l'établissement d'une approche unique, fondée sur une base de sondage aréolaire, pour un programme d'enquêtes auprès des ménages. Simultanément, le Bureau a reconnu que des approches communes pourraient aussi s'appliquer aux enquêtes auprès des entreprises et a donc développé une approche commune appuyée par un système d'enquête généralisé, appelé General Survey System (GSS). Le but était de procurer un système complet d'exécution des enquêtes-entreprises réalisées à l'ABS, allant de l'élaboration du plan de sondage à l'estimation (y compris les erreurs-types) et la production de tableaux, en passant par la sélection de l'échantillon, l'acheminement, la collecte des données et leur vérification. Les enquêtes utilisant le GSS étaient appelées collectivement enquêtes-entreprises. Elles comprenaient trois enquêtes trimestrielles, sur les stocks, les dépenses en immobilisations et les bénéfices. De nombreux autres programmes d'enquêtes-entreprises de l'ABS ont choisi de ne pas utiliser le GSS.

Le concept du GSS était puissant et constituait une excellente plateforme d'apprentissage, mais le système n'offrait pas la souplesse que les programmes d'enquête jugeaient nécessaire pour répondre à leurs divers besoins concernant l'estimation ou le traitement des nouvelles entreprises et des entreprises disparues, *etc*. Lorsqu'un programme choisissait d'utiliser le GSS, il entrait dans un tunnel et n'avait accès à aucune autre stratégie jusqu'à l'obtention des tableaux à la sortie. L'expérience a montré qu'il fallait offrir la capacité de s'adapter à de nouvelles méthodes et approches, ainsi qu'à de nouvelles sources de données auxiliaires susceptibles de devenir efficaces pour les enquêtes au fil du temps. Elle a également montré qu'il était nécessaire d'adopter des approches paramétrables pour offrir la souplesse voulue.

Au cours des années 1980, l'ABS a dû transférer toutes ses applications sur une nouvelle plateforme et, dans le cadre de cette migration, a développé une deuxième génération d'outils d'enquête généralisés, appelés fonctions d'enquête (*survey facilities*), qui consistaient en une série de modules. Les outils de conception et d'estimation ont été étendus afin d'offrir une grande gamme d'options et, forts des leçons tirées de la rigidité du GSS, nous avons décomposé la fonctionnalité des nouveaux outils en très petits éléments pouvant être assemblés pour construire une application. L'une des préoccupations étant qu'on ne pouvait prédire quelle technique d'estimation pourrait effectivement être utilisée cinq à dix ans plus tard, l'approche de conception était de veiller à ce que, quelles que soient les techniques appliquées, ces petites composantes (par exemple, sommes des carrés et produits croisés) puissent être utilisées comme éléments de base.

Cependant, nous étions allés trop loin. Chaque composante résolvait une tellement petite partie du problème global que l'assemblage d'une application demandait beaucoup de travail, surtout si l'interfaçage était difficile. Des spécialistes étaient nécessaires pour construire des systèmes qui fonctionnaient. Nous étions bien loin de la souplesse de type « brancher et utiliser » que nous souhaitions offrir aux utilisateurs. Un élément positif a toutefois été que, dans le cadre du grand programme de transfert du travail au nouvel environnement informatique, l'élaboration des fonctions d'enquête a été gérée de manière très serrée, avec une participation importante de la haute direction. Des normes ont été établies et utilisées par tous les programmes d'enquêtes pertinents, et la documentation sur les fonctions d'enquête était excellente. Cette dernière caractéristique signifiait qu'il serait possible d'adapter les fonctions à de nouveaux environnements bien au-delà de que ce qui aurait été considéré normalement comme une date limite d'utilisation. De surcroît, l'utilisation générale des fonctions dans toutes les applications rendait le passage à de nouvelles fonctions plus simples que cela ne l'aurait été autrement.

Au cours des années 1980, à part les trois enquêtes qui continuaient d'utiliser l'ancien GSS, les programmes d'enquêtes-entreprises utilisaient des systèmes propres à leurs applications, qui s'articulaient sur des outils de conception et d'estimation communs. Cependant, à mesure que se multipliaient les efforts en vue d'accroître l'efficacité grâce à une plus grande utilisation des technologies disponibles, par exemple aux étapes de l'entrée et de la vérification des données, la demande d'un plus grand usage partagé des méthodes et des outils destinés à faciliter ces efforts s'est accrue. Fin des années 1980, début des années 1990, l'ABS a créé l'environnement de traitement des enquêtes SPEED (pour *Survey Processing Environment for the Economic Division*). L'objectif principal était d'offrir aux statisticiens responsables des enquêtes-entreprises une unité de présentation pour les diverses enquêtes et la capacité de construire des systèmes sur mesure en utilisant des composantes normalisées pour la conception et la sélection, l'acheminement et la collecte, l'entrée des données, ainsi que la vérification et l'estimation. Le nouvel environnement offrait des interfaces conviviales à partir desquelles il était possible d'appeler les diverses composantes requises dans un système piloté par menu.

L'analyse de rentabilité de l'environnement SPEED promettait une approche normalisée mais souple d'exécution des enquêtes-entreprises permettant de mettre en œuvre la technologie la plus récente de manière efficace et de réaliser des économies grâce à l'amélioration des processus. Toutefois, les efforts et la plupart du financement étaient axés sur la fourniture d'un environnement plutôt que d'outils dans cet environnement, même si certains nouveaux modules d'imputation et d'estimation faisaient partie du projet de développement. Par conséquent, bien que SPEED ait porté fruit dans une certaine mesure, il n'a pas servi de tremplin à de nouvelles approches efficaces d'utilisation de la technologie pour le traitement des données, en raison de la poursuite limitée des efforts en vue d'aboutir aux résultats ayant trait aux outils fournis et utilisés dans des domaines tels que le traitement et la vérification. Les divers secteurs opérationnels étaient libres d'appliquer leur propre approche d'utilisation de l'environnement, et un contenu propre s'articulant sur la collecte et le traitement était élaboré pour les applications.

Par conséquent, le coût du développement et de la maintenance des applications a continué d'augmenter, et plusieurs approches légèrement différentes ont continué de coexister de manière inefficace dans l'environnement « normalisé ». Alors que l'intention était de moderniser l'ensemble complet de fonctions d'enquête dans le cadre de la création de cet environnement, un manque de fonds et la réduction de la portée du projet en raison de dépassements des coûts n'ont pas permis d'y arriver. GENINT et GENEST ont été développés afin d'offrir une gamme d'options d'estimation et d'imputation plus faciles à utiliser que les composantes très fines qui constituaient les fonctions d'enquête antérieures.

Parallèlement au développement des systèmes généralisés pour les enquêtes-entreprises, le Bureau a procédé à celui d'un registre des entreprises. Un registre élémentaire avait été utilisé dans le vaste programme des recensements

économiques menés par l'ABS à la fin des années 1960. Ce registre a fait l'objet d'importantes mises à jour durant les années 1980, y compris le soi-disant ajout d'un programme très ambitieux de « fioritures », tels que le marquage de la date. En raison de plusieurs problèmes de performance, nombre de ces améliorations ne pouvaient pas être activées. Le système du registre manquait aussi considérablement de souplesse pour ce qui était de l'adaptation ou de la modification du modèle d'unité opérationnelle et de la sauvegarde de toute information supplémentaire, telle que des données auxiliaires administratives, au niveau de l'unité. Au cours de la deuxième moitié des années 1980, la complexité du registre a également réduit la capacité de l'ABS de s'adapter à la saisie des données provenant des fichiers électroniques de données fiscales au lieu de la saisie de données sur support papier. La complexité du registre faisait que les modifications apportées pour prendre en charge la saisie de données électroniques a causé des problèmes de système qui ont entraîné d'importants délais et pertes d'information durant la mise à jour des données sur les entreprises. Les répercussions de ces délais et pertes sur les principales séries d'indicateurs économiques étaient un important sujet de préoccupation, et les corrections appropriées ont dû être gérées manuellement pendant de nombreux mois.

Étant donné l'expérience associée à ce premier registre très complexe, une refonte gérée de très près a été entreprise au cours des années 1990. Durant cette refonte, le Bureau a délibérément évité d'essayer d'ajouter des fonctions astucieuses à celles qui fonctionnaient déjà bien dans le registre précédent, pour se concentrer sur la création d'un plus grand degré de souplesse, particulièrement en ce qui concernait les modèles d'unité grâce à une approche orientée objet. Bien qu'elle ait atteint son but, cette refonte reposait malheureusement sur une technologie de pointe qui n'a pas été adoptée comme norme de l'industrie et elle a été suivie, peu après, d'une refonte plus conventionnelle.

Au moment où se posaient ces problèmes particuliers à la fin des années 1980, étant donné la lenteur des mises à jour, surtout dans le cas unités complexes, ainsi que les délais et les pertes d'entités commerciales résultant du passage au transfert électronique des données provenant du bureau de l'impôt, les méthodologistes ont commencé à élaborer des règles opérationnelles communes à toutes les enquêtes-entreprises afin de préciser comment les opérateurs des programmes d'enquête devaient traiter les disparitions et les créations d'entreprises, les changements d'industrie, les fusions, *etc*. de certaines unités sélectionnées dans les échantillons, y compris l'utilisation ou non de nouvelles modalités opérationnelles. Différentes pratiques étaient mises en œuvre dans les diverses enquêtes, généralement pour des raisons historiques plutôt que rationnelles.

Pour tenter de résoudre les problèmes qu'entraînait l'utilisation du registre, les secteurs opérationnels ont pris part activement à ces travaux, et un comité formé de membres de la haute direction des secteurs des statistiques économiques et sur la main-d'œuvre de l'ABS a été formé pour superviser l'élaboration et la mise en œuvre des nouvelles approches communes dans le cadre d'un programme d'intégration des enquêtes.

Le programme a débuté à la fin des années 1980, mais s'est poursuivi au début des années 1990, avec le mandat révisé d'assurer la cohérence des diverses enquêtes, comme les mesures trimestrielles et annuelles du secteur manufacturier. Jusque-là, on s'était limité à expliquer les écarts entre les enquêtes en se fondant sur les différences entre les pratiques utilisées. Toutefois, étant donné la plus grande importance accordée au nouvel énoncé de mission de l'ABS consistant à fournir de l'information pour la prise de décision, nous avons reconnu plus clairement que notre tâche n'était pas simplement de produire un certain nombre de publications d'après différentes enquêtes, mais plutôt d'informer les membres du public. Au lieu d'expliquer simplement les écarts, nous sommes passés à l'élimination de ceux sur lesquels nous exercions un contrôle.

Il n'en reste pas moins que le changement culturel requis à l'échelle de l'organisme ne s'est produit que lentement. Pour arriver à appliquer des méthodes et des processus normalisés, les divers secteurs devaient renoncer à exercer un contrôle et mettre en œuvre des changements qu'ils n'avaient pas inventés eux-mêmes, tout en assurant la continuité opérationnelle. Les procédures élaborées pour tenir à jour les échantillons et les bases de sondage soulevaient des problèmes relativement complexes et il était difficile de convaincre les intéressés qu'elles marcheraient toutes. Les secteurs possédaient leurs propres priorités. Toutefois, le problème a culminé au début des années 1990, quand un tournant de l'économie n'a pas été révélé aussi rapidement qu'il aurait dû l'être par les données produites par l'ABS. Il s'agissait en partie d'un problème administratif lié aux mises à jour des données fiscales dans le registre et en partie d'un problème découlant du traitement particulier des créations et disparitions d'entreprise qui continuait d'être appliqué à certaines enquêtes importantes.

La haute direction a appuyé vigoureusement le projet d'intégration des enquêtes. Des définitions normalisées des périodes de référence des enquêtes, des périodes normalisées pour le tirage des bases de sondage du registre et une approche normalisée pour l'adoption de nouvelles modalités opérationnelles, conformes à la définition de la période de référence de l'enquête, ont été appliquées. On a créé une nouvelle section responsable des bases de sondage communes qui travaillait en collaboration avec le programme du registre. Cette section a élaboré des mesures de la qualité afin de surveiller la mise à jour des données sur les nouvelles entreprises dans le registre et a produit la nouvelle modalité opérationnelle modélisée en vue de tenir compte des entreprises en exploitation durant la période de référence, mais se trouvant encore dans le circuit de mise à jour au moment de la sélection de l'échantillon. Elle a tiré du registre des bases de sondage trimestrielles, a procédé à l'assurance de la qualité des bases de sondage, ainsi qu'à la sélection et à l'acheminement des échantillons pour toutes les enquêtes. Les programmes d'enquête n'avaient plus la liberté de choisir la période de sélection de leur base de sondage, ni la capacité d'ajuster leur liste d'entreprises en fonction de « connaissances locales » ou les échantillons sélectionnés. La section produisait et surveillait aussi les statistiques démographiques sur les entreprises, à titre de vérification de la qualité et de produits statistiques.

Des procédures normalisées de tenue à jour des échantillons et des bases de sondage ont été établies pour toutes les enquêtes et devaient être mises en œuvre durant la collecte. Au cours des années 1990, les opérations de collecte ont été décentralisées et les procédures ont dû être introduites dans les divers programmes d'enquête. Il fallait que ces derniers arrivent tous à comprendre et à appliquer ces procédures, et les premiers efforts dans ce sens, en s'appuyant sur un grand manuel de référence, n'ont pas été très fructueux. À la suite d'un examen et de discussions avec les utilisateurs, une autre approche, fondée sur un questionnaire et un diagramme de cheminement, a donné de nettement meilleurs résultats, car elle était plus facile à comprendre pour les programmes d'enquête et pouvait être utilisée directement durant le processus de collecte. La centralisation subséquente de la collecte des données des enquêtes-entreprises a facilité la bonne application des règles normalisées.

Au cours de la deuxième moitié des années 1990, reconnaissant qu'il était précieux de pouvoir extraire facilement des données sur un sujet donné des diverses enquêtes et d'autres sources, la haute direction s'est vivement intéressée à une approche intégrée de la gestion des données. Cette approche comprenait la création d'un entrepôt de données contenant les données publiées et d'un système normalisé de publication permettant de puiser des données dans cet entrepôt et de les publier. Le programme de développement de la gestion des données était axé au départ sur de grands ensembles de données transversales relativement simples. Le temps était intégré dans le modèle de données simplement comme une autre dimension, analogue à l'industrie, et, pour ces ensembles de données, l'approche donnait de bons résultats. Toutefois, quand le modèle de données a été étendu afin d'englober les séries chronologiques, et que des fonctions de manipulation des séries chronologiques ont été nécessaires, l'approche susmentionnée s'est avérée très inefficace. Ce constat a suscité un important débat et ce n'est qu'après de vives discussions et un examen indépendant qu'il a été admis qu'un modèle tenant compte des propriétés intrinsèques du temps permettrait un traitement nettement meilleur des données chronologiques. Un entrepôt de données spécial pour les séries chronologiques doté de fonctions pour la manipulation de ces dernières a donc été acquis. Par la suite, lorsque des outils permettant de gérer et de diffuser les ensembles, grands et complexes, de données des enquêtes-ménages ont été nécessaires, l'ajout d'un autre ensemble de fonctions fondé sur Blaise et Supercross en tant qu'infrastructure intégrée a été facilement accepté. Malheureusement, il n'a jamais été possible d'interfacer convenablement ces éléments avec le reste de l'infrastructure de gestion des données de l'ABS.

Au début des années 2000, l'ABS a poursuivi l'intégration des enquêtes-entreprises accomplie au début du processus grâce au registre des entreprises et à une unité de base de sondage commune, et à la fin de celui-ci, au moyen de l'entrepôt de données du système de publication. Nous avons d'abord créé un entrepôt commun de données d'entrée, puis nous avons adopté une approche centralisée de collecte des données pour les enquêtes-entreprises. Cette dernière mesure visait à répondre aux demandes constantes de réduire les coûts, y compris ceux de la tenue à jour du grand nombre de systèmes disparates de traitement de la collecte et de l'entrée des données qui avaient proliféré dans l'environnement SPEED et qui nécessitaient tous une maintenance et une mise à niveau, par exemple, pour tirer parti de la technologie Internet.

Reconnaissant que le passage d'une approche axée sur les enquêtes à une approche fonctionnelle nécessiterait un important changement organisationnel, la haute direction a appuyé fermement un programme de gestion du changement de grande portée et bien financée, appelé Business Statistics Innovation Program (BSIP). Ce programme, géré par la haute direction, s'occupait des questions relatives aux ressources humaines, des changements

touchant les processus opérationnels, des structures organisationnelles, des processus de gouvernance et de l'infrastructure technique. Dans l'ensemble, le programme de changement, qui comprenait le déplacement de fonctions entre les huit bureaux, a remporté un franc succès. En ce qui concerne la réduction considérable du nombre de systèmes, les bénéfices ont été presque instantanés; par contre, comme la réalisation de certaines autres économies espérées prend du temps, les processus de gouvernance et l'attention de la haute direction se sont portés sur d'autres priorités, ce qui met en péril la réalisation finale de ces économies. Une petite unité a été maintenue en vue d'assurer le suivi de la réalisation de ces économies, mais les ressources disponibles sont limitées.

Un nouvel ensemble de fonctions d'enquête a également été élaboré dans le cadre du BSIP. Pour l'estimation, diverses options offertes dans les fonctions d'enquête initiales du Système GENEST avaient déjà été mises à jour, mais nous avions aussi appris une autre leçon. Appliqué au nombre de plans de sondage possibles, le nombre d'options offertes multipliait le nombre d'options d'estimation de la variance requises en cas d'estimation directe de la variance. Compte tenu de ce fait et afin de produire une gamme appropriée d'options pour le plan de sondage et l'estimation, tout en s'assurant que l'estimation de la variance demeure gérable, il a été décidé en toute connaissance de cause d'adopter une approche d'estimation de la variance par repondération dans les nouvelles fonctions d'enquête. Autrement dit, la capacité de généraliser la méthode est devenue un critère pour choisir une méthode normalisée. Cette approche a donné de très bons résultats en pratique.

3. Pourquoi recourir à la normalisation ?

Le survol très partiel de l'évolution de l'infrastructure statistique à l'ABS qui précède montre qu'au cours des cinq dernières décennies, le Bureau a procédé à d'importants investissements dans le développement de méthodes et d'outils généraux pour appuyer son travail. Cet effort de développement se poursuit aujourd'hui grâce aux nouvelles générations d'infrastructure statistique dont l'élaboration est en cours. Les avantages de la normalisation des méthodes et des outils sont nombreux. Différents facteurs peuvent jouer un rôle essentiel selon la situation, mais il existe au moins cinq bonnes raisons, qui sont apparentées, de procéder à la normalisation.

La première est l'efficacité. Si les utilisateurs sont nombreux, les coûts de développement et de maintenance par utilisateur demeurent faibles et la maintenance est particulièrement importante ici. Le coût de la maintenance d'une myriade de systèmes différents dans un organisme et de leur mise à niveau à mesure que s'offrent de nouvelles possibilités technologiques ou scientifiques est souvent prohibitif et est le déterminant le plus visible de l'adoption d'approches normalisées. Les coûts de formation sont également plus faibles avec une approche normalisée, car un moins grand nombre d'outils de formation doivent être développés et, avant tout et par-dessus tout, les employés qui acquièrent de l'expérience dans une partie de l'organisation peuvent réutiliser leurs compétences concernant les méthodes et les outils dans une autre partie de celle-ci, ce qui réduit le coût de la mobilité du personnel. Si les travaux de développement sont axés sur une méthode ou un outil standard, ils peuvent être plus rigoureux et mieux gérés, décrits et mis à l'essai, ce qui réduit la portée des erreurs et le besoin de refonte. Toutefois, l'avantage le plus important en ce qui concerne l'efficacité pourrait être la capacité accrue de procéder à des mises à niveau à l'aide de nouvelles méthodes et de nouveaux outils à l'échelle de l'organisme à mesure que l'occasion se présente. Les mises à niveau des composantes normalisées peuvent être réparties facilement à travers l'organisme si elles sont mises en œuvre d'une manière commune.

L'exactitude est un deuxième domaine dans lequel la normalisation est avantageuse. Si les coûts de développement sont partagés entre de nombreux utilisateurs, il est possible de se doter d'une meilleure méthode ou d'un meilleur outil, et une norme bien éprouvée est moins susceptible d'entraîner des erreurs. En outre, comme le nombre de façons de procéder est réduit, les méthodes et les outils normalisés peuvent être bien documentés, compris et gérés au cours du temps. Ils sont donc moins susceptibles de devenir des boîtes noires mal comprises et éventuellement mal utilisées.

La normalisation favorise aussi une plus grande cohérence et une plus grande comparabilité des données au sein de l'organisme. En éliminant la variation inhérente à la façon dont les choses sont faites, il est plus facile de se concentrer sur la variation de la statistique d'intérêt sous-jacente, ou sur les régularités dans les relations entre les statistiques sous-jacentes. Cet avantage est manifeste lorsque les normes ont trait à des aspects tels que les définitions et les classifications des items. Cependant, il en est également ainsi lorsqu'elles s'appliquent aux méthodes et aux

outils utilisés, par exemple pour tenir compte des créations et des disparitions d'entreprises, au traitement des valeurs aberrantes ou à la désaisonnalisation.

Quatrièmement, la normalisation augmente le niveau de contrôle exercé sur le processus, donc les données de sortie. Si le nombre de méthodes et d'outils utilisés est plus faible, il est plus facile d'établir une politique intégrée, par exemple en ce qui concerne la sécurité, la gestion des données et la diffusion. Il est également plus facile de surveiller et de vérifier la mise en œuvre de cette politique.

Une cinquième bonne raison de normaliser les méthodes et les outils est que la normalisation met l'organisme dans une situation où il peut réagir aux nouvelles exigences et maintenir la pertinence. Elle facilite le développement rapide de nouvelles applications. Au lieu de partir de rien, la normalisation donne la possibilité de choisir des méthodes et outils pertinents et de les intégrer en définissant les paramètres appropriés pour une nouvelle utilisation.

3.1 La normalisation a-t-elle un côté négatif ?

Certains inconvénients peuvent être associés à la normalisation. S'il était facile de mettre en œuvre des méthodes et des outils normalisés, nous ne serions pas encore en train d'y consacrer de si grands efforts tant de décennies après les premières tentatives de mise en place de systèmes d'enquête généralisés.

Habituellement, l'établissement de normes demande un compromis. L'optimum global n'est pas l'optimum local. Il est, certes, possible d'offrir des moyens d'adapter une méthode ou un outil normalisé à un cas particulier en fixant les paramètres à des valeurs particulières, mais très souvent, le fait d'offrir suffisamment de souplesse pour permettre de bien adapter les méthodes et les outils à tous les types de situation accroît considérablement la complexité de la norme. Cela peut, à son tour, nuire à la performance au détriment de tous les utilisateurs. Par conséquent, il faut parfois limiter au minimum les fonctions accessoires fournies avec une méthode ou un outil normalisé, ce qui réduit l'adaptation à certaines applications.

Parfois, le coût de la généralisation est simplement trop élevé pour que l'effort en vaille la peine. Un processus commun n'est pas nécessairement une raison suffisante pour adopter une méthode ou un outil commun, si les circonstances diffèrent trop. Par exemple, à de nombreux égards, le processus de contrôle de la circulation est en grande partie le même qu'il s'agisse de circulation aérienne, routière, ferroviaire ou nautique, mais les circonstances diffèrent fortement et, bien qu'il existe une certaine terminologie et une certaine logique communes, fort peu d'outils sont communs à ces divers types de circulation. Ou bien, pour en revenir à la statistique, les tentatives de l'ABS en vue de généraliser les éléments de la gestion des données pour de grands ensembles de données transversales ainsi que de relativement petites séries de données chronologiques ont montré que, si le temps pouvait effectivement être considéré comme une simple dimension utilisée pour décrire les données, ses attributs particuliers rendaient l'approche fort inefficace pour gérer les séries chronologiques comme données de sortie. Le coût de l'utilisation du système généralisé d'entrepôt de données de l'ABS pour la gestion et la manipulation des séries chronologiques de statistiques publiables était prohibitif et le coût de la généralisation aux divers types de données était trop important.

Outre le fait que l'optimum global n'est pas forcément l'optimum local, un autre inconvénient découle de la perte de contrôle local sur la maintenance et l'amélioration essentielles à l'efficacité des outils. Si le financement des nouveaux travaux de développement est centralisé, il n'existe aucune ressource pour les mises à niveau locales lorsque le besoin s'en fait sentir. De surcroît, l'usage partagé des outils intégrés peut poser des problèmes de performance selon la charge qui leur est imposée. Les programmes locaux perdent le contrôle des paramètres qui ont une incidence sur leur capacité de livrer leurs données dans les délais prévus.

Cette perte de contrôle locale de l'ajustement des méthodes et des outils peut aussi avoir des conséquences négatives sur la créativité et l'innovation, et réduire la motivation à rechercher de meilleurs moyens d'effectuer les tâches. Cette situation peut s'empirer à mesure que les secteurs opérationnels sont de moins en moins familiarisés avec les méthodes et outils intégrés, de sorte qu'ils sont moins au courant des possibilités d'améliorer les processus et les outils quand la technologie évolue. S'il n'est pas suffisamment conscient des problèmes opérationnels, l'organisme propriétaire de la norme peut lui aussi manquer des occasions d'améliorer les résultats opérationnels. Il importe donc d'établir une collaboration étroite entre ceux qui sont responsables des méthodes et des outils et les secteurs opérationnels.

Il faut aussi trouver des moyens de continuer à favoriser l'innovation locale et le partage des idées fructueuses au sein de l'organisme. La stratégie d'Apple, qui consiste à encourager le développement et la propagation d'applications qui reposent sur les normes et de permettre à celles qui remportent un succès de survivre tandis que d'autres ne font que passer, donne de bons résultats dans un contexte commercial dont le moteur est la recherche d'un profit. L'adoption par l'ABS de l'environnement SPEED, qui permettait aux secteurs opérationnels de soumissionner pour obtenir les fonds nécessaires pour développer des applications dans l'environnement normalisé, a été bien accepté par les utilisateurs en ce qui concerne la création locale de systèmes dans une infrastructure intégrée, mais elle ne récompensait pas l'usage partagé ni la promotion des nouveaux développements locaux à l'échelle de l'ABS.

Un autre inconvénient de la mise en œuvre d'une nouvelle norme tient au changement qu'elle nécessite. Un secteur qui possède déjà une application locale satisfaisante de son point de vue pourrait hésiter à assumer le coût (en termes monétaires ainsi que d'efforts) du passage à une nouvelle norme intégrée. Si le changement doit être appliqué dans un grand nombre de secteurs, le coût peut être très important. Une source de financement, telle qu'une injection de capital, peut être nécessaire. Si les organismes statistiques peuvent souvent s'adresser aux utilisateurs pour financer l'exécution de nouveaux travaux statistiques importants, il leur est parfois beaucoup plus difficile de trouver de nouvelles sources de financement pour développer et mettre en place une infrastructure statistique destinée à un usage plus généralisé.

De surcroît, il est rare que le changement requis soit simplement d'ordre technique, tel que la suppression d'un système pour en raccorder un autre. Le changement peut au contraire nécessiter des modalités de gouvernance différentes comprenant un contrôle centralisé de choses qui étaient gérées localement auparavant. Il pourrait impliquer que des employés remplissent des fonctions différentes, que différentes compétences, de nouvelles structures organisationnelles, de nouveaux moyens de travailler au sein de l'organisme, et ainsi de suite, soient nécessaires. Autrement dit, l'élaboration et la mise en œuvre d'une norme peut requérir une gestion importante du changement dans tous les secteurs touchés et, si une norme doit être appliquée à une grande partie de l'organisme, le changement doit être traité comme un changement organisationnel important, comportant tous les aspects complexes d'un changement culturel, organisationnel, ainsi que technologique.

Lors du lancement d'un programme important de normalisation des méthodes et des outils, il convient de planifier et de gérer l'ampleur de la tâche jusqu'à la mise en œuvre complète et la réalisation des résultats comme s'il s'agissait d'un programme de changement. L'ampleur du programme qui est lancé doit être proportionnée aux avantages que l'on pourra en tirer et aux ressources disponibles pour réaliser le changement.

4. Planification d'un important projet de développement et de mise en œuvre d'outils et de méthodes normalisés au sein d'un organisme : six décisions clés qui doivent être prises

L'entreprise d'un programme de normalisation des outils et des méthodes au sein d'un organisme statistique représente un grand projet de changement qui entraînera des coûts importants sur le plan tant monétaire que de l'utilisation de ressources limitées. L'organisme visera à intégrer dans les méthodes et outils normalisés les meilleures possibilités qu'offrent les nouvelles technologies et les nouveaux développements dans le domaine de la statistique, ce qui demandera le concours de membres clés du personnel pour le développement et la mise en œuvre de ces outils et méthodes, tout en nécessitant le maintien des « opérations habituelles » fondées sur les anciennes méthodes et technologies.

Voici certaines décisions de gestion importantes qui doivent être prises :

1. Quels sont les principaux déterminants de la normalisation ? Par exemple, l'objectif est il de réaliser un gain d'efficacité, d'établir un contrôle ou d'obtenir d'autres avantages liés à la qualité ?
2. Quelle sera la portée de la normalisation ? Quelle part des processus complets de l'organisme sera couverte ? À quels secteurs spécialisés/programmes d'enquête, *etc*. les outils s'appliqueront-ils ? Et quel sera l'horizon temporel des normes qui sont élaborées (par exemple, doivent-elles permettre d'assurer le lien avec les possibilités techniques qui, selon les prévisions, atteindront le stade de maturité, disons, dans cinq ans) ?

3. Qui assume la responsabilité et dirige le développement et la mise en œuvre ? En particulier, qui prend les décisions importantes d'aller ou non de l'avant, établit les priorités d'utilisation des ressources, détermine les changements de portée du projet au besoin, *etc*., est associée à cela la détermination des structures de gouvernance nécessaires tout au long du déroulement du programme jusqu'à la réalisation des avantages escomptés.
4. Quel est le degré de normalisation qu'il faut atteindre et quel est le degré de complexité des solutions qui sera toléré (combien de souplesse sera offerte pour les solutions de rechange, éventuellement au détriment du coût, de la performance et de la facilité d'utilisation, et quelles sont les exigences de maintenance) ?
5. L'utilisation des méthodes et outils normalisés par les secteurs opérationnels sera-t-elle obligatoire ou facultative ?
6. Quel degré de granularité (niveau de détail) faudra-t-il utiliser dans la conception des modules (tension entre souplesse et utilité) ?

Suit une discussion de chacune de ces questions.

4.1 Déterminants de la normalisation

Les avantages de la normalisation ont été examinés plus haut. L'approche de normalisation adoptée dépendra de la combinaison de ces avantages qui est essentielle à la mise en œuvre d'un programme proposé. S'il est impératif de réduire immédiatement les coûts, la priorité sera accordée aux éléments de la normalisation qui produiront facilement le plus d'économies, tandis que si la qualité et le contrôle sont les principaux motifs, l'attention se portera d'abord sur les domaines présentant le risque le plus important. Si des compromis doivent être faits, par exemple entre l'efficacité (nombre d'outils, fonctions accessoires, *etc*.) et la qualité (par exemple, niveaux de fonctionnalité nécessaires pour répondre à divers besoins locaux), ils doivent l'être de manière systématique en suivant une direction globale convenue.

Puisqu'un programme de normalisation est un programme de gestion du changement et qu'il nécessitera vraisemblablement un changement de culture et l'engagement personnel des employés d'un certain nombre de secteurs, il est utile de communiquer clairement l'objectif. Si un objectif clé est d'accroître l'efficacité, il sera tentant de présenter aussi le projet comme un effort en vue d'accroître la qualité, particulièrement si l'amélioration de la qualité est un résultat probable. Le danger dans ce cas est que la réalisation de l'avantage escompté, c'est à dire le gain d'efficacité, soit voilée par une tendance des employés à opter pour des compromis en faveur de la qualité qui ne sont pas nécessairement en harmonie avec l'exigence principale de la direction.

4.2. Portée

La portée doit être déterminée en fonction des résultats clés souhaités, des ressources disponibles (monétaires et capacités) et de tout obstacle, par exemple d'ordre culturel, qui doit être surmonté.

Cela peut sembler être une évidence, mais nous voyons fréquemment des exemples de mauvais alignement des ambitions et des capacités faisant que des travaux se voulant de grande envergure au départ voient leur portée réduite considérablement par après, de façon, par exemple, à fournir « au moins la fonctionnalité du système précédent », ce qui limite considérablement les avantages réalisés comparativement aux promesses du plan d'activités initial. Un inconvénient majeur de cette « ambition exagérée » est que l'on a tendance à affecter une quantité excessive de ressources aux premières étapes en vue d'explorer et d'élaborer un large éventail d'options et à sous investir aux dernières étapes en vue de réaliser les résultats grâce à une mise en œuvre et une surveillance efficaces.

Quelques causes éventuelles d'une ambition exagérée concernant la portée ?

- Simple avidité : Nous voyons des possibilités et nous n'aimons pas dire non, à nous mêmes ni à d'autres, de sorte qu'il nous arrive d'avoir les yeux plus gros que le ventre, tant du point de vue des capacités que du budget.
- Absence d'une estimation honnête du niveau réel de capacité et de la rapidité avec laquelle le niveau requis peut être atteint.
- Niveau général d'optimisme excessif, par exemple penser que les choses marcheront du premier coup ou que le taux de roulement du personnel sera faible.

- Manque de conviction que les coûts estimés sont raisonnables. (Ont-ils été gonflés volontairement en guise de stratégie d'atténuation du risque, ou parce que l'on s'attend à n'obtenir qu'un financement partiel seulement ? Le dernier point mène à un cycle de surestimation des coûts et de sous-financement qu'il peut être difficile de rompre.) Ce manque de confiance dans les coûts pousse souvent la direction à réduire les budgets tout en souhaitant maintenir la portée complète du projet.
- Un problème connexe peut se poser si la direction, soumettant une demande de fonds à une source externe, fait une demande pour un projet d'une portée qui dépasse sa capacité parce qu'elle s'attend à ne recevoir qu'un financement partiel et qu'elle sera obligée de réduire la portée du projet. Si le financement complet est obtenu, l'organisme risque de manquer de capacité.
- Manque d'appréciation du fait que le programme doit être planifié, géré et doté en ressources des premières étapes jusqu'à la réalisation finale du bénéfice. Cela se traduit par une tendance à ne tenir compte que des coûts de développement pour déterminer la portée du projet, en supposant que les coûts de la mise en œuvre opérationnelle seront relativement faibles et pourront être « absorbés ».
- En rapport avec le point précédent, le fait éventuel de ne pas apprécier l'effort qu'il faut déployer pour faire évoluer la culture et l'entendement d'un organisme jusqu'à un point où les secteurs opérationnels comprennent les projets de développement intégré, y accordent de la valeur et y contribuent.

Les quatre premiers points susmentionnés sont les principales raisons pour lesquelles, souvent, les produits d'un programme de normalisation des méthodes et des outils n'atteignent pas les objectifs du plan initial en ce qui concerne la couverture ou la fonctionnalité. Cependant, un plan d'activités justifie les coûts en regard des bénéfices réalisables, et non des produits, et ce sont les deux derniers points qui représentent vraisemblablement la raison principale pour laquelle, même si les produits sont livrés, les bénéfices précisés dans l'analyse de rentabilité (par exemple, réduction des coûts de traitement, production plus rapide) sont nettement inférieurs au niveau promis et théoriquement réalisable.

Le niveau de pertinence du dernier point varie d'un organisme à l'autre. Certains sont bien structurés pour appuyer des approches intégrées et possèdent déjà une solide culture organisationnelle. L'ABS rentre relativement bien dans cette catégorie. D'autres, peut être formés par la fusion récente de différents organismes, ou financés sur la base d'opérations compartimentées, peuvent avoir plus de difficulté à créer une culture où les individus comprennent et apprécient les avantages à l'échelle de l'organisme et ont la motivation de contribuer à la réalisation des résultats.

Lors de l'examen des ressources requises pour réaliser les bénéfices, il est important de déterminer le coût de tous les éléments du processus de changement, y compris ceux nécessaires pour que les changements prévus soient adoptés par les divers secteurs opérationnels de l'organisme et qu'ils le demeurent. Il est parfois préférable de limiter la portée du projet au départ, mais de veiller à ce que les ressources soient disponibles et que la gouvernance soit maintenue tout au long du projet jusqu'à la réalisation des bénéfices.

Durant les grands travaux de développement de l'ABS des années 1970 et du début des années 1980, l'une des préoccupations fréquentes était le piège de l'investissement, de grands projets prometteurs de bénéfices considérables accusant en réalité des dépassements importants des délais et des coûts. Le seul moyen qu'avait alors la direction de tirer une certaine valeur de l'investissement effectué était de poursuivre les dépenses sous la promesse que les bénéfices étaient imminents.

Une stratégie efficacement mise en œuvre par l'ABS au cours des années 1980 a consisté à imposer la règle selon laquelle chaque nouveau projet de développement devait donner un produit livrable autonome dans les six mois, sans que plus de dix personnes à la fois travaillent sur le projet. Les projets de plus grande envergure étaient décomposés en morceaux plus petits capables de produire d'eux mêmes de la valeur.

Cependant, cette stratégie comporte certains pièges qu'il convient d'éviter. En particulier, elle favorise le développement progressif d'un système normalisé à partir d'applications « prototypes » particulières, ou encore les approches fondées sur une « validation de principe » qui ne sont pas nécessairement bien adaptées à toutes les applications destinées à utiliser les normes. Si le premier produit livrable dans les six mois est un ensemble de méthodes et d'outils pour la vérification des opérations de collecte d'une enquête particulière et que l'intention est de poursuivre le développement de cet ensemble pour en faire une norme applicable à de nombreuses opérations de collecte, il faut veiller à ce que l'approche de développement suivie s'avère vraiment efficace pour la série complète d'opérations prévues.

En général, afin de déterminer la portée du projet qui concorde avec la disponibilité des ressources, il faut entreprendre une certaine forme de programme en plusieurs phases, et savoir comment définir la portée de chaque phase sous l'angle des parties du système statistique qui doivent être couvertes, non seulement en ce qui concerne les domaines spécialisés et les opérations de collecte, mais aussi les éléments de la chaîne de valeur statistique qui doivent être inclus dans les travaux de développement. Voici les questions auxquelles il convient de répondre ici :

- Où se situent les points d'arrêt logiques dans le processus statistique ? Comment les raccordements entre les processus remaniés et les anciens processus seront-ils construits et comment peut-on réduire au minimum le coût de ces raccordements ?
- Existe-t-il des dépendances ? Est-il possible de réaliser les bénéfices prévus dans l'analyse de rentabilité avec la portée choisie ? Toutes les composantes nécessaires pour réaliser les bénéfices sont elles incluses dans la portée (y compris les composantes de formation et de mise en œuvre) ?
- Où se situent les plus grandes possibilités d'atteindre les résultats escomptés (par exemple, les secteurs les plus inefficaces ou ceux présentant le risque le plus élevé à l'heure actuelle) ? Où se situent les possibilités les plus faciles à exploiter (par exemple, processus moins complexes ou secteurs opérationnels plus engagés) ?
- Dans quels secteurs les possibilités d'amélioration évoluent elles encore, par exemple grâce à des technologies naissantes ? Il pourrait être raisonnable de retarder les travaux dans ces secteurs jusqu'à ce que l'évolution arrive à un certain stade de maturité.

Quelle que soit la façon dont la portée est déterminée afin qu'elle concorde aux ressources disponibles, elle doit être complètement définie d'avance, afin de s'assurer, par exemple, que si les premiers travaux se concentrent sur les fruits se trouvant sur les branches basses, le développement puisse néanmoins être étendu efficacement de manière à cueillir tous les fruits que l'on veut placer dans la coupe. Il est vraisemblable qu'essayer de faire entrer dans un moule qui ne convient pas des applications plus complexes appartenant à des secteurs moins consentants entraînera d'importants problèmes de performance et empêchera de réaliser les bénéfices.

4.3. Qui fait siens et dirige les travaux de développement et de mise en œuvre ?

Si l'on veut effectuer d'importants changements au sein d'un organisme, il est important de comprendre quelle sera l'incidence de ces changements sur le personnel et quels seront les éléments motivateurs pour divers groupes. Les organismes statistiques sont souvent régis par des valeurs fortes qui peuvent faciliter ou entraver les processus de changement. Ces valeurs s'articulent sur l'intégrité, la qualité et le service. Les coûts et l'efficacité ont tendance à occuper un rang moins élevé. Les coûts deviennent un facteur quand l'une de ces valeurs fondamentales est menacée. Par exemple, une compressions budgétaire importante peut devenir un motivateur essentiel de la mise en œuvre d'approches plus rentables si elles permettent de maintenir les niveaux d'intégrité, de qualité et de service qui, autrement, pourraient être compromis. Quel que soit le facteur, pour que le changement soit effectué efficacement, les secteurs opérationnels doivent être parties prenantes, de l'étape du développement à celle de la mise en œuvre, à tous les niveaux hiérarchiques en partant de la haute direction. Cela ne veut pas dire que le développement doit se faire dans les secteurs opérationnels, mais que ceux ci doivent avoir la conviction qu'il est effectué pour eux, avec leur engagement.

Les secteurs opérationnels sont facilement motivés par leurs cadres supérieurs et par leurs clients. Ils ne le sont pas aussi facilement par une idée ou un outil génial venant d'un autre secteur de l'organisme, tel que la méthodologie ou les services techniques. Si les secteurs opérationnels ne font pas leur le projet proposé et ne sont pas fermement convaincus que les résultats escomptés sont à la fois réalisables et utiles, l'appui pour le développement sera faible, et la mise en œuvre sera très difficile. Quel que soit le secteur qui entreprend les travaux de développement, les cadres supérieurs des secteurs opérationnels doivent s'approprier le projet, l'appuyer en fournissant l'expertise appropriée, poser les questions difficiles concernant l'avancement des travaux et l'établissement des priorités, et participer aux prises de décisions concernant tout rajustement nécessaire de la portée du projet. La haute direction doit demeurer partie prenante, du développement à la réalisation des bénéfices en passant par la mise en œuvre.

Les compétences et les connaissances des secteurs opérationnels sont essentielles à l'élaboration de nouvelles approches et à la résolution des problèmes qui se posent. Les questions pratiques concernant ce qui pourrait ou non fonctionner, tant du point de vue des processus que des problèmes statistiques et des besoins des clients, doivent être prises en compte durant les travaux de développement et de mise en œuvre. Par ailleurs, les anciennes façons de procéder ne devraient pas restreindre le développement, de façon qu'idéalement, les personnes qui connaissent les

secteurs opérationnels puissent sortir des sentiers battus et réfléchir à ce qu'il serait possible d'accomplir à l'aide de nouvelles technologies et méthodes. L'infrastructure normalisée doit tenir compte des divers besoins au sein de l'organisme. Par conséquent, il est généralement efficace d'isoler les travaux de développement des pressions quotidiennes en vue de maintenir le statu quo.

Les méthodologistes et les secteurs techniques sont des membres importants de l'équipe responsable de créer l'infrastructure intégrée, de même que les secteurs opérationnels. Les méthodologistes, surtout s'ils ont participé au cours de leur carrière aux activités de divers secteurs opérationnels et à divers travaux de développement pratique, auront une vision transversale de la façon dont les choses pourraient se faire et, culturellement, une propension à accepter le changement et à saisir les occasions qui se présentent. Par ailleurs, les méthodologistes qui sont demeurés isolés des secteurs opérationnels, dans des fonctions hautement spécialisées, ou qui se considèrent uniquement comme les gardiens d'une vision étroite de la qualité peuvent résister à la modification des méthodes ou des processus passés aussi résolument que n'importe qui d'autre.

En général, l'approche idéale consiste à former une équipe multidisciplinaire bien motivée, « appartenant » à la haute direction des secteurs opérationnels, située en un seul endroit et participant à l'unisson à un programme de travail intégré dont les priorités sont déterminées par un seul ensemble de mécanismes de gouvernance.

Les questions qui se posent au sujet d'une telle équipe sont les suivantes :

- Comment libérer l'importante capacité opérationnelle et la capacité méthodologique et technique qui sont également nécessaires pour poursuivre les opérations ordinaires (c'est-à-dire comment rediriger des ressources limitées des opérations quotidiennes vers les travaux liés au changement stratégique) ?
- Dans quelle mesure faut-il séparer les équipes chargées du développement de l'équipe qui s'occupe des opérations « ordinaires », surtout si les ressources sont limitées ?

Ces deux points se recoupent, car une séparation importante des travaux de développement et des opérations « ordinaires » peut nécessiter plus de spécialistes du domaine, lesquels sont peu nombreux, que la combinaison des travaux courants et des travaux de développement sous une seule structure de gestion. La meilleure approche dépend de la portée du projet entrepris, mais dans le cas de très grands projets, il semble logique de créer la « nouvelle » version de l'infrastructure sous une structure de gestion distincte de celle utilisée pour les opérations courantes, afin de pouvoir se concentrer spécialement sur les travaux de développement et assurer une participation de haut niveau à ces derniers. Les projets apparentés peuvent être regroupés dans un programme et s'appuyer les uns les autres sous ce modèle en maintenant le tonus du projet. Le transfert des anciennes opérations dans le nouvel univers peut alors demander beaucoup d'efforts si l'équipe s'occupant des opérations « ordinaires » n'a pas suffisamment participé durant la phase de développement. L'ABS a souffert de ce genre de séparation dans certains grands projets, dont le projet de gestion des données des années 1990. Les équipes d'autres projets, tels que le *Business Statistics Innovation Program*, se sont efforcées par tous les moyens de maintenir cette participation à tous les niveaux en partant de la haute direction, ce qui a donné de bons résultats.

Une autre approche, adoptée à l'ABS dans le cadre du programme de modernisation courant, consiste à créer une structure de gestion distincte pour le programme de développement, mais d'inclure dans le contrôle de la haute direction un groupe de secteurs opérationnels les plus touchés par les aspects clés du programme de développement à un moment particulier, en prévoyant que ces secteurs entrent dans la structure de gestion et en sortent à mesure que progressent les travaux de développement. L'équipe de gestion est alors responsable des opérations ordinaires ainsi que des travaux de développement du groupe de secteurs les plus touchés par la modernisation, mais dont la composition évolue au cours du temps. La communication avec les autres secteurs opérationnels doit être maintenue, car les outils généralisés doivent satisfaire à des exigences plus générales, et une leçon importante que nous avons tirée de notre expérience passée est qu'il faut se concentrer sur la gestion durant tout le processus jusqu'à la réalisation des bénéfices, particulièrement pour ce qui est de décider quand et sous quelle modalités de gouvernance les secteurs opérationnels quittent le groupe de modernisation.

4.4. Degré de normalisation à atteindre et degré de complexité des solutions qui sera toléré

Deux questions connexes dont doit tenir compte la direction sont la mesure dans laquelle une méthode et/ou un outil sera utilisé par toutes les applications possibles et le degré de complexité qui sera toléré, géré et appuyé par des ressources dans une normalisation afin de permettre qu'elles soient adaptées à différentes applications.

Il existe une bonne raison de vouloir « opter pour la simplicité ». Les systèmes simples ont tendance à mieux fonctionner et à être plus faciles à utiliser, à maintenir et à mettre à niveau. Cependant, si l'outil ou la méthode doit être adaptable à diverses applications, sans perdre trop d'optimalité locale, il devra vraisemblablement offrir la possibilité de déterminer un certain nombre de paramètres pour arriver à un ajustement adéquat. Et cela augmentera la complexité. Il s'agit essentiellement d'une question de compromis entre le coût et le bénéfice, et les décisions doivent être prises en s'appuyant sur une bonne analyse des coûts de la complexité et des bénéfices d'une mise en œuvre locale plus efficace. La question importante est de veiller à ce que les coûts ainsi que les bénéfices aient été bien vérifiés. Le cas du développement du registre de l'ABS durant les années 1980, où les importantes fonctions intégrées dans le registre n'ont jamais pu être activées en raison de la perte de performance connexe, est une bonne illustration de la nécessité d'évaluer à fond la performance tout au long du cycle de développement.

Lorsque la complexité est le résultat de la fourniture d'une plus grande fonctionnalité en vue de traiter, en fait, des circonstances complexes qui se présentent en réalité, on peut être obligé d'accepter la complexité et de la gérer grâce à une bonne conception. Chaque décennie, les problèmes que nous cherchons à résoudre en tant que statisticien deviennent plus complexes. Nous voyons augmenter la demande de données intégrées et cohérentes provenant de diverses sources, à divers niveaux, observées à différents points dans le temps, en minimisant le fardeau imposé aux fournisseurs de données, en offrant divers modes de collecte et en maximisant l'utilisation de données par les analystes grâce à des méthodes de diffusion multiples et évolutives, tout en assurant le respect de la confidentialité malgré le nombre croissant de sources d'information ayant une incidence sur les possibilités d'identification. Si nous ne résolvons pas ces problèmes de plus en plus complexes au moyen de nos ensembles de méthodes et d'outils normalisés, c'est aux secteurs locaux qu'en incombera la responsabilité, à un coût nettement plus élevé pour l'organisme.

Donc, au lieu du mot d'ordre « opter pour la simplicité », il est plus pertinent de réfléchir en termes de gestion de la complexité grâce à une bonne conception et une bonne mise en œuvre (y compris une bonne formation) et à une évaluation appropriée des répercussions sur la performance et de l'affectation des ressources pour la résolution des problèmes. Il suffit de nous souvenir que l'on a pu gérer la complexité des tâches nécessaires pour envoyer un homme sur la lune. Cela a simplement coûté très cher. La complexité en soi ne devrait pas être évitée, mais le degré de complexité devrait être fondé sur l'analyse des coûts et des avantages de la gestion de cette complexité, et cette dernière ne devrait pas être perçue par l'utilisateur. Dans une approche de « prêt à l'emploi », les spécialistes doivent pouvoir soulever le couvercle et comprendre la complexité, mais les utilisateurs ordinaires ne devraient pas en avoir besoin. La gestion de la complexité requiert une expertise pertinente considérable et en l'absence de celle ci, la complexité devrait définitivement être évitée.

4.5. L'utilisation des méthodes et des outils développés sera-t-elle obligatoire ou facultative pour les secteurs opérationnels?

Une autre décision de gestion qui doit être prise est celle de savoir si les méthodes et les outils normalisés devront obligatoirement être adoptés et utilisés par tous les secteurs opérationnels pertinents ou s'ils seront mis à la disposition des secteurs opérationnels qui auront l'option de les utiliser ou non. Cette décision dépend en partie de la raison de la normalisation. Si le but est d'exercer un contrôle, par exemple pour s'assurer que tous les produits publiés et publiables sont diffusés de manière appropriée et satisfont aux normes organisationnelles de qualité et de gestion de la confidentialité, l'utilisation des outils et méthodes normalisés sera obligatoire.

Par contre, si l'objectif est de réaliser des gains d'efficacité et si l'on peut montrer que les secteurs opérationnels sont les mieux placés pour juger de l'applicabilité des méthodes et outils à leurs opérations, l'organisme peut considérer approprié de laisser à ces secteurs la décision d'adopter ou non les outils et méthodes normalisés. Cela éliminera aussi la situation monopoliste dont pourraient autrement jouir les concepteurs, et les inciter à s'assurer que les produits qu'ils développent répondent aux besoins du secteur opérationnel et qu'ils puissent être mis à niveau au besoin par ce dernier. Toutefois, ce raisonnement repose sur la notion que le secteur opérationnel est fortement

motivé à choisir l'option qui répond le mieux aux besoins de l'organisme plutôt qu'aux besoins locaux. Si un secteur opérationnel possède déjà des ressources adéquates et qu'il ne bénéficiera pas lui même d'un accroissement de l'efficacité, que ce soit au niveau local ou de l'organisme, il pourrait n'avoir aucune raison de s'engager dans un processus de changement qui pourrait comporter une perte de contrôle, comme il est mentionné plus haut.

Sous la contrainte d'utilisation, il incombe à la haute direction qui a pris cette décision de veiller à ce que les choses se passent bien. Cette situation responsabilise davantage la fonction de gestion de l'organisme et est moins confortable pour la haute direction.

4.6. Degré de modularité des outils

Les méthodes et outils normalisés sont souvent construits sous forme de modules afin que les personnes qui développent une application puissent choisir parmi une série d'options qu'elles joignent au moyen d'interfaces conviviales, d'où l'analogie des applications prêtes à l'emploi (« brancher et utiliser »). Cette approche offre la souplesse de pouvoir adapter les outils et méthodes à des applications particulières et de répondre à des besoins encore inconnus qui pourraient survenir dans l'avenir.

Le système général d'enquête développé par l'ABS au cours des années 1970 était antérieur à l'existence de cette approche modulaire et le système livré à l'utilisateur était complet, mais en grande partie dépourvu de souplesse du début à la fin (quoique certains moyens souples étaient également fournis pour l'interrogation et la modification des bases de données). L'approche des fonctions d'enquête adoptées par l'ABS au cours des années 1980 reconnaissait la nécessité d'offrir une certaine souplesse pour répondre à divers besoins, y compris certains besoins futurs, et que cette souplesse pouvait être obtenue au moyen de modules, mais la modularisation est allée trop loin... chaque module résolvait une trop petite partie du problème. Les modules devaient être assemblés par un expert. Étant donné la grande mobilité du personnel, les outils doivent pouvoir être adoptés et utilisés de manière intuitive, afin d'obtenir des résultats utiles en possédant un minimum d'expérience.

5. Gestion de grands projets de normalisation – certaines choses à faire et à ne pas faire inspirées de l'expérience passée

Lorsque la direction d'un organisme entreprend un programme important de normalisation des méthodes et des outils, et définit la portée du programme et prend les diverses décisions qui établiront dans les grandes lignes de l'approche suivie pour le développement, certaines stratégies qui se dégagent des expériences antérieures semblent permettre d'aboutir aux résultats souhaités. À mon avis, les éléments qui suivent sont essentiels :

- Voir grand et définir les possibilités de façon générale en regard d'un budget plus important que celui qui pourrait être disponible et d'un horizon temporel éloigné en ce qui concerne les possibilités.
- Revenir à la réalité et définir les diverses étapes en essayant dans la mesure du possible de déterminer l'ordre des composantes sur la base d'un petit projet à livraison rapide, sans exclure des intentions à plus long terme, et en veillant à ce que tous les membres de la haute direction comprennent le projet et donnent leur aval.
- Cerner les parties du programme présentant d'importantes interdépendances et planifier minutieusement... l'application initiale dans ces domaines peut fournir rapidement certains produits livrables, mais il est important de s'assurer que la forme de la solution conviendra à toutes les interdépendances et pas seulement à la première application.
- Planifier et gérer les résultats et non les produits, comprendre les changements de comportements et de rôles individuels et organisationnels nécessaires pour réaliser les bénéfices du projet.
- Adopter de bonnes approches de gestion de projet pour s'assurer que les objectifs soient clairs, que les processus de prise de décision soient appropriés, que les responsabilités soient bien définies et que les systèmes de surveillance soient efficaces tout au long du projet jusqu'à l'obtention des résultats; ne pas abandonner la gestion du projet une fois que les produits ont été développés.
- Veiller à ce que la haute direction continue de piloter le projet jusqu'à la livraison des résultats.
- Veiller à ce que les secteurs opérationnels adhèrent au projet; idéalement, ils devraient déterminer les problèmes qui doivent être résolus et discuter des solutions, par exemple durant des discussions de groupe.

- Être réaliste en ce qui concerne les ressources et les capacités, même si cela est désagréable; veiller à ce que les communications soient claires et honnêtes; si certaines composantes, telles que la mise en œuvre, doivent être financées au moyen du budget courant, expliquer explicitement ce qui sera supprimé du programme de travail.
- Amener l'équipe de projet à ne faire qu'un, à adopter un ensemble commun d'objectifs, de priorités, d'échéanciers et de langage, que ses membres proviennent des domaines techniques et méthodologiques ou opérationnels.
- Faire des essais très tôt, faire des essais fréquents et être honnête (obtenir la participation du secteur opérationnel, établir des relations de confiance et demander son avis pour résoudre le problème).
- Reconnaître ses erreurs quand les choses commencent à aller mal et y remédier rapidement (ne pas se dérober).
- Modifier l'ordre des priorités au besoin pour livrer la valeur maximale et être clair au sujet de tout changement et de ses répercussions sur la réalisation des bénéfices. Ne pas permettre que s'établisse un manque de confiance de la part des secteurs opérationnels clients.
- Communiquer régulièrement, clairement et honnêtement.
- Au départ, éviter que la perfection soit l'ennemi du bien.
- Vers la fin, éviter qu'un « résultat presque suffisant soit considéré comme suffisamment bon », maintenir le tonus jusqu'à ce que le résultat soit obtenu; cela signifie de fournir le soutien promis en parlant la même langue que les secteurs opérationnels.
- Veiller à ce que les secteurs opérationnels, et pas seulement les développeurs, soient responsables; les secteurs opérationnels doivent se concentrer sur la résolution du problème durant la phase de mise en œuvre et ne pas adopter une attitude du genre « Je vous avais prévenu ».
- Surveiller (disons 12 mois, trois ans après la mise en œuvre) et prendre des mesures pour s'assurer que les bénéfices escomptés soient réalisés et que les nouvelles méthodes et les nouveaux outils soient adoptés et le demeurent.

Par ailleurs, l'histoire nous dit qu'il faut faire attention à certains signaux de danger :

- La solution d'un problème qui n'existe pas (l'idée géniale de quelqu'un est souvent géniale, mais les secteurs opérationnels doivent s'y intéresser et l'adopter comme résolvant un problème réel).
- Quelque chose de mal défini... de flou et qui se trouve là, mais qui résiste à une définition et à une gestion claire.
- Un conseil donné par une partie intéressée.
- La dernière pointe du progrès (prévoir un budget et une capacité beaucoup plus importants si, pour une raison quelconque, l'on veut expérimenter avec la pointe du progrès).
- Une équipe divisée.
- Un projet dont la dotation en ressources est insuffisante, particulièrement si le manque de ressources concerne l'étape de la mise en œuvre, et surtout si des raccourcis sont pris pour des composantes telles que la documentation et la mise à l'essai.
- Des problèmes de performance.
- Arrêt de la gouvernance une fois que le produit est prêt, mais avant que le résultat ne soit atteint.

6. Conclusion

La normalisation a de nombreux avantages et, par conséquent, il n'est pas étonnant que les organismes statistiques s'y soient employés de manière très fructueuse, mais en tirant certaines leçons au cours de plusieurs décennies. Cette expérience aide à orienter les nouveaux efforts à mesure qu'évolue la technologie et que surviennent d'autres possibilités et besoins. Nous connaissons dès le départ certains des problèmes qu'il convient de résoudre et des décisions qu'il faut prendre, et nous pouvons nous préparer à prendre les décisions sagement. Nous connaissons aussi certains pièges à éviter et certaines choses qui semblent donner d'heureux résultats.

Évidemment, de nouveaux problèmes se poseront et de nouvelles réponses devront être trouvées à mesure qu'évolue le contexte environnemental, mais la leçon peut être la plus importante tirée de l'expérience est vraisemblablement que les efforts de normalisation doivent demeurer entièrement pertinents. Une technologie ou une méthodologie ingénieuse ne suffit jamais. Pour arriver à des résultats durables, un important projet de normalisation sous tendant une infrastructure statistique doit être considérée comme un processus de changement. Pour que les efforts soient couronnés de succès, la haute direction doit comprendre clairement les résultats qu'il faut atteindre, être convaincue

de leur bien fondé et fournir un leadership ferme afin de s'assurer que leur réalisation perdure. Le succès tient davantage aux personnes qu'à des solutions techniques.

SÉANCE 2A

ARCHITECTURE OPÉRATIONNELLE DU BUREAU : BASES DE SONDAGE DES ENQUÊTES-MÉNAGES

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Remaniement à Statistics Netherland

Frank Hofman[1,2]

Résumé

Statistics Netherlands (SN) doit relever plusieurs grands défis, à savoir améliorer l'efficacité et la qualité des statistiques clés, tout en réduisant le fardeau administratif. Pour y arriver, SN a mis en place une architecture intégrée, des services opérationnels génériques et une boîte à outils standard. Par la suite, les processus statistiques doivent être remaniés un à un, conformément aux principes architecturaux et grâce aux services génériques et à la boîte à outils. Une approche de remaniement a été élaborée, en vue de faciliter les projets de remaniement individuels.
L'approche de remaniement est fondée sur le processus de conception statistique, dans l'optique de l'architecture intégrée, et est combinée au processus RUP pour le développement de logiciels. Deux principes supplémentaires complètent le processus principal de conception statistique et de développement de logiciels. Il s'agit des suivants : 1) une étude préliminaire et un lancement de projet précèdent le processus principal ; 2) les éléments les plus importants du projet de remaniement font l'objet d'un examen centralisé.
En général, l'approche de remaniement s'est révélée utile, même si certains commentaires s'imposent.

Mots clés : Approche ; architecture ; processus de conception ; document.

1. Introduction

Statistics Netherlands est en voie de procéder à des changements radicaux. Un certain nombre de facteurs de changement très hétérogènes présentent des défis majeurs qui ne pourront être relevés que si le mode de fonctionnement de l'institut fait l'objet d'un remaniement majeur. Van der Veen (2007) donne un aperçu de la situation actuelle et des défis qui se poseront pour l'avenir dans un contexte large. De façon plus particulière, l'efficience et la qualité des statistiques clés doivent être améliorées, et le fardeau administratif doit être réduit considérablement.

Pour pouvoir relever ces défis, un programme ambitieux de modernisation, le plan directeur « Counting on Statistics » (Ypma et Zeelenberg, 2007) a été lancé en 2005. Dans le cadre de ce plan directeur, l'architecture opérationnelle, certains services opérationnels génériques aux entreprises et une boîte à outils standard ont été élaborés (voir Braaksma, 2009). L'objectif ultime du plan directeur est de remanier tous les processus statistiques selon les principes architecturaux, ce qui signifie l'adoption à grande échelle de la boîte à outils et des services génériques. Le remaniement de tous les processus statistiques, qui sont au nombre de plusieurs centaines, représente une tâche énorme et ne peut se faire qu'étape par étape. SN a élaboré une approche de remaniement afin d'aider au déroulement des projets de remaniement individuels.

Dans le présent article, nous explorerons cette approche de remaniement, qui a été élaborée par des équipes affectées à cette tâche et qui est maintenant généralement adoptée à l'échelle de SN. Nous commencerons par décrire le processus de remaniement, avant de mettre l'accent sur les produits les plus importants à livrer : le document relatif à l'analyse des activités (DAA), le document consultatif sur la méthodologie (DCM) et le document relatif à l'architecture logicielle (DAL). Dans la dernière section, nous réfléchirons aux expériences liées à l'approche de remaniement dans le cadre de projets de remaniement et nous examinerons les développements actuels et futurs.

Le présent document représente une version abrégée et légèrement mise à jour d'un article antérieur : Hofman et Leerintveld (2010).

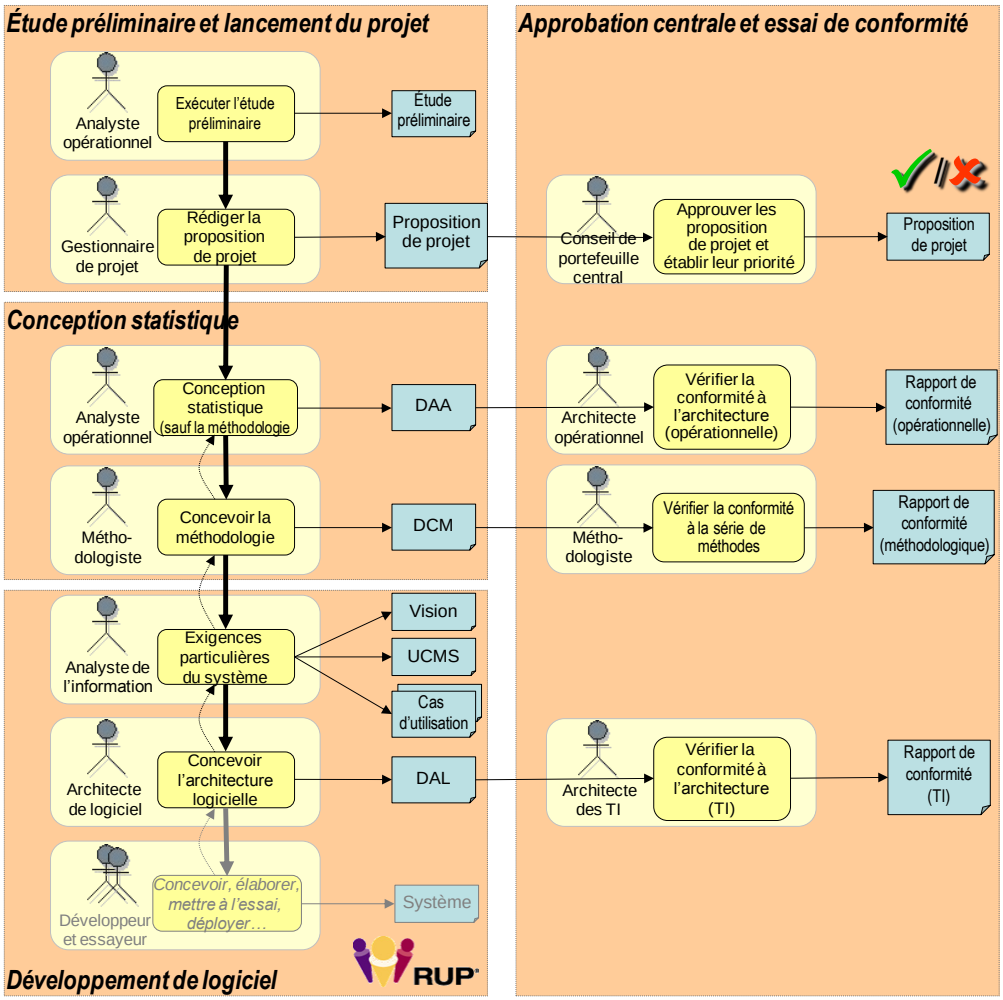
[1]Frank Hofman, Statistics Netherlands, Henri Faasdreef 312, 2492 JP La Haye, Pays-Bas, f.hofman@cbs.nl.
[2]Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas nécessairement la politique de Statistics Netherlands.

2. Processus de remaniement

Le processus de remaniement a été développé pour appuyer les projets de remaniement au moyen d'une méthode de travail uniforme. Le processus comporte un certain nombre d'activités connexes qui doivent être menées pendant un projet de remaniement moyen. Chaque activité est prise en charge par un ensemble de compétences et produit un document (ou d'autres éléments). Même si la séquence des activités peut laisser supposer un ordre séquentiel strict, les activités peuvent aussi être menées en parallèle ou de façon itérative.

La portée du processus de remaniement est large et, avant même qu'un projet de remaniement ait été lancé, le processus commence par une étude préliminaire, pour aller jusqu'à la production du nouveau système d'information. Le présent article est axé sur la première partie du processus jusqu'à la « conception statistique » et son lien avec le développement du système. SN a adopté certaines normes bien établies pour des parties du processus de remaniement : RUP (Rational Unified Process[3] pour le développement du système, TMap[4] pour la mise à l'essai et Prince2[5] pour la gestion de projet. Le présent article est axé sur les parties non uniformes du processus de remaniement.

Figure 2-1
Processus de remaniement



[3] http://en.wikipedia.org/wiki/IBM_Rational_Unified_Process
[4] http://eng.tmap.net/Home
[5] http://en.wikipedia.org/wiki/PRINCE2

La figure 2-1 décrit le processus de remaniement. Un survol rapide montre qu'avant d'amorcer un projet de remaniement proprement dit, la proposition de projet doit être rédigée et approuvée. Afin de remplir correctement une proposition de projet, il est souvent nécessaire de procéder à une étude préliminaire au sujet de la méthodologie, des produits statistiques, des outils, *etc*., en mettant l'accent sur les changements nécessaires et en définissant le déroulement du projet de remaniement proprement dit. Au cours de la conception statistique, les produits statistiques sont conçus, de même que le processus et la méthodologie servant à la transformation des données d'entrée en ces produits. Les processus statistiques ont généralement besoin de systèmes de soutien (logiciels) qui doivent être développés.

Certains documents qui sont produits au cours d'un projet de remaniement doivent être approuvés au niveau central ou doivent être vérifiés au chapitre de la conformité à l'architecture.

Dans la présente section, nous examinerons chaque sous-processus du processus de remaniement.

2.1 Étude préliminaire et lancement du projet

Pour chaque projet de remaniement, le gestionnaire de projet doit rédiger une proposition de projet décrivant notamment l'analyse de rentabilisation et les objectifs. Comme SN utilise Prince2 pour la gestion de projet, la proposition de projet comprend le mandat du projet, le résumé du projet et/ou le document de lancement du projet.

Une proposition de projet peut être précédée par une étude préliminaire si le gestionnaire de projet ou le comité directeur désigné (ses membres) estime qu'il est nécessaire de la faire. L'étude commence habituellement par une description de la situation, c'est-à-dire un bref énoncé du processus en vigueur, afin de faire ressortir les problèmes et/ou les changements souhaités. Elle peut aussi comprendre une analyse des changements externes dont on doit tenir compte. Au cours de l'étude, les options de rechange pour l'avenir sont examinées. Après avoir pesé les pours et les contres de chaque option, le comité directeur doit faire un choix, en vue de définir le déroulement du projet de remaniement proprement dit.

L'enquête est habituellement effectuée par un analyste opérationnel, mais d'autres spécialistes, comme un méthodologiste, un analyste de l'information et un architecte de logiciel, peuvent aussi intervenir au besoin. Une présentation définie pour le rapport d'étude préliminaire est en voie d'élaboration.

2.2 Conception statistique

Comme cette partie du processus est propre à un institut national de statistique, nous abordons la conception statistique de façon plus détaillée que les autres étapes. L'architecture opérationnelle de SN comporte cinq étapes pour la conception et le remaniement de données statistiques, les quatre premières étant englobées dans la première étape de la figure 2-1 :

1. Déterminer les besoins d'information statistique
2. Concevoir le produit statistique
3. Concevoir les sources de données
4. Concevoir le modèle de processus
5. Concevoir la méthodologie

En ce qui a trait à l'ensemble du processus de remaniement, ces cinq étapes ne doivent pas nécessairement être menées de façon séquentielle. Elles peuvent aussi se dérouler en parallèle ou de façon itérative. Par exemple, il peut sembler difficile ou coûteux d'obtenir les données d'entrée exactes au moment de la conception des sources de données, alors qu'un registre comportant des renseignements légèrement différents est facilement disponible. Dans un tel cas, un analyste opérationnel doit se pencher à nouveau sur la conception du produit statistique et même consulter à nouveau les clients pour évaluer si un produit légèrement différent réussira à combler leurs besoins.

Un analyste opérationnel est responsable de la rédaction du DAA. La rédaction du DCM est du ressort d'un méthodologiste.

Déterminer les besoins d'information statistique

Comme principe de base, le processus de conception est axé sur les produits. Nous commençons donc par déterminer les besoins de données statistiques. À cette étape, nous identifions les clients et leurs besoins.

Cette étape permet d'intégrer les clients internes et externes dans le diagramme de contexte pour les statistiques en voie d'élaboration. Elle est décrite dans le DAA.

Concevoir le produit statistique

Lorsque nous connaissons les besoins de nos clients, nous pouvons concevoir le produit statistique proprement dit. Nous déterminons le ou les tableaux à produire, la population, les variables, les niveaux d'agrégation, ainsi que les métadonnées sur la qualité (indicateurs et normes), de même que la fréquence de publication des statistiques.

Pour réduire les publications en double (ou presque) et maximiser la cohérence entre les produits statistiques, il est important d'être bien informé sur les produits existants et les produits intermédiaires sur le même thème statistique.

Généralement, les statistiques sont destinées à des clients externes; elles peuvent toutefois servir à de nombreux utilisateurs internes aussi. Dans ce cas, nous concevons aussi le ou les produits intermédiaires qui seront réutilisés par les utilisateurs internes à cette étape.

Le résultat de cette étape est une description des états stables du produit et, peut-être, des produits intermédiaires dans le DAA. Les états stables sont un élément clé de l'architecture de SN et comprennent des ensembles de données dont la qualité est garantie et qui peuvent être réutilisés par l'entremise d'un des services opérationnels génériques, le Centre de service de données.

Concevoir les sources de données

Lorsque nous savons ce que le processus statistique peut produire, nous pouvons nous pencher sur les méthodes permettant d'obtenir le produit souhaité. Cela signifie que nous devons concevoir les sources de données d'entrée nécessaires pour produire le produit. La description des sources de données d'entrée est comparable à celle du produit : métadonnées conceptuelles, métadonnées sur la qualité.

Encore plus que pendant la conception du produit statistique, il est important d'avoir un aperçu des données d'entrée et des produits intermédiaires disponibles au moment de la conception des sources de données. Afin de réduire le fardeau administratif pour les citoyens et les entreprises, de nouvelles enquêtes ou de nouveaux questionnaires peuvent être envisagés uniquement lorsqu'aucune des sources disponibles dans le SN ou d'autres organismes gouvernementaux ne sont pertinents pour le produit statistique souhaité.

Le résultat de cette étape est une description des états stables des données d'entrée dans le DAA.

Concevoir le modèle de processus

Dans le modèle de processus, nous décrivons les étapes (activités) devant être menées pour produire le produit statistique (données de sortie) à partir des sources de données (données entrée), en intégrant les services opérationnels génériques. À partir de chaque étape, nous décrivons l'objectif, les données d'entrée et les données de sortie. Les données d'entrée peuvent comprendre des renseignements auxiliaires, par exemple, le calcul des poids nécessaires pour « majorer » les résultats de l'enquête sur échantillon, afin qu'ils soient représentatifs de la population cible.

Le déroulement de toutes les étapes est aussi important que les étapes proprement dites, particulièrement lorsqu'il est complexe et qu'il comprend des ramifications et des boucles. Les critères utilisés pour les décisions, comme les critères d'interruption des bouches, sont souvent établis à partir des besoins en matière de qualité du produit statistique.

Dans le cas des processus plus complexes, nous décrivons non seulement le processus de production, mais aussi le processus de gestion, particulièrement lorsque les statistiques font partie d'un ensemble de processus statistiques. Nous énonçons les indicateurs nécessaires dans le processus de production pour gérer ce dernier avec exactitude.

Les résultats de cette étape sont un modèle de processus, ainsi que des produits intermédiaires (additionnels) dans le DAA.

Concevoir la méthodologie

La dernière étape de la conception de statistiques consiste à choisir la méthodologie appropriée pour le produit, selon la qualité requise des données d'entrée sélectionnées. En pratique, la méthodologie principale est souvent explorée avant que le modèle de processus soit conçu. Puis, au moment des itérations, on élabore la méthodologie et le modèle de processus.

Un autre principe de base consiste à utiliser uniquement des méthodes validées. Par conséquent, SN a élaboré la série de méthodes, des méthodes scientifiques, bien documentées et ayant fait leur preuve qui sont privilégiées dans un processus statistique particulier.

Le résultat de cette étape est une description de l'ensemble de la méthodologie décrite dans le DAA.

2.3 Élaboration du logiciel

Pour passer de la conception statistique au système d'information, SN combine le processus de conception statistique et le RUP, comme processus de développement de système. Dans le présent article, nous mettons l'accent sur la discipline des exigences et la partie de l'architecture logicielle de l'analyse et de la discipline de la conception.

La discipline des exigences est cruciale pour assurer une transition sans heurt de la conception statistique au développement du système. L'analyste de l'information entre les exigences du système dans le document de vision et élabore l'utilisation des systèmes dans l'enquête sur le modèle de cas d'utilisation. La description détaillée de la fonction du système est intégrée dans plusieurs cas d'utilisation.

Les architectures logicielles des projets individuels jouent un rôle crucial pour la gestion de l'environnement global des TI. Un principe directeur important est la réutilisation du logiciel existant à SN, avant l'achat de logiciel standard et le développement de logiciel sur mesure.

L'architecte de logiciel décrit l'architecture logicielle des projets dans un document relatif à l'architecture logicielle (DAL).

2.4 Approbation centralisée et vérification de la conformité

À l'intérieur du projet de remaniement, tous les documents sont passés en revue. Par ailleurs, certains documents doivent être passés en revue par un responsable central. En ce moment, il s'agit des suivants : proposition de projet, DAA, DCM et DAL.

Les propositions de projet de remaniement doivent être approuvées par un organisme central, le « conseil de portefeuille central », qui est constitué du directeur général adjoint, des directeurs des trois divisions statistiques, du directeur des TI et du directeur de la méthodologie. Le conseil vérifie la légitimité de l'analyse de rentabilisation et établit la priorité des projets. Une fois une proposition de projet de remaniement approuvée et les ressources nécessaires affectées, un projet de remaniement peut être lancé.

Une fois le DAA complet dans une certaine mesure, il est soumis aux architectes opérationnels pour examen. Ces derniers déterminent si le processus statistique proposé est conforme à l'architecture opérationnelle de SN et documentent leurs constatations dans un rapport de conformité à l'intention du comité directeur du projet de remaniement. Si on détermine que le DAA est conforme à l'architecture opérationnelle, le projet de remaniement se poursuivra. Autrement, le comité directeur peut décider de corriger le DAA ou de demander au conseil d'architecture

central la permission de s'écarter de l'architecture opérationnelle. Le conseil d'architecture central est constitué du vice-directeur de la division de la méthodologie et des gestionnaires des services de développement.

Tout comme le DAA, le DCM et le DAL sont vérifiés par la division de la méthodologie (intégrée) et les architectes des TI afin de déterminer s'ils sont conformes à la série de méthodes et à la partie des TI de l'architecture opérationnelle, respectivement.

3. Principaux documents de conception : DAA, DCM et DAL

Dans la section qui précède, nous avons montré à quelles étapes de la conception correspondent les documents. Dans la présente section, nous mettrons l'accent sur les documents proprement dits : le DAA, le DCM et le DAL.

3.1 Document relatif à l'analyse des activités (DAA)

Le DAA rend compte de la conception statistique, sauf la méthodologie (qui fait l'objet du DCM). Les principaux sujets du DAA sont les suivants :

- **Contexte**
  Le contexte décrit l'environnement d'un service statistique. Il comprend les utilisateurs externes et internes des données statistiques, ainsi que les fournisseurs des sources de données. Par ailleurs, il englobe les services opérationnels génériques qui sont utilisés dans le processus, comme le service de collecte des données et le centre de service de données.
- **Produits**
  Cette section décrit tous les états stables compris dans le processus. Elle fait état non seulement du produit, mais aussi de la source des données et des produits intermédiaires (aux niveaux micro et macro). Pour chaque état stable, les métadonnées conceptuelles et de qualité pertinente sont explorées, tout comme la population, les variables, la fréquence et la qualité. Pour les statistiques plus complexes, nous élaborons aussi un modèle objet opérationnel. Ce modèle fournit un aperçu des objets du « monde réel » (unités) visés par ces statistiques, ainsi que de leurs rapports. Par exemple, il montre comment un emploi est lié à la fois à une entreprise et à une personne. Il peut aussi être utilisé pour établir un lien entre les unités d'entrée et les unités statistiques, par exemple, le lien entre une entité juridique et une unité commerciale. Conformément au RUP, nous utilisons un diagramme de catégories UML, qui englobe les catégories statistiquement pertinentes et leurs principaux attributs.
- **Modèle de processus**
  Le modèle de processus fournit un aperçu de l'ensemble du processus et de ses liens avec les états stables. Généralement, l'ensemble du processus est réparti en sous-processus, d'un état stable au suivant. Le modèle de processus montre aussi comment les services opérationnels génériques sont utilisés dans le processus. Pour le fonctionnement exact de chaque étape du processus, il faut se reporter au DAA. Dans les processus plus complexes, le processus de production et le processus de gestion et/ou le processus de conception (ou des parties) sont compris dans le modèle. Celui-ci rend compte de l'information (de processus) nécessaire pour contrôler et rajuster le processus de production.
- **Conformité à l'architecture**
  Dans cette annexe, l'analyste opérationnel peut rendre compte des écarts par rapport à l'architecture opérationnelle. Même si ces écarts sont bien fondés, le DAA peut quand même être approuvé.

Il faut plusieurs mois pour rédiger un DAA pour un projet de remaniement moyen, et le document compte de 30 à 80 pages. Si une étude préliminaire est menée, le DAA nécessitera généralement moins de temps et d'effort.

Le conseil d'architecture central a récemment approuvé un nouveau modèle pour le DAA. Outre les sujets mentionnés précédemment, ce modèle englobe les exigences (ou les conditions préalables) du processus, la planification du processus et la stratégie de migration initiale. Dans une annexe, nous devons préciser la conformité à l'architecture de la conception. La conception est-elle conforme à l'architecture opérationnelle? Sinon, quelles sont les raisons de l'écart?

3.2 Document consultatif sur la méthodologie (DCM)

La méthodologie est abordée dans le DCM. La méthodologie comprise dans le DCM englobe l'ensemble du processus statistique, de la collecte des données à la diffusion du produit final, et définit le fonctionnement de chaque étape du processus. Les thèmes méthodologiques comprennent les suivants :

- Enquête sur échantillon
- Conception du questionnaire
- Correction pour tenir compte de la non-réponse
- Imputation
- Correction pour tenir compte des variations saisonnières
- Prévention de la divulgation

Le DCM énonce non seulement les méthodes choisies, mais indique aussi pourquoi elles l'ont été et comment elles sont appliquées. C'est l'établissement des paramètres.

Étant donné que le DCM a été adopté plus récemment, il n'existe pas encore de modèle ni de lignes directrices concernant son contenu. Toutefois, le développement futur du DCM est en marche.

3.3 Document relatif à l'architecture logicielle (DAL)

Conformément au RUP, nous utilisons le modèle de vision 4+1 de Kruchten[6] pour le DAL. Une attention spéciale est accordée à la réutilisation des outils existants, ainsi qu'à l'utilisation ou au développement de nouveaux services logiciels réutilisables. La rentabilité est considérée comme très importante et le développement de logiciels personnalisés doit être évité, chaque fois que l'un des outils COTS[7] ou des services réutilisables sont disponibles. Un DAL type complet comporte de 40 à 70 pages.

Comme le DAA, le DAL comprend aussi une annexe appelée « Conformité à l'architecture ». Un modèle de DAL est disponible.

4. Réflexions et développements (pour l'avenir)

Dans la présente section, nous évaluons le processus de remaniement actuel et nous mentionnons les nouveaux développements visant à l'améliorer ou à l'élargir.

4.1 Lien entre la conception statistique et le développement des TI

Le processus général de remaniement de la section 2 laisse supposer un lien sans heurt entre la conception statistique et le développement des TI. Toutefois, le RUP ne comporte pas de lien explicite entre le document de vision (ou les autres documents d'exigences) et la conception du processus. Le document de vision comporte plutôt une définition des intervenants et de leurs besoins, ainsi que des caractéristiques du système des TI. Nous relions de façon explicite les caractéristiques et les cas d'utilisation aux étapes du modèle de processus, en indiquant le soutien TI nécessaire (ou souhaité) pour exécuter chaque étape du processus. Ce faisant, nous pouvons établir un lien entre toutes les exigences fonctionnelles et le processus.

[6] http://www.ibm.com/developerworks/ wireless/library/wi-arch11/

[7] Commercial sur étagère : logiciel sous emballage facilement disponible ; seule une configuration est nécessaire pour sa mise en œuvre.

4.2 Développement itératif par rapport au modèle en cascade

Il est facile de voir qu'il existe une certaine tension entre la nature itérative[8] du RUP et l'ensemble du processus de remaniement, d'une part, et l'approbation séquentielle des documents (propositions de projet, DAA, DCM et DAL) par les responsables centraux, d'autre part. Par exemple, si les expériences relatives au logiciel au moment du développement montrent que le modèle de processus proposé mène à un mauvais rendement pour les grands ensembles de données, on aura tendance à accroître le rendement en utilisant du matériel plus puissant plutôt qu'en repensant le modèle de processus proprement dit, celui-ci ayant été approuvé officiellement.

Évidemment, les gestionnaires de projet et les comités directeurs sont réticents à soumettre à nouveau des documents approuvés précédemment, et les analystes opérationnels, analystes de l'information et architectes de logiciel hésitent par conséquent à leur conseiller de le faire. Que cette crainte soit justifiée ou non, elle est peu pertinente en pratique. Le simple fait qu'un processus d'approbation formel représente nécessairement un obstacle et soit perçu comme tel mène à des solutions potentiellement sous-optimales.

Si ces obstacles sont tels qu'il est impossible de revenir en arrière (intentionnellement ou psychologiquement), le processus de développement ressemblera au modèle appelé en cascade[9], qui a été largement critiqué. L'équipe de remaniement multidisciplinaire vise à alléger les problèmes liés à ce modèle, tout comme l'étude préliminaire. Toutefois, le retour en arrière et la réévaluation des décisions sont difficiles et nécessitent une connaissance du processus statistique, ainsi que du développement de logiciel.

4.3 Étapes de processus normalisées

Un projet de recherche concernant les possibilités des étapes de processus normalisées (EPN) a été entrepris. Le concept de base des EPN est de concevoir et d'élaborer des éléments préfabriqués qui peuvent être utilisés au moment de concevoir des statistiques. Ces éléments englobent une ou plusieurs étapes du processus statistique et fournissent une solution méthodologique uniforme. L'analyste opérationnel qui remanie les statistiques combine idéalement plusieurs de ces éléments et les configurent pour qu'ils conviennent aux statistiques particulières. L'objectif des EPN est de réduire le temps et les coûts de conception et de remaniement des processus statistiques, ainsi que de leur maintien. Pour plus de renseignements, voir Renssen et coll. (2009) et Hofman, Camstra et Renssen (2011).

4.4 Différenciation de l'approche de remaniement

À Statistics Netherlands, la différenciation de l'approche de remaniement pour les différents types de projets représente un problème majeur. Deux questions se posent : 1) « Comment faire une distinction entre les différents types de projet de remaniement? » et 2) « Comment adapter l'approche générale aux projets de remaniement individuels? »

Le conseil de portefeuille central élabore actuellement une classification des projets de remaniement. L'idée est de faire une distinction entre trois types de projets, selon leur taille et l'importance des statistiques. Néanmoins, les critères exacts de classification sont toujours à l'étude. À cet égard, on tente aussi de déterminer si le développement décentralisé, c'est-à-dire les projets utilisant des outils standard dont l'effectif est recruté localement, devrait utiliser cette approche plus formelle.

On n'a pas encore répondu à la deuxième question. Certaines personnes plaident pour la réduction du nombre de documents, tandis que d'autres réduiraient plutôt le niveau de détails en conservant tous les documents.

[8] http://en.wikipedia.org/wiki/ Iterative_and_incremental_development
[9] http://en.wikipedia.org/wiki/ Waterfall_model

4.5 Assurance de la qualité du processus

Dans le cadre du système d'assurance de la qualité, qui est obligatoire pour (certains) organismes gouvernementaux, la sécurité de l'information de chaque processus statistique doit être décrite, et une évaluation des risques indiquant les mesures préventives possibles doit être effectuée.

Même si la description du processus pour l'assurance du processus est assez similaire à celle de la conception du processus, elles étaient autrefois différentes. Cela est dû en partie au fait que des outils différents étaient utilisés pour la conception de processus et l'assurance de la qualité. SN est maintenant passé à un seul outil (Mavin) pour ces deux fins et a entrepris un projet, en vue d'aligner davantage les deux descriptions de processus. L'objectif consiste à concevoir un processus une fois seulement et à élargir sa conception sur la base de l'information nécessaire pour la gestion du risque et pour le maintien à jour de la description du processus, lorsque des révisions mineures se produisent. Ainsi, nous obtenons des descriptions de processus détaillées et à jour pour tous les processus statistiques, pendant leur cycle de vie.

Bibliographie

Braaksma, B. (2009), « Redesigning a Statistical Institute: The Dutch case », *Proceedings of MSP2009, workshop on Modernisation of Statistics Production 2009*.

Hofman, F. et B. Leerintveld (2011), « Remaniement à Statistics Netherlands », *Recueil du Symposium 2011 de Statistique Canada*, Ottawa, Canada.

Hofman, F., Camstra, A. et R. Renssen (2011), « Standardisation of Processes », *Proceedings of World Statistics Congress (ISI2011)*, Dublin, Irelande.

Renssen, R., Morren, M., Camstra, A. et T. Gelsema (2009), « Standard processes », document non publié, Statistics Netherlands, La Haye, Pay-Bas.

Van der Veen, G. (2007), « Changing Statistics Netherlands: driving forces for changing Dutch statistics », document présenté au Seminar on the Evolution of National Statistical Systems, New York, États-Unis.

Ypma, W.F.H. et K. Zeelenberg (2007), « Counting on Statistics; Statistics Netherlands' Modernization Program », document présenté au Seminar on Increasing the Efficiency and Productivity of Statistical Offices at the plenary session of the Conference of European Statisticians, Genève, Suisse.

Vers la normalisation à Statistics New Zealand

John Lopdell et Gary Dunnet[1]

Résumé

Statistics New Zealand emploie des concepts normalisés pour la conception de haut niveau de ses enquêtes et applique un certain nombre de processus et de méthodes communs, mais non normalisés. Toutefois, l'infrastructure pour appuyer ces méthodes et processus est unique à chaque produit. Notre « carnet de route de la normalisation » décrit une trajectoire vers une normalisation plus poussée des méthodes, des processus, de la gestion des données et de la technologie.

Plusieurs projets ont été entrepris en vue de faire progresser la normalisation à Statistics New Zealand. L'un des projets clés est l'élaboration de « plateformes » fondées sur une infrastructure commune et sur des « groupes » d'enquêtes. Cinq plateformes sont élaborées pour soutenir les diverses étapes du modèle générique des processus opérationnels (gBPM pour generic business process model), à savoir une plateforme des enquêtes ménages pour la production des statistiques sociales transversales, une plateforme des statistiques commerciales et économiques (BESt pour Business and Economic Statistics) pour les statistiques microéconomiques, une plateforme basée sur la gestion des relations avec la clientèle pour la collecte des données, une plateforme pour les statistiques macroéconomiques et une plateforme fondée sur les statistiques de l'OCDE pour la diffusion des données. En plus des gains d'efficacité, ces plateformes augmenteront la cohérence de nos produits statistiques et faciliteront l'accroissement de l'usage et de l'analyse des données.

La communication décrira l'approche des « plateformes » qui est élaborée à Statistics New Zealand, les progrès d'intégration des opérations de collecte dans les plateformes réalisés jusqu'à présent et les futurs travaux, en se concentrant principalement sur la plateforme des enquêtes ménages et la plateforme des statistiques commerciales et économiques. La communication présentera 1) les principes appliqués pour élaborer les plateformes, 2) l'approche utilisée pour mettre en œuvre les plateformes, 3) les exigences et les fonctionnalités en ce qui concerne les méthodes et les outils intégrés dans ces plateformes (par exemple vérification et imputation des données statistiques) et 4) les phases effectives d'élaboration.

[1]John Lopdell et Gary Dunnet, Statistics New Zealand, Nouvelle-Zélande.

Développement d'une base de sondage commune pour les enquêtes-ménages à Statistique Canada

Larry MacNabb, Martin St-Pierre et Marco Grenier[1]

Résumé

Afin de continuer d'être conforme aux normes élevées de qualité visées par ses programmes statistiques, l'initiative de l'Architecture opérationnelle du Bureau (AOB), mise en œuvre en 2009, répond à certains défis que doit relever Statistique Canada. L'AOB vise principalement à améliorer l'efficacité organisationnelle et la robustesse des systèmes et des processus, tout en accélérant l'exécution des nouveaux projets et programmes. Une des recommandations de l'AOB est la création d'une base de sondage commune pour les enquêtes-ménages et la normalisation des processus nécessaires à sa création et à sa maintenance. Cette base commune sera centrée autour du Registre des adresses (RA), une base de logements qui est déjà utilisée pour le recensement et quelques enquêtes-ménages. Au RA, on associera de l'information de contact et de l'information sociodémographique provenant du recensement et de diverses sources administratives. Pour ce faire, des outils de normalisation et de validation seront développés afin de traiter les différents fichiers de données. L'utilisation de cette base commune mènera à des gains d'efficacité, entre autres, par l'élimination du dédoublement de tâches entre les différentes enquêtes, la réduction des coûts de collecte et par l'élaboration de plans d'échantillonnage plus efficaces. De plus, l'excellente couverture de cette base, jumelée à une stratégie de collecte plurimodale, offrira des solutions de rechange aux enquêtes à composition aléatoire de numéros de téléphone. Dans cette communication, nous présentons d'abord l'historique et la justification du projet. Ensuite, nous traitons du développement des différentes fonctionnalités de la base commune, notamment des outils de normalisation et des étapes de traitement des sources de données intégrées à la base, et nous donnons un aperçu des utilisations potentielles de cette base pour les enquêtes-ménages.

Mots clés : Architecture opérationnelle du Bureau ; base de sondage ; registre des adresses ; enquêtes-ménages.

1. Architecture opérationnelle du Bureau (AOB) à Statistique Canada

En avril 2009, un groupe de travail de Statistique Canada a produit un rapport faisant état de la possibilité de réaliser des économies de 5 % dans le budget de fonctionnement de Statistique Canada (StatCan), grâce à une architecture opérationnelle améliorée. À la suite de ce rapport, StatCan lançait officiellement l'initiative de l'Architecture opérationnelle du Bureau. L'initiative, dirigée par un groupe de travail de cadres supérieurs et relevant directement du Comité des politiques de SC, comportait trois objectifs principaux :

1. réaliser une économie exploitable de 5 % sur les coûts de fonctionnement permanents d'ici cinq ans;
2. élaborer un ensemble de systèmes et de processus réduits et sans chevauchement faisant l'objet d'une mise à jour et d'une documentation appropriées;
3. accroître la rapidité de réaction du point de vue de l'exécution des nouveaux programmes statistiques.

Pour atteindre ces objectifs, plusieurs principes clés de l'AOB ont été adoptés au Bureau. Les principaux principes mis de l'avant étaient les suivants :

- décisions concernant les processus devant être optimisés au niveau intégré plutôt que localement;
- utilisation optimisée et, dans nombre de cas, obligatoire des services intégrés;
- dans la mesure du possible, réutilisation des processus opérationnels et des systèmes informatiques existants;
- promotion de l'utilisation d'applications et d'outils opérationnels intégrés; et
- processus élaborés dans l'intention de produire et d'utiliser des métadonnées pour documenter et diriger (contrôler) le processus.

[1]Larry MacNabb, Larry.MacNabb@statcan.gc.ca; Martin St-Pierre, Martin.St-Pierre@statcan.gc.ca; Marco Grenier, Marco.Grenier@statcan.gc.ca, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

2. Initiative de la base de sondage commune

2.1 Principes

Dans l'environnement de collecte actuel, il est de plus en plus difficile de concevoir des stratégies d'échantillonnage permettant non seulement d'assurer une représentation appropriée dans les enquêtes auprès des ménages, mais fournissant aussi des façons efficaces de communiquer avec les répondants aux enquêtes. À cette fin, un ensemble de méthodologies sont requises pour répondre à la fois à la nécessité de plans d'échantillonnage perfectionnés et de vecteurs multiples pour le contact avec les répondants. Link et coll. (2009) croient que l'échantillonnage fondé sur les adresses (EFA) est la base d'échantillonnage à partir de laquelle un ensemble de méthodologies peut être élaboré, fournissant une base d'échantillonnage stable, une source riche de caractéristiques et de données géographiques pour faciliter l'établissement de plans d'échantillonnage perfectionnés, et la possibilité d'utiliser plusieurs modes pour communiquer avec les ménages et mener des enquêtes auprès d'eux.

Statistique Canada a lancé le projet de base de sondage commune, dont les objectifs généraux sont de fournir un moyen de relever les défis liés à la conception de stratégies d'échantillonnage non biaisées, tout en servant de base pour appuyer plusieurs modes de collecte. L'objectif principal du projet était d'élaborer une base de sondage fondée sur les adresses devant être utilisée pour le recensement et la plupart des enquêtes auprès des ménages à Statistique Canada. Une part importante de l'infrastructure nécessaire à cette fin existait déjà à l'appui du Registre des adresses (RA), qui devait servir de fondement pour l'infrastructure de base de sondage commune restante.

Parmi les autres objectifs liés au projet figurait la centralisation de toutes les activités liées à la base de sondage dans une unité fonctionnelle. Cela comprend l'uniformisation des fichiers administratifs, l'entreposage des fichiers, l'élaboration d'une base de sondage et la mise à jour et la coordination des opérations de listage sur le terrain, aux fins de la gestion de la qualité de la base de sondage commune. Dans la mesure du possible, le projet visait à assurer l'harmonisation des processus et l'utilisation d'outils communs, en rapport avec la normalisation des adresses et le traitement des numéros de téléphone extraits de sources administratives.

2.2 Description et rôles du Registre des adresses

Le Registre des adresses (RA) est constitué d'une liste d'adresses couvrant la majorité des logements privés et collectifs au Canada. Avant le Recensement de 2011, le RA contenait environ 15 millions d'adresses. Le RA maintient les adresses selon deux concepts : l'adresse de l'emplacement et l'adresse postale. L'adresse de l'emplacement est utilisée pour déterminer l'emplacement du logement sur le terrain. Les adresses des emplacements peuvent être regroupées en deux grandes catégories, soit les adresses municipales et les adresses non municipales, selon la présence ou l'absence d'un numéro municipal unique figurant sur le logement. L'adresse postale est l'adresse utilisée par la Société canadienne des postes (SCP) pour distribuer le courrier à cette adresse et elle peut être représentée différemment de l'adresse de l'emplacement d'un même logement au Canada. Il est nécessaire pour le RA de détenir une adresse postale associée à l'adresse de l'emplacement d'un logement pour pouvoir effectuer l'envoi postal des questionnaires du recensement. On doit noter cependant que lorsque l'adresse postale est de type non municipal (par exemple, une case postale), elle n'est pas disponible sur le RA. Pour ce type de logement, seule l'adresse de l'emplacement est disponible. Plus d'informations sur les différents concepts d'adresses (emplacement, municipale ou postale) sont données dans un rapport interne de McClean et Charland (2011). À la suite du Recensement de 2006, le RA a pu déterminer une adresse municipale pour plus de 95 % des logements au Canada et une adresse postale pour près de 87 % des logements au Canada. La proportion d'adresses municipales et postales est de beaucoup supérieure dans les régions urbaines par rapport aux régions rurales. En fait, dans les régions urbaines, les adresses sont presque toutes de type municipal avec adresse postale associée.

Chaque logement couvert par le RA est également lié, par l'entremise de son adresse, à la hiérarchie des unités géographiques statistiques ou administratives. Statistique Canada gère, de concert avec Élections Canada, un fichier du réseau routier national appelé Base nationale de données géographiques (BNDG). Statistique Canada a conçu un logiciel qui relie ou « géocode » les adresses aux côtés d'îlot figurant dans la BNDG. Les logements géocodés à un côté d'îlot sont par extension codés à un îlot particulier. Les îlots eux-mêmes sont regroupés à des niveaux géographiques supérieurs tels que les aires de diffusion (AD), les secteurs de recensement (SR) et les subdivisions de

recensement (SDR). Ces entités font partie de la hiérarchie des unités géographiques statistiques utilisée par Statistique Canada aux fins de la diffusion des données du recensement.

Le RA est mis à jour trimestriellement à l'aide de deux méthodes : le traitement des données de sources administratives et la vérification sur le terrain. La mise à jour du RA à partir de sources administratives vise à repérer les nouvelles adresses résidentielles. Les adresses obtenues de ces sources sont appariées trimestriellement au RA existant. Les sources administratives utilisées actuellement sont les suivantes :

- Fichiers de facturation des compagnies de téléphone;
- Fichiers d'annuaires téléphoniques;
- Fichiers d'impôt de l'Agence du revenu du Canada;
- Relevés des mises en chantier et des achèvements de la Société canadienne d'hypothèques et de logement;
- Points de remise de la Société canadienne des postes;
- Diverses sources municipales.

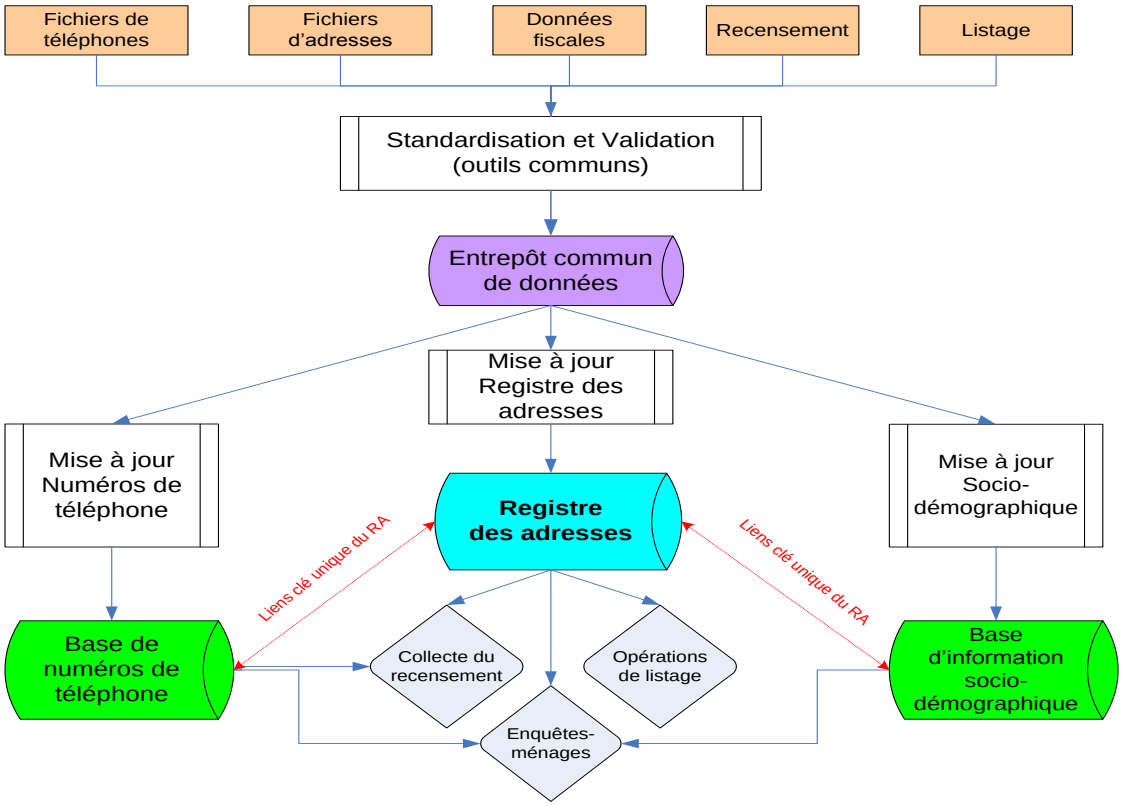
On a également recours à une activité de vérification sur le terrain, appelée listage, pour mettre à jour le RA. On remet au personnel sur le terrain des listes d'adresses pour un petit secteur (unité de listage) et une carte montrant les îlots de listage. Le personnel sur le terrain vérifie ensuite la liste des adresses, ajoute et/ou supprime des adresses au besoin, et vérifie les numéros d'îlot. Le listage est une activité permanente et un ensemble d'unités de listage est fourni au personnel sur le terrain chaque trimestre.

Jusqu'à maintenant, le rôle principal du RA est de fournir une base liste d'adresses aux fins du Recensement de la population du Canada qui a lieu tous les cinq ans. Depuis 2006, la méthode de collecte utilisée pour le recensement consiste à envoyer des questionnaires par la poste à un pourcentage élevé de logements. Le RA a servi à créer la liste d'envoi par la poste. Pour le Recensement de 2011, on a utilisé une méthode d'envoi par la poste dans des secteurs couvrant environ 80 % des logements au Canada. Le RA est également utilisé comme intrant à la base aréolaire de l'Enquête sur la population active (EPA). Cette base, aussi utilisée par d'autres enquêtes-ménages, est constituée d'une liste de grappes représentant de petites régions géographiques. Pour la collecte de l'EPA et autres enquêtes, le RA fournit la liste initiale de logements pour la majorité des grappes sélectionnées dans les échantillons. Par contre, un listage complet des logements est nécessaire dans d'autres grappes quelques mois avant la collecte. D'ailleurs, cette activité de listage sert également à mettre à jour le RA. Finalement, le RA est utilisé par l'Enquête canadienne sur les mesures de la santé (ECMS) afin de compléter sa base de sondage de logements.

Avec la nouvelle initiative de la base commune pour les enquêtes-ménages, le rôle du RA s'étendra considérablement au cours des prochaines années. Bien que son rôle principal demeurera de fournir une liste d'adresses pour l'envoi par la poste des questionnaires du recensement, le RA deviendra une base de sondage commune pour la majorité des enquêtes-ménages. Les enquêtes-ménages, notamment les enquêtes à composition aléatoire de numéros de téléphone, pourront recourir à une méthodologie d'échantillonnage fondée sur les adresses permettant d'améliorer la couverture de ces enquêtes. Afin de permettre à certaines enquêtes de continuer à utiliser principalement le téléphone comme mode de collecte, il sera primordial que le RA soit relié de façon permanente aux numéros de téléphone provenant des sources administratives. Également, afin d'accroître l'efficacité des plans d'échantillonnage des enquêtes-ménages utilisant le RA, on associera aux logements du RA de l'information sociodémographique provenant du recensement et des fichiers de données fiscales. Sous la nouvelle initiative, le RA continuera à être mis à jour sur une base trimestrielle et il devra fournir à la même fréquence une liste représentant l'univers des logements à partir de laquelle les méthodologistes responsables des enquêtes-ménages pourront sélectionner les échantillons. Finalement, le RA deviendra la source initiale et finale de logements pour la majorité des grappes de la base aréolaire de l'EPA, permettant de réduire voire d'éliminer les activités de listage pour les enquêtes utilisant cette base.

La figure 2.2-1 donne un aperçu général du processus trimestriel de mise à jour du RA, y compris les nouvelles fonctionnalités de la base de sondage commune. Les différentes sources de données passeront dans un module commun de normalisation et de validation des données, puis les fichiers normalisés seront conservés dans un entrepôt commun de données (ECD). L'ECD sera la source de trois processus distincts. Premièrement, au centre de la figure, on retrouve le processus de mise à jour du RA qui ressemblera au processus actuellement en place. Ensuite, deux nouveaux processus seront conçus afin de créer une base liste de numéros de téléphone ainsi qu'une base d'information sociodémographique. Ces deux bases seront reliées au RA à l'aide de la clé unique du RA.

Processus trimestriel de mise à jour de la base de sondage commune



3. Fonctionnalités de la base de sondage commune

3.1 Processus et outils communs

Comme démontré à la figure précédente, tous les fichiers de données servant à mettre à jour la base commune, dont le RA, devront passer dans un module commun de normalisation et de validation des données. Entre autres, on y fera la normalisation des adresses, la normalisation des noms et la validation des numéros de téléphone. Certains de ces processus, ou outils communs, sont développés à partir de procédures existantes à Statistique Canada, d'autres requièrent l'élaboration de nouvelles procédures. De plus, afin de permettre l'appariement des fichiers entre eux ou avec le RA, un processus commun de couplage des adresses devra être mis en place. Cette sous-section donne une description de quatre outils communs qui serviront à la construction de la base commune.

3.1.1 Normalisation des adresses

Les adresses provenant de différentes sources de données arrivent dans un format propre à chaque source. Afin d'uniformiser l'information sur l'adresse entre les sources et d'améliorer par la suite l'appariement par adresse entre ces sources, il est nécessaire de procéder à une normalisation des adresses, et ce, à l'aide d'un processus commun. Un processus de normalisation des adresses était déjà en place pour procéder au traitement des données de sources administratives servant à la mise à jour du RA. Par conséquent, il suffit tout simplement de revoir les étapes de traitement de ce processus et d'apporter des changements et des améliorations afin de répondre aux nouveaux besoins de la base commune. De plus, il faut optimiser le processus afin de le rendre générique et plus simple à utiliser.

L'outil de normalisation des adresses consiste à décomposer l'adresse provenant des fichiers sources en plusieurs champs comme la municipalité, le nom de la rue, le numéro d'immeuble, le numéro d'appartement, le type de rue,

etc. Chacun de ces champs de l'adresse est ensuite normalisé selon un ensemble de valeurs normalisées. Également, cet outil lie les adresses à la base de la Société canadienne des postes (SCP) afin d'obtenir l'adresse postale correspondante (y compris le code postal) dont le format peut être différent de celui de l'adresse originale. Étant donné que la base de la SCP est mise à jour mensuellement, il est important que les adresses provenant des différents fichiers administratifs soient liées à la même version de la base. Par conséquent, tous les fichiers administratifs sont liés à nouveau à la base de la SCP à chaque trimestre et cela même si le fichier administratif n'a pas eu de mise à jour depuis le dernier trimestre. Finalement, comme indiqué à la section 2.2, les adresses sont ensuite liées à la BNDG, ce qui permet de leur associer un côté d'îlot et par extension un îlot particulier. Ainsi, on peut lier les adresses à des niveaux géographiques supérieurs tels que les AD, les SR et les SDR.

3.1.2 Couplage des adresses

Afin de relier les adresses sur les fichiers de données administratives aux adresses sur le RA, un processus de couplage des adresses est absolument nécessaire. Un processus du genre existe déjà et il a été conçu pour permettre d'identifier des adresses sur les sources de données administratives qui ne sont pas sur le RA (processus d'identification de la croissance). Ainsi, un module commun de couplage d'adresses pourra être conçu à partir du processus actuel, et nécessitera seulement quelques améliorations mineures. L'objectif est de rendre ce processus générique et accessible aux utilisateurs désirant faire le couplage d'adresses provenant de n'importe quel ensemble de fichiers contenant des adresses normalisées.

3.1.3 Validation des numéros de téléphone

Un nouvel outil est élaboré afin de valider les numéros de téléphone provenant des sources administratives de numéros de téléphone. L'objectif principal de l'outil est d'identifier les numéros de téléphone qui sont valides et qui devraient donc être conservés sur la nouvelle base de numéros de téléphone. Un numéro est considéré valide si l'indicatif régional et le préfixe (CRP = 6 premiers caractères du numéro) sont valides au Canada. La source utilisée pour déterminer la liste des CRP valides est le fichier NPANXX maintenu par l'Administrateur de la numérotation canadienne (ANC). Un autre objectif de cet outil est d'associer un indicatif régional aux numéros pour lesquels il serait manquant ou invalide. C'est l'adresse (province ou code postal) qui est utilisée pour déterminer l'indicatif régional à associer au numéro, mais ce n'est pas toujours possible d'associer un indicatif régional étant donné qu'il y a parfois plus d'une valeur possible. Finalement, l'outil associe de l'information complémentaire aux numéros de téléphone. Par exemple, à l'aide de différents fichiers externes, on tente d'identifier si le numéro de téléphone est associé à une ligne terrestre ou à un cellulaire.

3.1.4 Normalisation des noms

Plusieurs des fichiers de données administratives utilisés pour la création de la base de sondage commune contiennent des noms de personnes. Afin d'accroître le taux de succès du couplage de données entre ces différentes sources administratives, il est avantageux de normaliser les noms provenant de ces sources et cela selon un processus identique pour toutes les sources. Ainsi, un module de normalisation des noms sera conçu et utilisé lors du traitement des sources de données. Quelques processus de normalisation des noms existent actuellement à Statistique Canada. Par conséquent, il suffira de choisir et d'implanter un de ces processus au traitement des fichiers utilisés pour la base de sondage commune. Cependant, quelques modifications et améliorations seront nécessaires afin de pouvoir traiter correctement toutes les sources de données administratives.

3.2 Base de numéros de téléphone

Une composante très importante de la nouvelle base de sondage commune est la création et la mise à jour d'une base de numéros de téléphone qui combinera l'information de tous les fichiers de données administratives contenant des numéros de téléphone, et qui sera reliée au RA. Il y a plusieurs objectifs visés par la création de cette base. D'abord, il faut s'assurer de relier les numéros de téléphone de toutes les sources au RA, afin que les enquêtes sélectionnant leurs échantillons à partir du RA puissent avoir accès à des numéros de téléphone pour contacter les ménages vivant dans les unités sélectionnées. Ce ne sont pas toutes les sources de numéros de téléphone qui étaient reliées au RA auparavant. De plus, l'objectif ne se limite pas à associer des numéros de téléphone aux adresses du RA, il faut également tenter d'associer les meilleurs numéros possible et plus qu'un numéro par adresse s'il y a lieu. Cela

permettra de générer des efficacités lors de la collecte. Un autre objectif de cette base est la création d'une baseliste indépendante et exhaustive de numéros de téléphone qui pourra être utilisée directement comme base de sondage, notamment dans le cadre d'une méthodologie de base duale. Dans certaines régions rurales, il est difficile d'associer des numéros de téléphone aux adresses sur le RA étant donné que l'adresse postale de type non municipal (casier postal ou route rurale) est souvent utilisée sur les sources de numéros de téléphone. La création de cette base liste de numéros de téléphone permettra de couvrir les logements de ces régions rurales et offrira aux enquêtes par téléphone une stratégie alternative d'échantillonnage et de collecte dans ces régions. Finalement, cette base associera des indicateurs de qualité et d'autres indicateurs à chacun des numéros de téléphone. Par exemple, la date de mise à jour, la liste des sources où le numéro est présent ou le type de service associé au numéro (ligne terrestre ou cellulaire).

Plusieurs sources de numéros de téléphone seront utilisées pour créer et mettre à jour la base de numéros de téléphone. Parmi ces sources, on retrouve le fichier mensuel Info-Direct qui contient les numéros de téléphone des annuaires téléphoniques des principaux fournisseurs de services téléphoniques au Canada (lignes terrestres seulement). Aussi, nous utiliserons les numéros de téléphone sur les fichiers de facturation de plusieurs fournisseurs de services téléphoniques qui sont habituellement disponibles sur une base trimestrielle. Puis, nous inclurons les numéros de téléphone fournis sur les questionnaires courts du recensement et ceux sur différents fichiers de données fiscales. Dans le futur, si nous obtenons accès à d'autres sources de numéros de téléphone, nous les inclurons dans la mise à jour de la base. Il est également prévu d'incorporer à cette base les numéros de téléphone provenant des répondants aux enquêtes-ménages afin, entre autres, d'éviter de contacter ces répondants à nouveau lors de prochaines enquêtes réduisant ainsi leur fardeau de réponse.

La base de numéros de téléphone contiendra seulement un enregistrement par numéro de téléphone et conservera, autant que possible, la meilleure adresse pour chacun de ces numéros. Voici un aperçu des étapes de création de cette base. D'abord, comme indiqué au diagramme de la Section 2.2, toutes les sources de numéros de téléphone devront avoir passé dans le module commun de normalisation et de validation, puis être ajoutées à l'ECD. La première étape consiste à lire les fichiers dans l'ECD et à exclure les numéros de téléphone invalides ou non résidentiels. Ensuite, à l'intérieur de chaque source, on procède à l'élimination des doublons afin de conserver seulement un enregistrement par numéro de téléphone. Si un numéro est associé à plus d'une adresse sur une même source, la meilleure adresse selon des critères de qualité de l'adresse est conservée. Lorsqu'il est impossible de choisir la meilleure adresse, seul le numéro de téléphone est conservé à cette étape (l'adresse est supprimée). Puis, on combine toutes les sources en appariant par numéro de téléphone. Lorsqu'un numéro de téléphone apparaît sur plus d'une source, un processus d'élimination des doublons a lieu une fois de plus afin de choisir la meilleure adresse pour chaque numéro. En plus des critères de qualité de l'adresse, d'autres critères de qualité sont utilisés, tels que la fiabilité de la source et la date de référence de la source, afin de déterminer la meilleure adresse associée au numéro. Après cette étape, la base liste est complète et contient un enregistrement par numéro de téléphone. Un point important à souligner est que dans le processus de sélection de la meilleure adresse, on pourra conserver pour un même numéro, la meilleure adresse municipale et la meilleure adresse non municipale (= adresse postale). Finalement, les numéros de téléphone sur la base sont reliés au RA. Ce lien est fait automatiquement étant donné que les sources de numéros de téléphone auront déjà été couplées au RA lors de l'étape de mise à jour trimestrielle du RA. Une fois les liens établis, la prochaine étape consiste à déterminer pour chaque adresse du RA, le ou les « meilleur(s) » numéro(s) de téléphone à utiliser au cours de la collecte. Ici, il s'agit de classer les numéros selon un ordre de priorité (1er numéro à appeler, 2^{e} numéro à appeler, *etc*.). Cet ordre de priorité tient compte de la fiabilité et de la date de référence des sources. Aussi, il peut tenir compte du type de service téléphonique, notamment en donnant priorité aux numéros de lignes terrestres par rapport aux numéros de cellulaires. Tout ce processus de création de la base de numéros de téléphone et de l'association au RA aura lieu sur une base trimestrielle. Chaque trimestre, la base contiendra seulement les numéros de téléphone se trouvant sur les versions les plus récentes des sources de numéros de téléphone. Elle ne conservera pas les numéros de téléphone n'apparaissant plus sur aucune des sources.

3.3 Base d'information sociodémographique

La base de sondage commune sera composée d'une autre base de données contenant de l'information sociodémographique sur les ménages. Ainsi, un nouveau processus doit être élaboré afin de créer cette base et d'associer l'information sociodémographique de celle-ci aux logements sur le RA. Cette base sera créée pour satisfaire à deux besoins principaux. Premièrement, l'information sur cette base servira en tant qu'information auxiliaire pour les enquêtes-ménages. Elle pourra être utilisée à la phase du plan d'échantillonnage pour servir à la création de grappes de logements, à la création de strates de logements ou encore pour déterminer la répartition de

l'échantillon. L'information sur la base pourra aussi servir à la phase d'estimation des enquêtes afin d'améliorer les ajustements faits pour la non-réponse ou pour procéder à l'imputation de certaines variables (par exemple, le revenu du ménage). Le deuxième besoin qui sera satisfait par la création de cette base est la centralisation du traitement et du couplage des fichiers contenant de l'information sociodémographique, en tirant avantage des outils communs décrits à la section 3.1. Au cours des dernières années, plusieurs enquêtes ou programmes utilisaient et traitaient séparément ces fichiers menant à une duplication des efforts et ainsi, à une perte d'efficacité à l'échelle de l'organisme. L'inclusion d'une base d'information sociodémographique dans la base de sondage commune vient donc répondre aux objectifs de l'AOB.

À Statistique Canada, quelques sources de données sociodémographiques sont disponibles. La source principale est le recensement quinquennal. Le questionnaire court du recensement contient de l'information sur les membres des ménages telle que l'adresse de résidence, le nom, la date de naissance, le sexe, la ou les langues parlées et le lien entre les personnes du ménage. Le dernier recensement a eu lieu en mai 2011 et les données sont disponibles aux utilisateurs depuis l'automne 2011. Pour la première fois en 2011, le contenu détaillé du questionnaire long du recensement a été recueilli à l'aide de l'Enquête nationale auprès des ménages (ENM) et a été administré à environ 1/3 des ménages sur une base volontaire. L'information sociodémographique détaillée de l'ENM est aussi une source utile d'information pour la création de la base commune. Un autre type de source de données sociodémographiques est fourni par les données fiscales des Canadiens, notamment à partir des déclarations annuelles de revenu connues sous le nom de fichiers T1. Ces fichiers contiennent de l'information sur les déclarants telle que l'adresse de résidence, le nom, la date de naissance, le sexe, l'état matrimonial, ainsi que l'information sur le revenu. À partir des fichiers T1 et d'autres sources de données administratives, le Fichier des familles T1 (T1FF) est créé à l'interne et est une version améliorée et plus complète du T1. En plus des déclarants, le T1FF contient de l'information sur les conjoints et enfants non déclarants et relie les personnes selon le concept de famille de recensement. Le T1FF fournit la composition des ménages d'une façon plus complète et précise que le T1. Cependant, le T1FF est disponible cinq à six mois plus tard.

À partir de ces sources de données, la base d'information sociodémographique sera créée et contiendra de l'information à deux niveaux principalement : au niveau des logements et au niveau des aires de diffusion (AD) du recensement. Premièrement, à partir des données du questionnaire court du recensement et du T1FF, de l'information au niveau des logements sera conservée dans un fichier et cela pour chaque logement (= adresse) sur le RA. Le fichier au niveau du logement contiendra de l'information sur le ménage (taille, type, revenu, *etc*.) et de l'information sur les caractéristiques des personnes vivant dans le logement (date de naissance, sexe, langue, revenu, *etc*.). La première version de ce fichier sera créée en utilisant les données du Recensement de 2011 auxquelles on ajoutera l'information de revenu provenant du T1FF 2010. Deux options sont disponibles pour le couplage des données du T1FF 2010 au Recensement de 2011. Le couplage pourra se faire au niveau des personnes donnant un taux d'appariement plus élevé et de meilleure qualité, mais nécessitant beaucoup plus d'efforts. Sinon, le couplage pourra se faire seulement à l'aide d'un appariement par adresse nécessitant beaucoup moins d'efforts, mais réduisant la qualité des résultats d'appariement. L'information du T1FF 2010 pourra aussi servir à compléter l'information manquante des logements non-répondants ou supposés vacants au moment du recensement. Ensuite, sur une base annuelle, le fichier à l'échelle logement sera mis à jour en se servant du dernier T1FF disponible. Le T1FF sera couplé à la base d'information sociodémographique en utilisant l'adresse seulement. Pour une adresse donnée, si un lien est établi avec le T1FF, alors toute l'information sociodémographique sera remplacée par l'information disponible sur le T1FF. Par contre, si aucun lien n'est établi avec le T1FF, l'information antérieure sera conservée dans la base. Ainsi, pour chaque adresse du RA, la base d'information sociodémographique contiendra toujours la dernière information disponible sur le ménage et les personnes de ce ménage.

Le deuxième fichier de la base d'information sociodémographique conservera de l'information à l'échelle des AD. L'AD représente une petite région composée d'un ou de plusieurs îlots de diffusion avoisinants et regroupant de 400 à 700 habitants. L'ensemble du Canada est divisé en AD. À partir des données provenant du Recensement de 2011 et de l'ENM, on effectuera l'agrégation de l'information ménage et personne à l'échelle des AD. Étant donné que l'information détaillée de l'ENM n'est disponible que pour un sous-ensemble de ménages (environ 20 %), alors il est plus judicieux d'agréger cette information à l'échelle des AD. Ce fichier contiendra des variables telles que le nombre de personnes et de ménages dans l'AD, le revenu médian, le niveau d'éducation médian, la proportion d'immigrants, et ainsi de suite. L'information dans ce fichier pourra être utile aux enquêtes-ménages pour la création de grappes ou strates de logements, ainsi que pour la modélisation de la non-réponse.

4. Conclusion

Au cours des dernières années, la qualité du RA maintenu à Statistique Canada pour le recensement de la population s'est considérablement améliorée et plusieurs processus de traitement et de couplage de données administratives ont été développés. La nécessité de trouver une autre base de sondage pour les enquêtes-ménages à composition aléatoire de numéros de téléphone combinée aux objectifs de l'AOB a mené à opter pour l'utilisation du RA comme base de sondage commune de logements pour le programme des enquêtes-ménages. Ainsi, ces enquêtes profiteront de l'infrastructure déjà en place pour la mise à jour du RA et de tous ses processus de traitement des données administratives. Les enquêtes pourront ainsi transiter à une méthodologie d'échantillonnage fondée sur les adresses et améliorer leur couverture de la population. L'utilisation d'une base commune facilitera du même coup la coordination des échantillons entre les enquêtes-ménages. Le processus de création d'une base liste de numéros de téléphone et son couplage au RA permettront d'utiliser plus efficacement l'information sur les numéros de téléphone provenant des différentes sources de données. Cela devrait résulter en de meilleurs taux de contact par téléphone des logements visés. De plus, cette base de numéros de téléphone représentera une base liste améliorée qui pourra servir à sélectionner des échantillons directement, notamment dans le cadre d'une méthodologie d'échantillonnage de base duale. La création d'une base de données sociodémographiques reliée au RA viendra enrichir la base de sondage commune et servira à développer des plans d'échantillonnage plus efficace. Finalement, la mise en place d'une base de sondage commune viendra répondre aux objectifs de l'AOB concernant l'uniformisation et la centralisation des processus de traitement, de stockage et d'accès des données.

Remerciements

Les auteurs aimeraient remercier Windie Gagné, Jocelyne Marion et Jean-Louis Tambay de leur contribution à ce projet, ainsi que Yves Lafortune et Denis Poulin de leurs commentaires.

Bibliographie

Link, M., Daily, G., Shuttles, C., Yancey, T., Burks, A. et C. Bourquin (2009), « Building a New Foundation: Transitioning to Address Based Sampling After Nearly 30 Years of RDD », *Proceedings of the Survey Research Methods Section, American Statistical Association*, p. 5654-5665.

McClean, K. et K. Charland (2011), « État du Registre des adresses suivant le Recensement de la population de 2011 », rapport non publié, Ottawa, Canada : Statistique Canada.

SÉANCE 2B

ÉCHANTILLONNAGE ET ESTIMATION

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Effets de plan différentiels dans les échantillons d'enquêtes en milieu scolaire

Caroline Dahmen et Marek Fuchs[1]

Résumé

Plusieurs enquêtes en milieu scolaire de grande échelle s'appuient sur l'échantillonnage en grappes et peuvent par conséquent pâtir d'effets de plan qui ont une incidence sur la taille effective de l'échantillon. Dans le présent article, nous démontrons que les effets de plan diffèrent considérablement pour divers sous-groupes de l'échantillon correspondant, par exemple, aux divers types d'écoles et aux diverses années d'études. Qui plus est, les effets de plan au niveau de l'école et au niveau de la classe varient en importance selon le type d'école au sein d'un même échantillon. Donc, aucune stratégie uniforme ne convient pour la plupart des enquêtes en milieu scolaire et il faut tenir compte des effets de plan différentiels lors de la planification d'une étude.

Mots clés : Échantillons complexes ; effet de plan, enquêtes en milieu scolaire ; taille effective nette d'échantillon.

1. Introduction

Fréquent dans le domaine des sciences sociales, l'échantillonnage en grappes s'avère plus rentable que l'échantillonnage aléatoire simple, mais les échantillons résultants pâtissent parfois d'effets de plan (*deff*). La composition des grappes est souvent fondée sur des groupes naturels, comme des classes ou des logements. Donc, les éléments d'un même groupe ou d'une même grappe ont tendance à se ressembler plus que ceux d'un échantillon aléatoire simple (EAS) tiré de la même population. Par exemple, ils peuvent avoir des caractéristiques sociodémographiques similaires ou être exposés aux mêmes effets environnementaux. Ces similarités font que le groupe ou la grappe présente un certain degré d'homogénéité. Cela, à son tour, affecte l'estimation de la variance, de sorte qu'il faut accroître la taille des échantillons en grappes pour obtenir le même niveau de précision qu'avec un EAS. Cet accroissement est essentiel parce que l'on obtient moins de nouvelle information lorsque l'on enquête auprès d'un grand nombre d'éléments appartenant au même groupe ou à la même grappe (Cornfield, 1951; Groves et coll., 2009; Kish, 1965; Lohr, 1999).

Ainsi, des enquêtes sur le rendement scolaire telles que le PISA (OCDE, 2009) ou le projet TIMSS et PIRLS (2007), sont réalisées auprès d'échantillons en grappes à deux degrés en échantillonnant plusieurs élèves dans des écoles sélectionnées aléatoirement ou plusieurs classes dans chaque école. Afin de tenir compte du plan de sondage complexe, l'effet de plan est inclus dans le calcul de la taille d'échantillon requise. Cet ajustement est nécessaire parce que la valeur statistique d'un échantillon en grappes est plus faible que celle d'un EAS de même taille.

En principe, l'effet de plan précise dans quelle proportion la variance de l'estimation est sous-estimée dans un échantillon en grappes comparativement à un échantillon aléatoire simple. Il donne donc une mesure de la précision gagnée ou perdue en utilisant un plan de sondage plus complexe plutôt qu'un EAS (Lohr, 1999, page 309). Par conséquent, l'effet de plan indique de combien il convient d'augmenter la taille de l'échantillon pour atteindre la précision souhaitée.

Au moment de la planification d'une enquête avec échantillon en grappes, on ignore souvent quels sont les effets de plan réels qui doivent donc être estimés. Dans le cas des échantillons en grappes, les *deffs* dépendent des caractéristiques individuelles de l'enquête, comme l'homogénéité des grappes ou la taille de celles-ci, et ne peuvent être estimés qu'en se fondant sur les données produites dans le cadre d'une enquête particulière. De surcroît, les effets de plan ne sont pas identiques pour toutes les estimations calculées d'après le même échantillon (Kish, 1995) et peuvent également différer d'un sous-groupe à l'autre dans un échantillon (Verma et Lê, 1996). Malgré cela, dans

[1]Caroline Dahmen, University of Technology - Institute of Sociology, Residenzschloss, 64283 Darmstadt, Allemagne; Marek Fuchs, University of Technology - Institute of Sociology, Residenzschloss, 64283 Darmstadt, Allemagne (fuchs@ifs.tu-darmstadt.de).

la plupart des enquêtes, on applique un effet de plan global constant pour calculer la taille d'échantillon requise. Or, dans certains cas, cette approche pourrait être inappropriée et entrainer un biais dans l'estimation de la variance ou des coûts plus élevés (Kish, 1965).

Afin d'obtenir un échantillon à la fois rentable et précis, il est souhaitable d'estimer minutieusement les *deff*s prévus. Toutefois, puisque les données réelles ne sont pas encore disponibles, les effets de plan ne peuvent pas être déterminés durant la phase de planification d'une enquête. Pour trouver un estimateur fiable du *deff* en vue d'obtenir des valeurs approximatives, les chercheurs qui planifient une enquête ont deux options : soit mener une étude pilote pour recueillir les données nécessaires pour estimer les *deff*s ou rechercher une étude réalisée antérieurement en se servant d'un plan de sondage, d'un groupe cible et de variables similaires.

Dans notre cas, le problème consistait à concevoir en Allemagne une enquête par sondage pour une étude comparative à grande échelle auprès des élèves des écoles professionnelles dans sept pays européens. Comme il n'existait aucune liste complète de tous les élèves inscrits dans les écoles professionnelles en Allemagne, nous avons décidé dès le départ de réaliser l'enquête selon un plan d'échantillonnage en grappes axé sur les écoles (élèves dans les classes dans les écoles). Afin de produire des estimations appropriées du *deff*, nous nous sommes servis d'une enquête en milieu scolaire existante réalisée auprès de la même population. Durant l'analyse, deux questions de recherche principales ont été soulevées : premièrement, dans quelle mesure les *deff*s diffèrent-ils entre les sous-groupes dans la population cible? Deuxièmement, s'il existe des différences de *deff* entre ces sous-groupes, quels sont les facteurs qui les déterminent?

2. Effets de plan différentiels

2.1 Données et méthode

Les données utilisées pour l'analyse proviennent de l'étude de 2010 sur la violence dans les écoles (Fuchs, 2009). Cette étude à grande échelle sur les tendances en milieu scolaire est réalisée tous les cinq ans en Bavière, en Allemagne. En 2010, la taille de l'échantillon était d'environ 6 000 élèves regroupés dans 173 écoles en tout. De manière générale, deux classes par école étaient incluses dans l'échantillon. Celui-ci comprenait des élèves provenant de quatre types d'écoles différents qu'on retrouve en Allemagne.

L'Allemagne est dotée d'un système d'enseignement secondaire avec groupement selon les aptitudes dans lequel les élèves sont répartis à l'âge de 10 ans en fonction de leur rendement scolaire au primaire. Les principaux types d'écoles sont les écoles secondaires générales de niveaux inférieur, intermédiaire et supérieur, qui accueillent chacune un groupe particulier d'élèves (par exemple, les élèves ayant de très bonnes notes au primaire sont admis à l'école secondaire de niveau supérieur). En outre, la durée des études varie selon le type d'école. Alors que les élèves des écoles secondaires inférieures terminent leurs études après la neuvième ou la dixième année, ceux des écoles secondaires supérieures terminent leurs études secondaires après 12 années (précédemment 13 années) de scolarité.

En plus de ces trois types d'écoles secondaires générales, l'échantillon comprenait les écoles professionnelles. Après être sortis de l'un des trois types d'écoles secondaires générales, les élèves peuvent s'inscrire à un programme de formation professionnelle, ou entrer à l'université, à condition de détenir le diplôme d'entrée requis. Les jeunes qui souhaitent obtenir une formation professionnelle s'inscrivent dans une école professionnelle à temps plein ou à temps partiel, selon le programme de formation.

Kish a proposé d'estimer l'effet de plan en divisant la variance de l'estimateur établi pour un plan d'échantillonnage complexe par la variance de l'estimateur pour un EAS de même taille (Kish, 1965). Nous avons estimé le *deff* au niveau de l'école et au niveau de la classe en appliquant la formule qui suit.

$$Deff = 1 + \rho(\mathrm{B} - 1) \qquad (2.1\text{-}1)$$

La formule 2.1-1 (Kish, 1965) inclut le coefficient de corrélation intraclasse (ρ), qui mesure l'homogénéité d'une grappe, ainsi que la taille de la grappe (B). Par conséquent, les effets dus au plan d'échantillonnage en grappes sont reflétés.

S'il est facile de déterminer la taille des grappes, l'estimation du coefficient de corrélation intraclasse est plus compliquée. Nous avons appliqué un modèle de régression à trois niveaux pour décomposer la variance qui peut être expliquée à chaque niveau de notre plan d'échantillonnage. Dans la formule 2.1-2, le coefficient de corrélation intraclasse est calculé en tenant compte des niveaux inter et intra-école. Ici, la variance inter-écoles est divisée par la somme de la variance inter-écoles et de la variance intra-école. Tandis que σ^2_B représente la variance inter-écoles, σ^2_W représente la variance intra-école (Foy, 2004).

$$\rho = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W} \tag{2.1-2}$$

L'intervalle des valeurs de ρ va de zéro à un, où zéro représente l'hétérogénéité totale de l'échantillon, par exemple un EAS, et un indique que l'échantillon est entièrement homogène.

Dans nos analyses, la formule a été légèrement modifiée parce que nous avons également inclus la variance qui pourrait être expliquée au niveau individuel σ^2_I (voir 2.1-3 et 2.1-4). Cette modification est nécessaire en raison des trois niveaux (c'est-à-dire école, classe et élève) présents dans notre échantillon. Ici, σ^2_I représente la variance qui peut être expliquée au niveau de l'élève.

$$\rho_{école} = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W + \sigma^2_I} \tag{2.1-3}$$

$$\rho_{classe} = \frac{\sigma^2_W}{\sigma^2_B + \sigma^2_W + \sigma^2_I} \tag{2.1-4}$$

En utilisant les équations 2.1-3 et 2.1-4 recommandées par Hox (2002), nous avons pu calculer ρ et donc le *deff* séparément au niveau de l'école et au niveau de la classe. Le calcul des valeurs pour 51 variables séparément au niveau de l'école et de la classe, pour toutes les années d'études et tous les types d'écoles, a produit, en tout, 816 estimations. Ces estimations ont servi de matériel de base pour notre analyse.

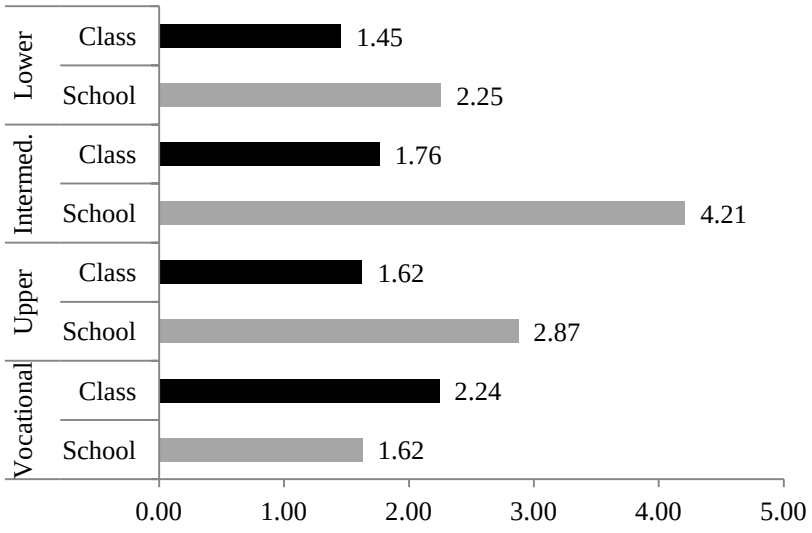
2.2 Résultats

Les valeurs inscrites à côté des barres de la figure 2.2-1 indiquent l'effet de plan pour chaque type d'école et niveau précis. Ici, les valeurs calculées correspondent au *deff* moyen pour les 51 variables.

Les barres grises de la figure 2.2-1 représentent les effets de plan au niveau de l'école (*School*). La comparaison de ces valeurs révèle d'importantes différences selon le type d'école. Ainsi, l'effet au niveau de l'école est plus de deux fois plus important pour les écoles secondaires de niveau intermédiaire que pour les écoles professionnelles (*Vocational*). Même si on limite la comparaison aux écoles secondaires générales, les effets de plan varient considérablement. En revanche, les différences entre les effets au niveau de la classe, représentés par les barres noires, sont plus faibles. En outre, les effets de plan au niveau de la classe sont généralement moins importants que ceux au niveau de l'école.

Les écoles professionnelles diffèrent des écoles secondaires générales qui affichent toutes la même tendance, à savoir des *deffs* au niveau de la classe de grandeur similaire et des *deffs* au niveau de l'école de grandeur variable, mais considérablement plus élevée que les valeurs observées au niveau de la classe. Par contre, pour les écoles professionnelles, le *deff* au niveau de la classe est plus important que le *deff* au niveau de l'école. En outre, les écoles professionnelles sont, de tous les types d'écoles, celles pour lesquelles le *deff* moyen est le plus faible. L'explication pourrait être qu'en Allemagne, les classes des écoles professionnelles sont composées d'élèves qui suivent principalement des cours menant à la même profession et qu'il existe des classes correspondant à différentes professions au sein d'une même école (par exemple, une école peut offrir des programmes de formation de cuisinier, de technicien en production chimique et de coiffeur). Par conséquent, les classes professionnelles sont plus homogènes que les écoles professionnelles.

Figure 2.2-1
***Deffs* moyens, selon le type d'école, au niveau de l'école et de la classe**



Comme l'étude sur la violence dans les écoles a également fourni des données pour différentes années d'études, nous avons pu déterminer les différences de grandeur des effets de plan pour divers groupes d'âge dans un type d'école et, par conséquent, obtenir des estimations encore plus précises des effets de plan. Nous avons donc réparti les élèves selon l'année d'études entre les cycles inférieur, moyen et supérieur.

Tableau 2.2-2
***Deff* moyen au niveau de l'école pour les différents types d'écoles, selon l'année d'études**

		Année d'études			
		Cycle inférieur (1)	Cycle moyen (2)	Cycle supérieur (3)	Total
Type d'école secondaire	Niveau inférieur (a)	1,82 (a2)(b1)	2,46 (a1)(b2)		2,25
	Niveau intermédiaire (b)	3,48 (a1)	4,48 (a2)(c2)		4,21
	Niveau supérieur (c)	2,82	2,34 (b2)	2,71 (d3)	2,87
	Professionnelle (d)			1,62 (c3)	1,62

Nota : Les combinaisons de lettres et de chiffres en exposants désignent des différences significatives ($p < 0,05$) par rapport à d'autres valeurs dans la même ligne ou colonne en utilisant le test post-hoc de Scheffé.

Comme le montre le tableau 2.2-2, nous avons constaté des différences d'effets de plan entre deux groupes d'âge pour un type d'école particulier. Cependant, seuls les *deffs* pour les écoles secondaires de niveau inférieur différaient de manière statistiquement significative entre les groupes d'âge indiqués par les exposants. Alors que le *deff* augmentait en passant des années du cycle inférieur à celles du cycle moyen dans les écoles de niveau inférieur et de niveau intermédiaire, les effets étaient plus faibles pour le groupe d'âge le plus âgé (années du cycle supérieur) que pour le groupe le plus jeune (années du cycle inférieur) pour les écoles secondaires de niveau supérieur. Donc, les effets de plan au niveau de l'école et au niveau de la classe diffèrent considérablement non seulement entre les types d'écoles, mais aussi entre les groupes d'âge, même si les différences ne deviennent statistiquement significatives que pour les écoles secondaires de niveau inférieur. Par conséquent, nous devons tenir compte dans l'échantillon de plusieurs sous-groupes pour lesquels la grandeur de l'effet de plan varie fortement.

Afin de déterminer les facteurs a l'origine des différences de *deff* au niveau de l'école, nous avons analysé les composantes qui déterminent le *deff.*

Tableau 2.2-3
Composantes déterminant le *deff* de différents types d'écoles au niveau de la classe et de l'école

			Taille de la grappe	Corrélation intraclasse (ρ)	*Deff* moyen
Type d'école secondaire	Niveau inférieur (a)	Classe	16	,029	1,45[(b)(d)]
		École	29	,053	2,25[(b)]
	Niveau intermédiaire (b)	Classe	23	,034	1,76[(a)(d)]
		École	46	,085	4,21[(a)(c)(d)]
	Niveau supérieur (c)	Classe	20	,034	1,62[(d)]
		École	39	,068	2,87[(b)(d)]
	Professionnelle (d)	Classe	20	,069	2,24[(a)(b)(c)]
		École	39	,017	1,62[(b)(c)]

Nota : Avec () = $p < 0,05$, test post-hoc de Scheffé

Au tableau 2.2-3, les composantes principales, c'est-à-dire la taille de la grappe et le coefficient de corrélation intraclasse (CCI), qui déterminent le *deff* selon l'équation 2.1-1, sont données séparément au niveau d'observation de la classe et de l'école pour quatre types d'écoles. L'examen du tableau révèle que les écoles diffèrent considérablement en ce qui concerne ces composantes. Alors que la taille de la classe est, en moyenne, de 16 élèves dans les écoles secondaires de niveau inférieur, elle est nettement plus élevée, atteignant 23, dans les écoles secondaires de niveau intermédiaire. Toutefois, les différences de CCI sont plus intéressantes. Dans l'ensemble, les valeurs paraissent assez faibles, la plus élevée étant égale à 0,085. Pourtant, elles diffèrent beaucoup, non seulement entre les écoles à un même niveau d'observation, mais aussi au sein des écoles selon le niveau d'observation. En particulier, les écoles professionnelles affichent une valeur de ρ assez faible, soit 0,017 seulement. Dans la dernière colonne, les *deffs* moyens pour chaque niveau d'observation et type d'écoles sont de nouveau indiqués et nous avons pu montrer que les différences de *deff* entre les écoles sont significatives dans la plupart des cas, comme l'indiquent les exposants. L'utilisation d'un test post-hoc nous a permis de démontrer que les *deffs* au niveau de l'école observés pour les écoles secondaires de niveau intermédiaire différent de manière fortement significative de tous les autres *deffs* observés au même niveau. En outre, les valeurs du *deff* diffèrent de manière significative d'au moins une autre valeur au même niveau.

Dans une analyse subséquente, nous avons examiné la question de savoir quels facteurs sont à l'origine des différences d'effets de plan entre les sous-groupes. Selon nous, ces différences sont imputables à deux facteurs, le premier étant la composition de l'effectif de chaque classe ou école en ce qui a trait au statut socioéconomique des familles, et le second étant le fait que, dans chaque classe ou école, les élèves passent beaucoup de temps ensemble et peuvent s'influencer les uns les autres et donc acquérir des attitudes et comportements similaires. Puisque nous ne pouvions pas analyser l'effet de l'interaction sociale dans les classes, nous nous sommes plutôt concentrés sur la composition de l'effectif des classes et des écoles. Comme nous l'avons montré plus haut, l'homogénéité des grappes varie considérablement selon les caractéristiques sociodémographiques. Nous avons supposé que la composition d'une grappe renseigne sur l'homogénéité de cette grappe.

Nous avons évalué six caractéristiques sociodémographiques, à savoir le niveau d'études du père, le niveau d'études de la mère, le revenu familial, et la taille de la collectivité dans laquelle vit l'élève. Comme nous avons émis l'hypothèse que la composition de l'effectif d'une classe ou d'une école est le facteur principal, plutôt que les caractéristiques individuelles des élèves, nous avons calculé l'écart-type des variables sociodémographiques dans chaque classe ou école. En outre, nous avons inclus dans les analyses le pourcentage de filles (écart en pourcentage par rapport à 50 %) et le pourcentage d'élèves ayant des antécédents d'immigration.

Les analyses indiquent que la composition de l'effectif d'élèves diffère selon le type d'école; toutefois, des différences importantes se dégagent aussi à l'intérieur des écoles. La comparaison des variables compositionnelles de l'effectif, par exemple, le pourcentage d'élèves ayant des antécédents d'immigration, entre les divers types d'écoles révèle de grandes différences. Alors que 33 % d'élèves des écoles secondaires de niveau inférieur (années du cycle inférieur) possèdent des antécédents d'immigration, la proportion est nettement plus faible dans les écoles

secondaires de niveau supérieur (environ 20 % seulement au cycle inférieur). Dans le cas du niveau d'études de la mère, l'écart-type pour les années du cycle inférieur est le même, soit environ 0,83, pour tous les types d'écoles. Pour les années du cycle moyen, la valeur diminue pour s'établir à 0,77 dans les écoles secondaires de niveau inférieur (et à 0,71 dans les écoles secondaires de niveau intermédiaire). Cela signifie que l'effectif d'élèves est légèrement plus homogène au cycle moyen qu'au cycle inférieur dans ces écoles. Dans les écoles secondaires de niveau supérieur, l'écart-type augmente en passant du cycle inférieur au cycle moyen. Ici, l'effectif d'élèves est plus hétérogène chez le groupe d'âge plus avancé.

Afin de déterminer quelles caractéristiques de la composition de l'effectif ont un effet significatif sur la grandeur des effets de plan au niveau de l'école, nous avons calculé des modèles de régression linéaire hiérarchiques (voir le tableau 2.2-4) en utilisant les écarts-types des variables sociodémographiques comme prédicteurs. Nous avons choisi les écoles secondaires générales de niveau intermédiaire comme catégorie de référence, et l'année comme variable de contrôle.

Tableau 2.2-4
Régression linéaire hiérarchique des *deffs* en fonction du niveau de l'école (coefficients de régression normalisés)

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7	Modèle 8
Niveau 1								
École de niveau inférieur†	-1,84***	-1,84***	-2,15**	-1,72***	-2,28**	-1,12	-1,82***	-0,84
École de niveau supérieur†	-1,36**	-1,48**	-0,85	-0,84*	-0,58	-1,25	-0,35	-0,67
École professionnelle†	-2,36***	-2,71***	-2,74**	-2,58***	-3,11***	-3,19+	-2,78***	-1,68
Année d'études		0,24	0,15	0,13	0,04	-0,33	0,12	-0,08
Niveau 2								
Niveau d'études du père$_{ET}$			-4,91					
Niveau d'études de la mère$_{ET}$				-7,38*	-8,06*	-6,86+	-11,12*	-6,48
Catégorie de revenu$_{ET}$					5,90			
Taille de la collectivité$_{ET}$						3,81		
Filles$_{\%1}$							-0,03	
Immigration$_{\%2}$								-5,89
Log-vraisemblance	-931,94	-931,95	-928,84	-927,79	-924,59	-924,51	-929,99	-924,08
n	408	408	408	408	408	408	408	408
Pseudo-$R^2_{McFadden}$	0,00703	0,00702	0,01033	0,01145	0,01486	0,01495	0,00911	0,01541

Nota : + p< 0,1; * p < 0,05; ** p < 0,01; *** p < 0,001, ET = écart-type, %1 = écart en pourcentage par rapport à 50 % de filles dans la grappe, %2 = pourcentage d'immigrants dans la grappe.
† Catégorie de référence = école secondaire de niveau intermédiaire

Nos résultats révèlent des effets négatifs significatifs pour tous les types d'écoles comparativement à la catégorie de référence. Ces effets demeurent significatifs pour la plupart des modèles, du moins pour les écoles professionnelles et les écoles secondaires de niveau inférieur. Dans les modèles comprenant les variables sociodémographiques, l'écart-type du niveau d'études de la mère est la seule variable produisant un effet significatif (modèle 4). L'ajout des autres variables réduit la force de cet effet, sauf celui du pourcentage de filles (modèle 7).

La variable dont l'effet est contrôlé dans le modèle 6, c'est-à-dire la taille de l'écart-type pour la collectivité dans laquelle vit l'élève, ne produit aucun résultat significatif, le type d'école et le niveau d'études de la mère n'ayant plus d'effet significatif. La variable « pourcentage d'immigrants » (modèle 8) rend entièrement non significatif l'effet des études et, dans ce modèle, nous constatons également que le type d'école n'a plus d'effet significatif non plus.

Ces résultats donnent à penser que l'homogénéité sociodémographique de la composition des grappes ne peut pas expliquer les différences de *deffs*, à l'exception du niveau d'études de la mère. À ce stade, il est nécessaire de

poursuivre les travaux en utilisant d'autres variables afin de déterminer si la composition des grappes a vraiment une influence.

2.3 Conclusion

Les résultats de nos analyses ont confirmé que les *deffs* varient considérablement entre les sous-groupes d'un échantillon. Comme l'effet de plan est un estimateur d'une variable particulière plutôt que d'un échantillon complet, les valeurs pour divers sous-groupes peuvent varier de manière différente. Nous avons montré que, dans les enquêtes en milieu scolaire, les différents types d'écoles et les différentes années d'études peuvent être considérés comme des sous-groupes pertinents, caractérisés par des *deffs* différents. Pour les échantillons sélectionnés en milieu scolaire, qui comprennent différents types d'écoles et différentes années d'études, il est recommandé d'utiliser des *deffs* personnalisés pour estimer la taille de l'échantillon. En outre, les travaux doivent se poursuivre afin de déterminer quelle est l'origine de ces différences.

Notre analyse ne nous a pas permis de confirmer notre hypothèse selon laquelle l'homogénéité des sous-groupes est la cause des différences entre les *deffs*. Bien que nous ayons réussi à déterminer les sous-groupes pertinents dans les échantillons sélectionnés en milieu scolaire, nous n'avons pas pu cerner les facteurs à l'origine des différences des *deffs*.

Les variables sociodémographiques prises en considération dans l'étude ne permettent pas d'expliquer les différences en ce qui a trait à l'homogénéité. Néanmoins, nous ne pouvons pas abandonner l'hypothèse voulant que la composition de l'effectif d'élèves donne lieu à des différences de *deff*. Au contraire, l'analyse devrait être poursuivie en utilisant d'autres variables sociodémographiques. Qui plus est, outre la composition des classes, d'autres sources pourraient être à l'origine de la variation des effets. On pourrait prendre en considération les différences de compétences des écoles en ce qui concerne l'intégration de leurs élèves dans une classe ou dans l'école.

Tout bien considéré, nous recommandons d'abandonner l'approche consistant à se fonder sur un *deff* global pour calculer la taille des échantillons. En particulier, une approche « standard » est déconseillée dans le cas de plans d'échantillonnage complexes comportant plusieurs sous-groupes dont l'homogénéité diffère sensiblement. Dans ces conditions, une plus grande souplesse est de rigueur afin de pouvoir ajuster au mieux la taille de l'échantillon, pour qu'il soit rentable et permette de calculer des estimations précises. Par conséquent, le calcul de la taille de l'échantillon devrait refléter les particularités de l'enquête et être fondé sur des *deffs* personnalisés.

Bibliographie

Cornfield, J. (1951), « Modern Methods in the Sampling of Human Populations », *American Journal of Public Health*, vol. 41, n° 6, p. 654 à 661.

Foy, P. (2004), *Intraclass Correlation and Varicance Components as Population Attributes and Measures of Sampling Efficiency in PIRLS 2001*, Hambourg: IEA Data Processing Center.

Fuchs, M. (2009), « Impact of school context on violence at schools – A multi-level analysis », *Journal International École et Violence,* vol. 7(2), p. 20 à 42.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J., Singer, E. et R. Tourangeau (2009), *Survey Methodology*, 2e édition, New York : Wiley Series.

Hox, J. (2002), *Multilevel Analysis : Techniques and applications*, Quantitative Methodology Series, New York : Psychology Press.

Kish, L. (1965), *Survey Sampling*, New York : Wiley Series.

Kish, L. (1995), « Methods for design effects », *Journal of Official Statistics*, vol. 11, n° 1, p. 55 à 77.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, 2e édition, Pacific Grove : Duxbury Press, 592 pages.

OCDE. (2009), *PISA 2006*, rapport technique, publication de l'OCDE.

TIMSS & PIRLS International Study Center (2007), *TIMSS 2007 Technical Report: Chapter 5. Sample Design*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Verma, V. et T. Lê (1996), « An Analysis of Sampling Errors for the Demographic and Health Surveys », *Revue Internationale de Statistique*, vol. 64, n^{o} 3, p. 265 à 294.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Enquêtes sur l'effectif des Forces canadiennes : Examen de la pondération et de la non réponse

François Larochelle, Tingting Gou et Irina Goldenber[1]

Résumé

Le Directeur général – Recherche et analyse (Personnel militaire) (DGRAPM) effectue des travaux de recherche pour appuyer l'élaboration de politiques et de programmes relatifs à la gestion de l'effectif des Forces canadiennes. Cette recherche est fondée en grande partie sur des données recueillies au moyen d'enquêtes sur l'effectif. Le présent article décrit le cheminement du DGRAPM dans sa recherche de méthodes permettant de calculer des estimations exactes et fiables d'après les enquêtes en vue de fournir des données probantes pour l'élaboration de programmes et de politiques concernant l'effectif. Bien que le recours à la pondération ajustée par rapport à la population ait des avantages dans le contexte de l'estimation en recherche opérationnelle, le présent article décrit les difficultés pratiques associées à la production d'estimations exactes. La principale de ces difficultés tient à la non réponse aux enquêtes. L'article expose les résultats d'une étude de cas entreprise en vue de déterminer quelles sont les variables démographiques associées à la non réponse et le moyen de surmonter cette importante difficulté.

Mots clés : Sondage « À vous la parole » auprès des Forces canadiennes ; non réponse ; pondération ajustée par rapport à la population.

1. Introduction

La gestion efficace de l'effectif des Forces canadiennes (FC) est essentielle du point de vue du recrutement, de la formation, de la préparation, du soutien et de l'hommage au personnel militaire pour services rendus au Canada. Au sein du ministère de la Défense nationale, le Directeur général – Recherche et analyse (Personnel militaire) (DGRAPM) appuie la gestion de l'effectif en menant des recherches qui permettent d'élaborer des politiques et des programmes fondés sur des données probantes. Les enquêtes auprès du personnel militaire représentent un outil de recherche essentiel en vue d'atteindre cet objectif et, chaque année, le DGRAPM conçoit, administre et analyse diverses enquêtes et en diffuse les résultats.

Le DGRAPM est un organisme assez petit qui travaille dans un environnement dans lequel le temps disponible pour épurer et analyser les données est généralement limité parce que le client doit connaître les premiers résultats dans un très bref délai après la collecte des données. La majorité des scientifiques qui travaillent au DGRAPM ont une formation en sciences sociales, habituellement en psychologie ou en sociologie. Ils possèdent ainsi les compétences et les connaissances nécessaires pour se pencher sur une vaste gamme de problèmes, tant quantitatifs que qualitatifs, concernant l'effectif. Bien qu'il possède des compétences spécialisées en élaboration et en analyse d'enquête, le DGRAPM ne compte pas à l'heure actuelle de scientifiques spécialisés en méthodologie d'échantillonnage dont le rôle officiel serait d'offrir ce genre d'encadrement et d'expertise.

Les taux de réponse aux sondages effectués auprès de l'effectif des FC sont généralement inférieurs à 40 %, ce qui rend l'analyse des données plus complexe et soulève des questions concernant le biais des estimations d'enquêtes. Étant donné le contexte de recherche opérationnelle dans lequel fonctionne le DGRAPM, résoudre ces difficultés n'est pas une tâche simple, puisqu'elle nécessite une connaissance et une expérience approfondies de la méthodologie d'enquête, ainsi que des ressources. Le DGRAPM est conscient de cette situation et reconnaît aussi le rôle critique que joue la recherche par sondage au sein de l'organisme. Par conséquent, ces dernières années, une plus grande attention a été accordée à la question de la non réponse, aux méthodes utilisées pour produire les

[1]François Larochelle, Directeur général – Recherche et analyse (Personnel militaire), ministère de la Défense nationale, 285, chemin Coventry, Ottawa (Ontario), Canada, K1A 1V0; Tingting Gou, Ph.D. en statistiques, Ottawa (Ontario), Canada; Irina Goldenberg, Ph.D., Directeur général – Recherche et analyse (Personnel miliaire), ministère de la Défense nationale, 285, chemin Coventry, Ottawa (Ontario), Canada, K1A 1V0.

résultats d'enquête, ainsi qu'à la validité des résultats et à la possibilité de les généraliser. Cet examen aide le DGRAPM à aller de l'avant en élaborant des méthodes plus efficaces et scientifiquement rigoureuses pour calculer les estimations, et pour mesurer leur fiabilité.

Le présent article passe en revue le cheminement du DGRAPM dans sa recherche de méthodes efficaces pour calculer des estimations exactes et fiables d'après des données d'enquête. Bien que la pondération ajustée par rapport à la population offre des avantages comparativement à d'autres méthodes de pondération pour l'estimation dans un contexte de recherche opérationnelle, l'article discute aussi des difficultés pratiques associées à la production d'estimations exactes. Enfin, il décrit les résultats d'une étude de cas entreprise pour déterminer quelles sont les variables démographiques associées à la non réponse.

2. Notation de base

N désigne la taille totale de la population cible à partir de laquelle est tiré un échantillon et, pour chaque indice i dans l'ensemble $\{1, 2, \ldots, N\}$, Y_i désigne la valeur d'une variable d'intérêt étudiée pour le i^{e} membre de cette population. Cette variable est mesurée d'après l'enquête pour un échantillon de n unités, et S_r désigne l'ensemble d'indices pour les n unités de population qui ont répondu à l'enquête.

3. Estimation au DGRAPM

3.1. Prise en considération des statistiques descriptives de l'échantillon

Une catégorie simple de statistiques qui peut être prise en considération pour présenter les résultats d'enquête est celle des « statistiques descriptives de l'échantillon ». Ces dernières sont utilisées pour décrire les caractéristiques des répondants à l'enquête et comprennent entre autres des statistiques comme la médiane de l'échantillon, le mode de l'échantillon et la moyenne de l'échantillon

$$\bar{y} = \frac{\sum_{i \in S_r} Y_i}{n}. \qquad (1)$$

Les statistiques descriptives de l' échantillon fournissent des renseignements sur l'échantillon de répondants et, tant qu'elles ne sont pas généralisées à la population plus grande dont a été tiré l'échantillon, aucune erreur d'échantillonnage n'y est associée. Dans ce contexte, ces statistiques sont très avantageuses si l'on veut fournir rapidement des résultats exacts au client, parce qu'elles peuvent être calculées sans examen particulier ni analyse de la non réponse et des erreurs de couverture. En outre, elles peuvent être diffusées sans fournir de mesure de l'erreur d'échantillonnage, telle que l'écart type ou le coefficient de variation.

Bien que la présentation de statistiques descriptives de l'échantillon pour décrire les répondants à l'enquête ait des avantages, dans un contexte de recherche opérationnelle, les clients s'intéressent généralement à l'obtention de résultats d'enquête qui décrivent exactement la population cible dont a été tiré l'échantillon. Par conséquent, des efforts supplémentaires doivent être investis dans la production d'estimations fiables et pertinentes qui peuvent être généralisées à l'ensemble de la population d'intérêt.

3.2. Prise en considération des estimateurs de la moyenne de la population

Dans le cas des enquêtes menées par le DGRAPM, le paramètre de population inconnu qui intéresse les clients est le plus souvent la moyenne de la population

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}. \qquad (2)$$

La moyenne de l'échantillon $\bar{y}$ peut servir à estimer $\bar{Y}$, mais contrairement au cas de son utilisation comme statistique descriptive de l'échantillon, elle est alors associée à une erreur d'échantillonnage et est également sujette à un biais dans ce contexte. Le biais et la variance de $\bar{y}$ doivent donc être étudiés afin de déterminer si $\bar{y}$ est un estimateur approprié ou si un autre estimateur, ayant un biais et (ou) une variance plus faible, peut être calculé moyennant un budget acceptable.

Quand on analyse le biais d'un estimateur, trois aspects importants doivent être pris en considération, à savoir les probabilités de sélection dans l'échantillon, ainsi que la non-réponse et le sous-dénombrement. Par exemple, Cochran (1977), ainsi que Biemer et Christ (2008) utilisent des cadres simples pour illustrer et analyser le biais éventuel de $\bar{y}$ résultant de la non-réponse et des erreurs de couverture. Afin de réduire le biais dû aux probabilités de sélection inégales, à la non-réponse et au sous-dénombrement, les statisticiens d'enquête ont proposé diverses techniques pour calculer des estimateurs pondérés de la forme $\bar{y}_w = \left(\frac{1}{N}\right)\sum_{i\in S_r} w_i Y_i$, où la somme des poids $\sum_{i\in S_r} w_i$ est égale à la taille totale de la population *N*, et le poids $w_i = \pi_i^{-1} a_{ri}\, a_{pi}$ est le produit du poids d'échantillonnage π_i^{-1} et de deux facteurs d'ajustement de la pondération, a_{ri} et a_{pi}, définis comme il suit :

1) π_i^{-1} est l'inverse de la probabilité de sélection dans l'échantillon pour la *i*e unité de population;
2) a_{ri} est un facteur d'ajustement de la pondération pour tenir compte de la non-réponse. L'inverse a_{ri}^{-1} est une estimation de la probabilité que la *i*e unité de population réponde à l'enquête, sachant qu'elle a été sélectionnée dans l'échantillon;
3) a_{pi} est un facteur d'ajustement pour la poststratification. L'inverse a_{pi}^{-1} est une estimation de la probabilité que la *i*e unité de population soit incluse dans la base de sondage à partir de laquelle a été tiré l'échantillon.

Statistique Canada (2003), ainsi que Biemer et Christ (2008) donnent une discussion plus détaillée de la construction des estimateurs pondérés de la forme

$$\bar{y}_w = \left(\frac{1}{N}\right)\sum_{i\in S_r} \pi_i^{-1} a_{ri} a_{pi} Y_i \ . \qquad (3)$$

Bien que la valeur de π_i soit connue et déterminée par le plan d'échantillonnage, le calcul des facteurs d'ajustement a_{ri} et a_{pi} requiert une analyse des mécanismes qui sous-tendent le sous-dénombrement et la non-réponse. La pondération est un sujet important en méthodologie d'enquête et, comme il est mentionné plus haut, diverses méthodes ont été proposées par les statisticiens d'enquête pour calculer a_{ri} et a_{pi} selon le type de renseignements auxiliaires disponibles sur la population cible, les répondants et les non-répondants (par exemple, voir Holt et Elliot, 1991; Deville et Särndal, 1992 et Lynn, 1996). En plus des exigences concernant les données, il convient de tenir compte de la complexité, de la robustesse, de l'incidence sur la variance et de l'intensité des calculs de ces méthodes pour décider d'une stratégie de pondération appropriée à utiliser.

La pondération ajustée par rapport à la population est présentée dans Lynn (1996) comme une méthode d'ajustement des poids pour tenir compte de la non-réponse qui peut être utilisée quand on ne dispose d'aucuns renseignements auxiliaires sur les non-répondants. Son exécution requiert la création de classes de pondération qui satisfont les conditions suivantes :

- les classes de pondération sont mutuellement exclusives et conjointement exhaustives;
- la population de taille N_k de chaque classe de pondération *k* est connue au moyen d'une source de données externe telle qu'une liste administrative;
- la classe de pondération à laquelle appartient un répondant peut être identifiée en utilisant l'information auxiliaire recueillie dans le cadre de l'enquête;

- la distribution de la variable d'intérêt *Y* chez les répondants qui appartiennent à une même classe de pondération est similaire à la distribution de *Y* dans la population pour cette classe.

L'estimation pondérée par rapport à la population a la forme générale

$$\bar{y}_{pb} = \left(\frac{1}{N}\right)\sum_{i\in S_r} \pi_i^{-1} a_i Y_i \,, \tag{4}$$

où a_i est le facteur d'ajustement du poids pour la i^{e} unité de population et satisfait

$$a_i = \frac{\sum_{k=1}^{L} \delta_i^k N_k}{\sum_{j\in S_r}\sum_{k=1}^{L} \delta_j^k \pi_j^{-1}} \,. \tag{5}$$

Dans l'équation susmentionnée, *L* désigne le nombre de classes de pondération et δ_i^k est une variable binaire qui est égale à 1 si la i^{e} unité de population appartient à la classe *k* dans {1, …, *L*} et à 0 autrement. De la définition de a_i dans l'équation (5), il découle que le facteur a_i est le même pour toutes les unités appartenant à une même classe de pondération et que la somme des poids $\sum_{i\in S_r} \delta_i^k \pi_i^{-1} a_{pi}$ pour les répondants appartenant à la classe *k* est égale à la taille de population N_k de la classe. Comparativement à l'estimation pondérée $\bar{y}_w$ définie dans l'équation (3), l'estimation pondérée par rapport à la population $\bar{y}_{pb}$ ne nécessite le calcul que d'un facteur d'ajustement a_i au lieu de deux (a_{ri} et a_{pi}) et, comme le mentionne Lynn (1996, p. 210), « en plus de corriger la non-réponse, la pondération faite de cette façon intègre simultanément un élément de poststratification ».

La pondération ajustée par rapport à la population est donc une option intéressante pour le DGRAPM, étant donné que les classes de pondération peuvent être construites en utilisant l'information démographique disponible pour la population des FC déjà recueillie par le Directeur – Système de gestion du personnel militaire (DSGPM) au sein du ministère de la Défense nationale. La base de données administrative tenue à jour par le DSGPM contient des renseignements tels que le sexe, la première langue officielle, l'état matrimonial, l'âge, le nombre d'années de service, le grade, le groupe professionnel militaire et l'unité des membres des FC. Lorsqu'une enquête est conçue, la base de données est utilisée pour extraire un instantané de la population cible à partir duquel sera créée la base de sondage. Les données extraites peuvent également être utilisées pour créer des classes de pondération à partir des variables choisies pour lesquelles des données sont disponibles par l'entremise du DSGPM ainsi que recueillies au moyen de questions d'enquête sur les caractéristiques démographiques.

3.2. Prise en considération de l'usage de la pondération ajustée par rapport à la population en pratique

Si la pondération ajustée par rapport à population semble répondre aux besoins du DGRAPM en tant que méthode efficace de calcul d'estimations exactes et fiables dans un contexte de recherche opérationnelle, en pratique la méthode pose certaines difficultés, dont les suivantes :

- les chiffres de population pour les classes de pondération pourraient ne pas être exacts à cause de la fiabilité de l'information sur les variables de pondération dans la base de données du DSGPM. Le cas échéant, il serait peut-être préférable d'utiliser un estimateur comportant un ajustement des poids d'échantillonnage pour la non-réponse seulement (pas d'ajustement de poststratification) à partir des renseignements auxiliaires recueillis auprès des répondants et des non-répondants (pondération ajustée par rapport à l'échantillon) ;

- afin de réduire efficacement le biais de non-réponse, les variables démographiques choisies pour définir les classes de pondération doivent être associées à la non-réponse et à la variable d'enquête *Y* d'intérêt. Pour faire un bon choix à cet égard, il faut solliciter la collaboration des spécialistes du domaine et, s'il existe

plus d'une variable d'enquête d'intérêt, il est parfois difficile de trouver un ensemble de variables raisonnables pour définir les classes de pondération ;

- comme l'a souligné Lynn (1996), les variables démographiques pour lesquelles des données du DSGPM sont disponibles pour créer les classes de pondération ne sont pas corrélées à la non-réponse. Le cas échéant, la pondération ajustée par rapport à la population ne réduira pas efficacement le biais de non-réponse ;

- comme en ont discuté Biemer et Christ (2008), Kim et coll. (2007), et Elliot (1999), certaines classes de pondération peuvent contenir un petit nombre de répondants et (ou) donner lieu à des ajustements extrêmes de la pondération. Dans ce cas, les spécialistes du domaine doivent être consultés afin de fusionner ces classes avec d'autres de manière à éviter un accroissement de la variance $\bar{y}_{pb}$ que ne justifie pas la réduction du biais résultant de la pondération ajustée par rapport à la population ;

- étant donné le faible taux de réponse aux enquêtes menées par le DGRAPM, il est très important d'évaluer l'effet de la non-réponse sur la fiabilité de $\bar{y}_{pb}$. Cependant, l'exercice est également difficile et coûteux, parce qu'il nécessite une analyse du mécanisme de non-réponse et du biais à l'intérieur classes de pondération si la non-réponse n'est pas répartie au hasard.

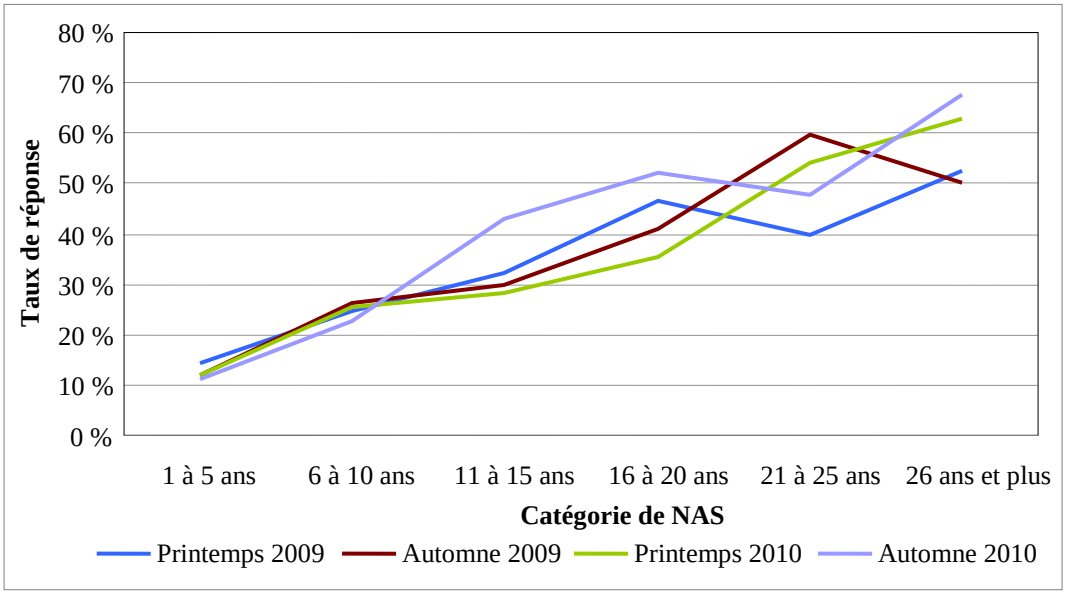
Malgré la simplicité relative de la pondération ajustée par rapport à la population, les difficultés susmentionnées soulignent l'importance de la planification et de l'affectation de ressources suffisantes au calcul des poids de sondage après la période de collecte des données. Elles font également ressortir l'importance de l'analyse de la non-réponse et des erreurs de couverture en vue d'évaluer la fiabilité de $\bar{y}_{pb}$ et de réduire son biais, ainsi que les avantages d'avoir accès à une expertise en méthodologie d'enquête (à l'interne ou à l'externe) pour pouvoir calculer efficacement des estimations exactes.

4. Prise en considération de la non réponse dans le sondage « À vous la parole » des FC

Comme nous l'avons mentionné à la section précédente, les faibles taux de réponse (généralement inférieurs à 40 %) aux enquêtes auprès de l'effectif des FC soulèvent des questions fondamentales concernant le biais de non réponse et la fiabilité des estimations en ce qui concerne leur généralisation à la population cible. L'évaluation de ce biais requiert l'investissement de ressources dans l'analyse du mécanisme de non réponse des membres des FC afin de pouvoir mesurer le biais résiduel dans les classes de pondération utilisées pour corriger les poids des estimateurs d'enquêtes pour la non-réponse. En guise d'étape initiale dans cette direction, le DGRAPM s'est servi d'un sondage auprès du personnel exécuté à intervalles réguliers pour analyser les courbes de non réponse afin de déterminer les variables démographiques associées à la non réponse aux enquêtes en ligne.

Le sondage « À vous la parole » (SAVP) est une enquête en ligne semestrielle qui est utilisée pour recueillir des renseignements sur les attitudes et les opinions des membres de la Force régulière concernant divers sujets importants qui concernent le personnel (Urban, 2007). La non réponse aux quatre SAVP effectués entre 2009 et 2010 (printemps 2009, automne 2009, printemps 2010 et automne 2010) a été modélisée par régression logistique. Cinq variables explicatives pour lesquelles des données existent pour les répondants (provenant du SAVP) et pour les unités figurant dans la base de sondage (provenant du DGSPM) ont été prises en considération pour l'analyse. Ces variables étaient le sexe, l'élément (Armée, Marine, Force aérienne), la première langue officielle (français ou anglais), le grade militaire et le nombre d'années de service (NAS). Parmi ces variables, le NAS s'est avéré être le meilleur prédicteur de la non réponse et, comme le montre l'examen de la figure 4 1, la probabilité de réponse la plus faible s'observe pour les membres des FC comptant moins de cinq années de service. En outre, la probabilité de réponse augmente en général parallèlement au nombre d'années de service d'un membre des FC. Étant donné l'association positive entre le grade et le nombre d'années d'ancienneté, le grade militaire s'est également avéré être une importante variable associée à la non réponse.

Figure 4-1
Taux de réponse selon la catégorie de NAS pour les quatre SAVP réalisés en 2009 et 2010



Les résultats de l'analyse des courbes de non-réponse au SAVP ont permis de prédire avec plus de précision la non-réponse aux enquêtes en ligne sur l'effectif des FC, et donc d'intégrer la non-réponse dans le calcul des tailles d'échantillon requises à l'étape de l'élaboration du plan d'échantillonnage. Par exemple, l'approche qui suit décrit comment un modèle de non-réponse (fondé sur une ou sur plusieurs variables démographiques telles que le grade et le NAS) peut être utilisé pour estimer les tailles d'échantillon requises par strate sous échantillonnage aléatoire stratifié quand une ou plusieurs variables explicatives ne figurent pas parmi les variables de stratification.

1) Déterminer le nombre de répondants souhaités dans chaque strate.
2) Utiliser le modèle de non-réponse pour calculer le taux de réponse prévu dans chaque strate en se fondant sur la distribution de la population de la strate en ce qui concerne les variables explicatives.
3) Calculer la taille d'échantillon requise pour produire le nombre souhaité de répondants dans chaque strate selon le taux de réponse prévu.

Grâce à l'intégration de modèles de non-réponse dans le plan d'échantillonnage, le DGRAPM peut prédire plus exactement la non-réponse aux enquêtes et calculer les tailles d'échantillon requises pour produire des estimations pondérées par rapport à la population ayant des variances d'échantillonnage qui satisfont les critères de précision. Cependant, une analyse plus approfondie de la non-réponse demeure nécessaire pour mesurer le biais persistant dans les classes de pondération.

5. Conclusion

Les initiatives entreprises récemment en vue d'examiner les méthodes d'estimation actuelles mènent le DGRAPM à perfectionner ces méthodes afin d'améliorer la qualité des résultats d'enquête. Bien que la pondération ajustée par rapport à la population soit une méthode d'estimation efficace dans un contexte de recherche opérationnelle, elle pose encore des difficultés, dont un grand nombre peuvent être résolues en consultant des spécialistes de la méthodologie d'enquête pour faciliter la fourniture rapide d'estimations fiables aux clients. À cet égard, le DGRAPM recherche et élabore des moyens d'intégrer efficacement ce genre d'expertise dans sa recherche par sondage.

La non-réponse des membres des FC aux enquêtes sur l'effectif demeure le principal défi que pose la collecte de données d'enquête de qualité. La détermination des variables démographiques qui sont associées à la non-réponse est certainement un bon point de départ en vue de comprendre la non-réponse. L'étape suivante pourrait consister à

analyser les « paradonnées » (données au sujet du processus de collecte des données) (Laflamme, 2008) provenant de futures enquêtes pour essayer de répondre aux questions qui suivent concernant la non-réponse :

- Comment le taux de réponse augmente-t-il avec le temps pendant le déroulement de la période de collecte des données?

- Comment le taux de réponse augmente-t-il à la suite des rappels par courriel durant la période de collecte des données?

- Les caractéristiques des répondants aux enquêtes varient-elles à mesure que le nombre de rappels augmente?

Les réponses à ce genre de questions pourraient aider à mieux comprendre la non-réponse et à découvrir des solutions rentables en vue de la réduire. En effet, comme l'ont souligné Holt et Elliot (1991, page 333), « la meilleure approche du problème de la non-réponse totale aux enquêtes consiste en premier lieu à déployer d'énormes efforts en vue de la réduire au minimum ».

Bibliographie

Biemer, P. et S. Christ (2008), « Weighting survey data » dans *International Handbook of Survey Methodology*, dans J. Hox, E. De Leeuw et D. Dillman (Éds.), European Association of Methodology Series, Taylor & Francis Group, p. 317 à 341.

Cochran, W. (1977), *Sampling Techniques: Third edition*, Wiley series in Probability and Mathematical Statistics, New York.

Elliot, D. (1999), « Report of the task force on weighting and estimation », *Government Statistical Series Methodology Series No 16*, Royaume-Uni : Office for National Statistics.

Holt, D. et D. Elliot (1991), « Methods of weighting for unit non-response », *The Statistician*, vol. 40, p. 333 à 342.

Kim, J.J., Li, J. et R. Valliant, (2007), « Regroupement de cellules lors de la poststratification », *Techniques d'enquête*, vol. 33, n°. 2, p. 157 à 170.

Laflamme, F. (2008), « Understanding survey data collection through the analysis of paradata at Statistics Canada », *American Association for Public Opinion Research 63rd Annual Conference, 2008 American Statistical Association, Proceedings of the Section on Survey Research Methods*.

Lynn, P. (1996), « Weighting for non-response », *Survey and Statistical Computing 1996*, Chesham: Association for Statistical Computing, p. 205 à 214.

Statistique Canada (2003), *Méthodes et pratiques d'enquête*, n° 12-587-XIF au catalogue de Statistique Canada.

Urban, S. (2007), « Your-Say: A Review of Current Administration Procedures and Survey Content », RDDC CARO TN 2007-28, Centre d'analyse et de recherche opérationnelle, ministère de la Défense nationale, Canada.

Élaboration d'un cadre d'échantillonnage intégré pour les enquêtes auprès des entreprises : Études de simulation pour évaluer les gains d'efficacité liés à un plan d'échantillonnage à deux phases

Yi Li et Frédéric Picard[1]

Résumé

En 2010, Statistique Canada a entrepris l'élaboration du Programme intégré de la statistique des entreprises (PISE), qui vise à remanier la plateforme existante de l'Enquête unifiée auprès des entreprises et à la transformer en modèle généralisé pour la production de statistiques sur les entreprises. Le PISE utilisera le Registre des entreprises (RE) comme base d'échantillonnage. Toutefois, en raison de la nature différente de certaines enquêtes qui seront comprises dans ce programme, il se peut que les données contenues dans le RE ne soient pas efficaces pour la stratification. Afin de répondre aux besoins spéciaux des différentes enquêtes, il est proposé d'avoir recours à un plan d'échantillonnage à deux phases comme option. Dans la présente étude, nous avons évalué l'efficacité d'un plan de sondage à deux phases par rapport à un plan de sondage à une phase.

Mots clés : Programme intégré de la statistique des entreprises ; plan de sondage à une phase ; plan de sondage à deux phases ; simulation.

1. Introduction

Le Programme intégré de la statistique des entreprises (PISE) est un nouveau projet qui intégrera la plupart des enquêtes auprès des entreprises de Statistique Canada, afin d'améliorer l'efficacité de la production des statistiques économiques. Un élément clé du PISE est l'élaboration d'un cadre d'échantillonnage intégré, qui peut être appliqué à une vaste gamme de programmes d'enquêtes auprès des entreprises de Statistique Canada. En raison de la nature différente de certaines enquêtes qui seront comprises dans ce programme, il se peut que les données contenues dans le Registre des entreprises (RE), qui sera utilisé comme base de sondage, ne soient pas efficaces pour la stratification de ces enquêtes. Afin de répondre aux besoins spéciaux des différentes enquêtes, une des options possibles consiste à avoir recours à un plan de sondage à deux phases. C'est donc dire que les données recueillies dans l'échantillon de première phase serviront à mettre à jour et à enrichir la base de sondage, un échantillon de deuxième phase étant sélectionné à partir de ces données pour mieux cibler la population, grâce au filtrage ou à une stratification et une répartition plus efficaces. Le lecteur est invité à consulter le chapitre 9 de Särndal et coll. (1992) pour un examen exhaustif de la théorie du plan de sondage à deux phases.

Afin de pouvoir comparer, du point de vue du coût et de la qualité des données, l'efficacité de l'approche à deux phases par rapport à l'approche à une phase, nous avons procédé à des simulations à partir de deux enquêtes du PISE, l'Enquête annuelle sur les manufactures et l'exploitation forestière (EAMEF) et l'Enquête sur les dépenses en immobilisations (EDI). Le présent article résumera les résultats de l'étude. Nous commencerons par un aperçu du PISE (section 2), suivi par la description de la méthodologie de simulation (section 3). Nous présenterons les résultats principaux dans la section 4 et nous conclurons par des recommandations et une liste de travaux à venir (section 5).

[1]Yi Li, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6, (Yi.Li@statcan.gc.ca) ; Frédéric Picard, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6 (Frederic.Picard@statcan.gc.ca).

2. Aperçu du PISE et de son plan d'échantillonnage

La méthodologie du plan d'échantillonnage du PISE est décrite brièvement dans la présente section. Plus de détails concernant le PISE se trouvent dans Godbout (2011).

Le PISE est un remaniement de l'Enquête unifiée auprès des entreprises, un programme annuel élaboré en 1997, qui englobe maintenant environ 60 enquêtes. La couverture du PISE sera élargie à environ 120 enquêtes, y compris des enquêtes annuelles et infra-annuelles (trimestrielles et mensuelles). On donnera la priorité à des optimums globaux, plutôt que des optimums locaux d'enquêtes individuelles, on élaborera des systèmes et des méthodes souples pour les enquêtes auprès des entreprises et on utilisera de façon optimale les données administratives.

Même si la plupart des enquêtes du PISE sont définies par industrie, certaines d'entre elles, comme l'EDI, sont définies par leurs activités. Certaines enquêtes du PISE ciblent les données au niveau de l'entreprise, et certaines autres recueillent des données à des niveaux plus bas, comme l'établissement ou l'emplacement. Les populations des différentes enquêtes du PISE peuvent se chevaucher. Par exemple, la population de l'enquête annuelle sur le transport et celle de l'enquête mensuelle sur le transport sont presque les mêmes. Une enquête axée sur les activités pourrait chevaucher des enquêtes du PISE axées sur l'industrie. Il est par conséquent nécessaire d'élaborer un plan pour la coordination et la rotation de l'échantillon, afin de contrôler le fardeau de réponse.

La population du PISE, qui englobe presque toutes les industries de l'économie canadienne, sera stratifiée par industrie et par province, afin que des estimations provinciales et industrielles fiables puissent être produites pour répondre aux besoins du Système de comptabilité nationale. La définition de l'industrie sera fondée sur le Système de classification des industries de l'Amérique du Nord. En raison de la nature très asymétrique de la population d'entreprises, celle-ci sera aussi stratifiée selon la taille, afin de veiller à ce que les grandes entreprises importantes aient une plus grande probabilité de sélection, et que les petites entreprises aient une probabilité plus faible de sélection. La ou les variables de taille pourraient être le revenu, les dépenses, les actifs, les terrains, *etc*.

On a proposé l'échantillonnage de Bernoulli stratifié comme méthode d'échantillonnage parce qu'il facilite la coordination et la rotation de l'échantillon. Les échantillons du PISE seront sélectionnés au moyen du Système généralisé d'échantillonnage élaboré à Statistique Canada. Les lecteurs qui souhaitent en savoir davantage sur le plan d'échantillonnage du PISE peuvent consulter Demnati et Turmelle (2011).

Il est aussi proposé d'avoir une option de plan à deux phases, afin que le plan d'échantillonnage soit aussi simple et souple que possible, d'une part, et qu'il puisse répondre aux défis et aux exigences particuliers de chaque enquête, d'autre part. Ainsi, un plan d'échantillonnage unifié sera appliqué à la première phase, et des données de base générales seront recueillies auprès de l'échantillon de première phase ; puis un sous-échantillon de deuxième phase sera sélectionné, en vue de recueillir des données financières et d'autres données détaillées. Les données recueillies à la première phase serviront à mieux cibler les répondants ou à restratifier les populations plus efficacement. La théorie de l'échantillonnage à deux phases est bien développée et on sait qu'elle peut réduire les coûts d'enquête et le fardeau de réponse dans certains cas. Toutefois, en raison de la complexité du PISE, ainsi que du fait que les enquêtes touchées ont plusieurs fins, nous avons dû effectuer une étude de simulation pour évaluer l'efficacité du plan de sondage à deux phases en comparaison avec le plan à une phase.

3. Description de la simulation et de sa méthodologie

3.1 Populations synthétiques

Deux populations synthétiques, une pour chacune des enquêtes sélectionnées, ont été construites à partir d'un mélange de données recueillies et imputées, afin d'imiter la population véritable. Elles comprenaient des variables de stratification (industrie, province et revenu pour les deux enquêtes, ainsi que pays de contrôle pour l'EDI) et des variables d'intérêt (livraisons de marchandises pour l'EAMEF et dépenses en construction d'immobilisations (CI) et en immobilisations de machine et de matériel (IMM) pour l'EDI). L'imputation a été fondée sur la méthode du plus proche voisin et a été mise en œuvre au moyen de BANFF, un système généralisé de contrôle et d'imputation élaboré à Statistique Canada.

3.2 Définition de l'unité d'échantillonnage et stratification de la population

Afin de pouvoir comparer les résultats de la simulation avec ceux de la production, les définitions de la stratification et de l'unité d'échantillonnage utilisées en production ont été appliquées aux deux enquêtes.

Tout d'abord, l'unité d'échantillonnage pour l'EAMEF était l'établissement et, pour l'EDI, tous les établissements des mêmes entreprises et à l'intérieur de la même cellule. Les unités d'échantillonnage étaient au nombre de 34 828 pour l'EAMEF et de 799 691 pour l'EDI dans les populations synthétiques. En deuxième lieu, on a calculé des seuils d'exclusion pour les deux enquêtes, afin d'exclure des populations les entreprises les plus petites qui représentaient jusqu'à 5 % à 10 % du revenu total d'une cellule. La partie des unités dont le revenu était inférieur aux seuils d'exclusion a été appelé strate à tirage nul. En troisième lieu, les populations des unités dont les revenus étaient supérieurs aux seuils d'exclusion ont été stratifiées selon l'industrie et la province (plus le pays de contrôle pour l'EDI). Enfin, l'algorithme de Lavallée-Hidiroglou (L-H) a servi à stratifier ces unités selon le revenu, en strates à tirage complet (TC), à tirage partiel grand (TPG) et à tirage partiel petit (TPP). L'algorithme de L-H est décrit dans Lavallée et Hidiroglou (1988). Alors que les unités de la strate à TC ont été sélectionnées avec certitude, celles de la TPG avaient une probabilité de sélection plus élevée que celles de la TPP. Dans le calcul des limites de taille, l'exigence de c.-v. a été établie comme étant la même que celle utilisée dans la production des deux enquêtes. Les unités auxquelles l'algorithme de L-H n'a pas permis d'attribuer de strates ont été étiquetées à « tirage obligatoire », ce qui signifie qu'elles ont été sélectionnées avec certitude.

3.3 Répartition et sélection de l'échantillon

On a eu recours à l'échantillonnage stratifié de Bernoulli pour la sélection des échantillons des plans de sondage à une phase et à deux phases.

Dans le plan de sondage à une phase, les tailles globales d'échantillon attendues étaient les mêmes que celles de la production : 13 500 pour l'EAMEF et 30 000 pour l'EDI. L'échantillon a été réparti au moyen de la répartition de puissance (avec la variable du revenu) pour l'EAMEF et de la répartition proportionnelle à la N pour l'EDI. Cette dernière méthode a été utilisée pour l'EDI en raison de la faible corrélation entre la variable de stratification (revenu) et les variables d'intérêt (CI et IMM).

Dans le plan de sondage à deux phases, les tailles globales d'échantillon attendues pour la première phase étaient de 18 000 dans le cas de l'EAMEF et de 30 000 dans le cas de l'EDI. Les échantillons ont été répartis à nouveau au moyen de la répartition de puissance selon le revenu pour l'EAMEF et de la répartition proportionnelle à la N pour l'EDI. La présente étude repose sur l'hypothèse de l'absence de non-réponse.

Dans une enquête à deux phases réelle, chaque unité sélectionnée à la première phase devra répondre à de brèves questions qui serviront au filtrage des unités de deuxième phase ou à la répartition de l'échantillon de deuxième phase. Par exemple, nous pourrions demander à l'unité de dresser une liste de ses trois principales marchandises (sans montants en dollars) pour l'EAMEF, ou lui demander si elle a des dépenses en CI ou en IMM (oui/non) pour l'EDI. Dans la simulation, pour chaque unité sélectionnée à la première phase, nous avons obtenu des renseignements à partir des populations synthétiques. Les unités de l'EDI qui ont répondu non aux deux questions ont été éliminées de la population pour la deuxième phase.

Pour la deuxième phase, l'échantillon de l'EAMEF a été réparti en minimisant la fonction objective multivariée $F = \sum_j \widehat{CV}\,(\hat{Y}_j)$ sous des contraintes de la taille de l'échantillon de deuxième phase et de fractions d'échantillonnage, où $\hat{Y}_j$ était l'estimateur pour le total de la variable d'intérêt, $\widehat{CV}(.)$ l'estimateur du coefficient de variation attribuable à la variance d'échantillonnage de deuxième phase, la somme ayant été calculée sur l'ensemble du domaine d'intérêt (marchandise*province). Nous avons utilisé les données sur les trois principales marchandises obtenues à la première phase pour estimer les coefficients de variation. Dans de cas de l'EDI, sur 1 000 répliques sélectionnées à la première phase, une moyenne de 19 286 unités faisaient partie du champ de l'enquête et ont été retenues. Puis, nous avons simulé deux autres scénarios, un fondé sur un recensement, et un autre, sur la sélection d'un sous-échantillon de 15 000 unités. On a eu recours à nouveau à la répartition proportionnelle à la N pour l'échantillon de deuxième phase.

3.4 Estimation

Des estimations des totaux pour les variables d'intérêt ont été produites et, à des fins de comparaison, des estimations des totaux pour la variable de taille (revenu) ont été aussi produites.

3.4.1 Calage des poids et niveaux d'estimation

Le poids initial d'une unité comprise dans l'échantillon est habituellement l'inverse de sa probabilité d'inclusion. Toutefois, comme la taille de l'échantillon selon l'échantillonnage stratifié de Bernoulli était aléatoire et que les tailles d'échantillon obtenues variaient d'un échantillon à l'autre, les poids ont été corrigés pour que la somme des poids des unités sélectionnées dans les strates pour la phase 1 (ou la phase 2) soit égale aux chiffres des strates de population correspondantes pour la phase 1 (ou la phase 2). Les poids calés ont servi à calculer les estimations pour chaque réplique.

3.4.2 Estimation de la précision

Pour évaluer la précision des plans d'échantillonnage à une phase et à deux phases, 1 000 répliques de Monte Carlo ont été sélectionnées pour chacun des plans. Une réplique de Monte Carlo d'un plan à deux phases consistait à sélectionner un échantillon de première phase, puis un sous-échantillon de deuxième phase. Pour chaque réplique, nous avons calculé les poids d'échantillonnage et estimé les totaux pour les variables d'intérêt.

Nous avons calculé la racine relative de l'erreur quadratique moyenne (RREQM), qui mesure la variation totale par rapport à la valeur de référence, de la façon suivante :

(1) $$RREQM = \frac{1}{Y}\sqrt{\frac{1}{1000}\sum_{r=1}^{1000}\left(\tilde{Y}_r - Y\right)^2}$$

où $\tilde{Y}_r$ représente l'estimation à partir de la r^{e} réplique et *Y*, la valeur du total de la variable d'intérêt à partir de la population synthétique.

3.4.2 Estimation des coûts

À partir des données sur les coûts opérationnels, on a appliqué les coûts unitaires suivants pour l'estimation des coûts de la collecte des données pour les plans à une phase et à deux phases.

Tableau 3.4-1
Coût unitaire de la collecte des données pour les plans à une phase et à deux phases

Une phase		Deux phases	
Collecte	**Coût unitaire**	**Collecte**	**Coût unitaire**
Contact préliminaire	1	Première phase (EAMEF)	1,5
		Première phase (EDI)	2
Questionnaire régulier	7	Deuxième phase : Questionnaire régulier	7
Questionnaire électronique	5	Deuxième phase : Questionnaire électronique	5

À noter que la première phase d'un plan d'échantillonnage à deux phases peut être considérée comme un contact préliminaire élargi du point de vue de la collecte des données. La raison du coût plus faible de première phase de l'EAMEF par rapport à l'EDI est qu'il y aura chevauchement significatif des échantillons d'une année à l'autre et le fait que les données recueillies à la première phase pour l'EAMEF (les trois principales marchandises) sont considérées comme suffisamment stables pour que l'on présume qu'elles n'ont pas à être mises à jour pour les unités qui étaient dans l'échantillon l'année précédente.

4. Résultats de la simulation

Dans la présente section, nous présentons les résultats de la simulation pour les plans à une phase et à deux phases pour chaque enquête. Seuls les résultats les plus intéressants sont présentés dans cette section. Des résultats détaillés se trouvent dans Pacquelet (2011) pour l'EAMEF et dans Xie et coll. (2011) pour l'EDI.

4.1 Analyse de la simulation de l'EAMEF

La qualité des données a été améliorée pour le plan à deux phases, comme on peut le voir dans le tableau 4.1-1, c'est-à-dire qu'un plus grand nombre de domaines (définis par marchandise*province) avaient une RREQM plus faible pour le plan à deux phases que pour le plan à une phase, avec à peu près les mêmes coûts totaux de collecte (tableaux 4.1-2 et 4.1-3). En outre, si nous additionnons la RREQM de ces 282 domaines, le total pour le plan à deux phases est de 1,26, ce qui représente une réduction substantielle par rapport au total de 3,99 obtenu pour le plan à une phase.

Tableau 4.1 -1
EAMEF : Comparaison de la RREQM entre les plans à une phase (RREQM1) et à deux phases (RREQM2)

Catégorie	Nombre de domaines	Pourcentage
RREQM1 > RREQM2	162	58 %
RREQM1 = RREQM2=0	32	11 %
RREQM1 < RREQM2	88	31 %
Total	282	100 %

Tableau 4.1-2
EAMEF : Coût du plan à deux phases avec une taille globale d'échantillon attendue de 18 600 pour la première phase et de 11 000 pour la deuxième phase

Collecte	Taille globale d'échantillon moyenne	Questionnaire électronique (1re phase : 0 %; 2e phase : 60 %)		Questionnaire régulier (1re phase : 100 %; 2e phase : 40 %)		Coût total global
		Coût unitaire	Coût total	Coût unitaire	Coût total	
1re phase	18 600			1,5	27 900	27 900
2e phase	11 000	5	33 000	7	30 800	63 800
Total						91 700

Tableau 4.1-3
EAMEF : Coût du plan à une phase avec une taille globale d'échantillon attendue de 13 500

Opération	Taille globale d'échantillon moyenne	Questionnaire électronique (contact préliminaire : 0 %; collecte : 60 %)		Questionnaire régulier (contact préliminaire : 100 %; collecte : 40 %)		Coût total global
		Coût unitaire	Coût total	Coût unitaire	Coût total	
Contact préliminaire	13 500			1	13 500	13 500
Collecte	13 500	5	40 500	7	37 800	78 300
Total						91 800

Nous voyons qu'avec des coûts similaires à ceux du plan à une phase, le plan à deux phases peut améliorer la qualité des estimations. Cela vient de ce que nous avons la possibilité, à la deuxième phase, de mieux cibler l'échantillon au moyen des données recueillies à la première phase. Nous avons procédé à d'autres simulations dans lesquelles le nombre de domaines d'intérêt variait, et nous avons noté qu'au fur et à mesure que le nombre de domaines d'intérêt

augmentait, le gain de précision à partir du plan à deux phases devenait moins important. Le plan à deux phases ne comporte pas d'avantages par rapport au plan à une phase lorsqu'il y a trop de variables d'intérêt.

4.2 Analyse de la simulation de l'EDI

Dans le cas d'un recensement pour la deuxième phase, la qualité des données du plan à deux phases est équivalente à celle du plan à une phase. Dans le cas d'une enquête sur échantillon pour la deuxième phase, la qualité des données du plan à une phase n'est que légèrement supérieure à celle du plan à deux phases lorsque l'on compare la RREQM. Toutefois, les tableaux 4.2-1 et 4.2-2 montrent une réduction de 16 % du coût, même lorsque la deuxième phase est un recensement. On peut économiser encore davantage en sélectionnant un échantillon pour la deuxième phase.

Tableau 4.2-1
EDI : Coût d'un plan à deux phases avec une taille globale d'échantillon attendue de 30 000 pour la première phase et un recensement pour la deuxième phase

Collecte	Taille globale d'échantillon moyenne	Questionnaire électronique (1re phase : 0 %; 2e phase : 60 %)		Questionnaire régulier (1re phase : 100 %; 2e phase : 40 %)		Coût total global
		Coût unitaire	Coût total	Coût unitaire	Coût total	
1re phase	30 000			2	60 000	60 000
2r phase	19 286	5	57 858	7	54 001	111 859
Total						171 859

Tableau 4.2-2
EDI : Coût d'un plan à une phase avec une taille globale d'échantillon attendue de 30 000

Opération	Taille globale d'échantillon moyenne	Questionnaire électronique (contact préliminaire : 0 %; collecte : 60 %)		Questionnaire régulier (contact préliminaire: 100 %; collecte : 40 %)		Coût total global
		Coût unitaire	Coût total	Coût unitaire	Coût total	
Contact préliminaire	30 000			1	30 000	30 000
Collecte	30 000	5	90 000	7	84 000	174 000
Total						204 000

5. Conclusion/Recommandations

Les résultats de l'étude ont démontré que le recours à un plan d'échantillonnage à deux phases peut réduire les coûts de la collecte et le fardeau de réponse par rapport à un plan à une phase, sans répercussions négatives sur la qualité des données, ou encore, qu'il peut aider à améliorer la qualité des données avec le même nombre de ressources.

À l'avenir, des études pourraient être menées pour d'autres enquêtes du PISE. Lorsque la corrélation entre les variables d'intérêt et le revenu n'est pas élevée, il peut être utile d'avoir une approche assistée par modèle pour produire des variables auxiliaires comportant des corrélations relativement élevées avec les variables clés d'intérêt, grâce à l'utilisation de toutes les données disponibles, plutôt que du revenu seulement. Les problèmes découlant de tailles d'échantillon réalisées aléatoirement pour les deux phases, en raison de l'échantillonnage de Bernoulli, doivent être résolus. Les répercussions de la non-réponse ou des erreurs de mesure à la première phase sur l'échantillon de deuxième phase devraient être aussi évaluées.

Bibliographie

Demnati, A. et C. Turmelle (2011), « Proposed Sampling and Estimation Methodology for the Integrated Business Statistics Program », rapport technique présenté au Comité consultatif des méthodes statistiques de Statistique Canada, Statistique Canada, Ottawa, Canada.

Godbout, S. (2011), « Normalisation du traitement des données après la collecte dans les enquêtes-entreprises à Statistique Canada », *Recueil du Symposium 2011 de Statistique Canada*, Statistique Canada, Ottawa, Canada.

Lavallée, P. et M. Hidiroglou (1988), « De la stratification des populations asymétriques », *Techniques d'enquête*, vol. 14, n^{o} 1, Statistique Canada, Ottawa, Canada.

Pacquelet, L. (2011), « Simulation de différentes méthodes d'allocation pour un plan de sondage à deux phases dans le contexte du PISE », document interne, Statistique Canada, Ottawa, Canada.

Särndal, C.-E., Swensson, B. et J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, 688 pages.

Xie, H., Li, Y. et J. Gaudet (2011), « A Preliminary Sampling Study for the Integrated Business Statistics Program Using the Capital Expenditures Survey Data », document interne, Statistique Canada, Ottawa, Canada.

Estimation de la variance par réplication peu dense et efficace pour les enquêtes complexes

Jae Kwang Kim et Changbao Wu[1]

Résumé

Il est pratique courante pour les organismes d'enquête de fournir des poids de réplication dans les fichiers de données d'enquête. Ceux ci servent à produire des estimations valides et efficaces de la variance, pour une gamme variée d'estimateurs, de manière simple et systématique. Toutefois, la plupart des méthodes d'élaboration des poids de réplication ne sont valides que pour des plans d'échantillonnage particuliers et exigent habituellement un nombre très important de répliques. Dans cette communication, nous démontrons d'abord comment produire des poids de réplication à partir de la méthode énoncée dans Fay (1984), de façon que l'estimateur de la variance par réplication en découlant soit équivalent au niveau algébrique à l'estimateur de la variance par linéarisation pleinement efficace, peu importe le plan d'échantillonnage. Afin d'atteindre simultanément l'efficience et la faible densité, nous proposons ensuite une nouvelle application de la méthode de calage pour les poids de réplication, de sorte qu'un petit nombre d'ensembles de poids de réplication puisse produire des estimateurs valides et efficaces de la variance par réplication pour des paramètres clés. La méthode proposée peut être utilisée en parallèle avec les techniques de rééchantillonnage existantes dans les enquêtes complexes à grande échelle. On aborde aussi un certain prolongement aux plans d'échantillonnage équilibrés. Les résultats de la simulation ont montré que la méthode proposée permet d'obtenir de très bons résultats. Les stratégies que nous proposons auront probablement des répercussions sur la façon dont les fichiers de données d'enquête à grande diffusion sont produits et dont ces ensembles de données sont analysés.

[1]Jae Kwang Kim, Iowa State University, États-Unis et Changbao Wu, Université de Waterloo, Canada.

SÉANCE 3A

HARMONISATION DES MÉTHODES DANS LE CADRE DE PROJETS DE NORMALISATION À GRANDE ÉCHELLE POUR LES ENQUÊTES AUPRÈS DES ENTREPRISES

Harmonisation de méthodologies dans le contexte d'un projet d'intégration de Systèmes : défis et leçons apprises

J. Andrews, F. Brisebois, I. Delahousse, C. Dochitoiu, M. Lachance, R. Philips et S. Pursey[1]

Résumé

En 2007, le projet de l'Examen de l'assurance de la qualité des données mené par Statistique Canada a indiqué que les systèmes de traitement des données de certaines enquêtes-entreprises mensuelles de la Direction de la statistique de l'industrie étaient un sujet de préoccupation. Certains de ces systèmes reposaient sur une technologie vieillissante, comportaient des schémas de traitement complexes, avaient fait l'objet de multiples personnalisations au fil du temps ou étaient dépourvus de certaines composantes fonctionnelles importantes. En réponse à cet examen, un projet de renouvellement des systèmes de traitement des données a été lancé, dont la vision était de livrer un système généralisé, simplifié et commun pouvant être utilisé par de nombreuses enquêtes mensuelles. Puisque les problèmes que pose un cycle de production mensuel étaient communs aux diverses enquêtes concernées, un environnement de traitement harmonisé devait profiter à toutes ces enquêtes en leur permettant de réaliser des gains d'efficacité et de partager les compétences pour résoudre les problèmes communs. Ce projet a également été considéré comme une excellente occasion d'harmoniser les méthodologies utilisées aux diverses étapes des enquêtes, comme la calendarisation des données déclarées, la vérification et l'imputation, ainsi que l'estimation. Pendant la création de ce nouvel environnement, un certain nombre de défis liés aux différences en matière de méthodologie, de communications et de ressources humaines ont dû être résolus. La communication discutera de ces défis et des leçons apprises durant le processus.

Mots clés : Harmonisation ; intégration ; système généralisé ; leçons apprises.

1. Introduction

Statistique Canada mène près de 350 enquêtes portant sur pratiquement tous les aspects de la vie des Canadiens, que ce soit au niveau social ou économique. Malgré la diversité de ces enquêtes, celles-ci doivent toutes répondre à des exigences communes de qualité et s'assurer de respecter la confidentialité des données. Pour la plupart de ces enquêtes, les opérations et méthodes utilisées à cette fin sont souvent très semblables, vues en survol. Par exemple, à l'exception de quelques recensements, la majorité des enquêtes procèdent par échantillonnage. Les données recueillies sont ensuite vérifiées et validées afin d'assurer certains critères de qualité. Le tout est ensuite diffusé par l'entremise de divers produits respectant les hauts standards à l'égard du respect de la confidentialité. Toutefois, en examinant les processus individuels de plus près, on constate rapidement que la similitude s'arrête là. Les différences, souvent légitimes, proviennent du fait que chaque enquête a des objectifs particuliers visant à répondre aux besoins spécifiques de son client et de ses utilisateurs. Par conséquent, malgré que ces différences soient inévitables, elles se traduisent souvent en complexités, réduisant ainsi l'efficacité globale de l'organisme statistique à livrer ses produits. L'harmonisation est donc une initiative souhaitable.

L'harmonisation vise à prévenir ou éliminer les différences dans les processus techniques ayant tous le même but. Pourquoi développer et maintenir différentes méthodes de traitement des données, si elles mènent toutes à des résultats semblables et de qualité équivalente? Dans un contexte de production continue, tel que celui présent dans les organismes statistiques, il est toutefois difficile de s'arrêter et d'examiner clairement la situation pour ensuite effectuer d'importants changements visant à harmoniser les méthodes d'enquête. L'harmonisation de processus d'enquête, quoique souhaitable, s'avère donc un exercice laborieux. Toutefois, les avantages peuvent être nombreux : réduction de la complexité et de la duplication des processus, amélioration de l'efficacité et réduction des coûts opérationnels, réduction des coûts opérationnels liés aux technologies informatiques en raison de l'infrastructure commune et d'une maintenance simplifiée, et finalement, instauration plus efficace des pratiques

[1]J. Andrews (jessica.andrews@statcan.gc.ca), F. Brisebois (francois.brisebois@statcan.gc.ca), I. Delahousse (ivelina.delahousse@statcan.gc.ca), C. Dochitoiu (catalin.dochitoiu@statcan.gc.ca), M. Lachance (martin.lachance2@statcan.gc.ca), R. Philips (robert.philips@statcan.gc.ca) et S. Pursey, Statistique Canada, Pré Tunney, 120, avenue Parkdale, Ottawa, Ontario, Canada, K1A 0T6.

exemplaires au sein des diverses équipes impliquées dans les différentes enquêtes, de même que la collaboration entre celles-ci.

Cet article donne un aperçu du Projet d'intégration des systèmes des enquêtes mensuelles de la Direction de la statistique de l'industrie, qui vise à intégrer trois enquêtes dans un même système de production, de même qu'à harmoniser les méthodes et pratiques employées. Les défis rencontrés et les leçons apprises tout au cours de ce projet sont présentés. La section 2 brosse un portrait global du contexte entourant la mise en oeuvre du projet. Ensuite, la section 3 décrit différents paramètres du projet, tels que son but, les enquêtes ciblées et l'échéancier; le projet peut être décomposé en quatre phases menant au produit final. La section 4 présente ces quatre phases en soulignant leurs propres défis et leçons apprises. Enfin, un sommaire de ces leçons, de même que certaines autres englobant l'ensemble du projet sont décrits à la section 5.

2. Contexte

Statistique Canada effectue des examens de l'assurance de la qualité visant à évaluer la rigueur des pratiques d'assurance de la qualité dans ses différents programmes. Ces examens ciblent précisément l'exécution d'un programme et non son plan, et vise à déterminer les risques qui pourraient entraver le programme à livrer ses produits de base. En 2007, neuf programmes ayant une mission cruciale ont été passés en revue, et des éléments à améliorer ont été proposés et mis en œuvre.

L'Examen de l'assurance de la qualité mené auprès de l'Enquête mensuelle sur les industries manufacturières (EMIM) a permis de déterminer que son système de traitement des données comportait un domaine à risque important en raison de son âge, de la complexité de son plan de sondage et de plusieurs fonctions spécialement conçues à des fins opérationnelles qui ont été intégrées au fil du temps. Le rapport fait mention que le renouvellement de son système de traitement des données est depuis longtemps nécessaire et que l'environnement de traitement actuel augmente de façon importante le risque d'erreurs, et de diffusion de données de piètre qualité.

L'examen effectué auprès d'une autre enquête mensuelle, l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD) indique que son système de production des données est un système plus à jour, et qu'il est intégré lentement aux normes de l'architecture opérationnelle. Toutefois, le système de production des données de l'EMCGD ne dispose pas de certaines composantes fonctionnelles importantes en ce qui a trait principalement à l'analyse des macrodonnées, telles que les séries chronologiques et les outils de révision.

Afin de relever les défis identifiés au moyen des examens de l'assurance de la qualité, une proposition a été soumise par la Direction de la statistique de l'industrie pour trouver une approche harmonisée de mener ses enquêtes-entreprises mensuelles. Le projet sera mis en œuvre par l'intégration des deux enquêtes examinées mentionnées ci-dessus (EMIM et EMCGD) et d'une autre enquête mensuelle importante couvrant l'industrie alimentaire, soit l'Enquête sur l'industrie des services de restauration et de débits de boisson mensuelle (EMSR).

3. Paramètres et déroulement du projet

3.1 But et objectifs du projet

Le but du projet d'intégration des systèmes est de livrer un système rationalisé, partagé et généralisé qui peut être utilisé par plusieurs enquêtes mensuelles. Étant donné la similitude des défis rencontrés au cours d'un cycle mensuel de production, ce nouveau système permettra aux enquêtes de bénéficier d'un environnement de traitement harmonisé tout en gagnant en efficacité et en partageant l'expertise lorsque des enjeux communs sont rencontrés.

Ce nouveau système permettra de satisfaire l'objectif principal de simplifier et d'harmoniser le processus actuel de traitement des enquêtes, d'automatiser le déroulement des opérations de traitement, de même que de répartir les rôles et responsabilités à l'ensemble des équipes de soutien des enquêtes mensuelles de la Direction de la statistique de l'industrie. Il est important de noter que la portée du projet inclut toutes les étapes de traitement suivant la collecte des données, excluant ainsi des étapes telles que l'échantillonnage et la collecte proprement dite.

Les efforts ont donc été dirigés vers la mise en oeuvre d'un nouveau système intégré. Un objectif secondaire consistant à harmoniser les méthodologies en place lorsqu'il est possible de le faire s'est aussi ajouté au projet. Cet objectif n'avait pas pour but de réinventer la méthodologie en place pour chacune des trois enquêtes, mais tentait plutôt de profiter de l'existence de systèmes généralisés et cherchait à identifier les occasions d'harmonisation pouvant être mises en oeuvre dans le futur. Cet objectif se traduit donc par une migration vers les systèmes généralisés de Statistique Canada lorsque ceux-ci ne sont pas déjà utilisés. Il est à noter que Statistique Canada possède et appuie un éventail de produits couvrant le spectre complet d'un cycle typique de production d'enquête : de l'échantillonnage (Système généralisé d'échantillonnage, SGECH) à l'estimation (Système généralisé d'estimation, SGE) en passant par la vérification et l'imputation (système BANFF), de même que plusieurs autres. Deguire, Reedman et Wenzowski (2011) présentent un aperçu de ces systèmes.

3.2 Description des enquêtes ciblées

Le projet d'intégration des systèmes cible trois enquêtes mensuelles de la Direction de la statistique de l'industrie de Statistique Canada. Même si ces enquêtes couvrent des populations différentes, elles partagent quand même plusieurs similitudes conceptuelles, opérationnelles et méthodologiques. Par exemple, les trois enquêtes sont à caractère obligatoire, utilisent le Registre centralisé des entreprises de Statistique Canada comme base de sondage, font appels aux dossiers fiscaux afin de réduire le fardeau de réponse des plus petites entreprises, *etc*. Elles se distinguent toutefois par l'état et la complexité de leur système respectif de production de données. Leur méthodologie diffère aussi dans quelques cas, mais ces différences résident souvent au niveau des paramètres utilisés dans les méthodes en place, et non dans les méthodes proprement dites. Les paragraphes qui suivent donnent un survol des trois enquêtes, tout en pointant les besoins de chacune en ce qui a trait aux systèmes.

L'EMIM produit des séries statistiques sur l'activité de l'industrie manufacturière, mesurant les ventes de biens fabriqués, les stocks, les commandes en carnet et les nouvelles commandes. Les données de l'EMIM servent d'indicateurs de la situation économique des industries manufacturières, de même qu'au calcul du produit intérieur brut du Canada. Le plus récent remaniement de cette enquête remonte à 1999. La technologie informatique a beaucoup évolué depuis, et la méthodologie a aussi été raffinée au cours des dernières années. Par exemple, un changement méthodologique important a été intégré à l'EMIM en 2004 afin de réduire le fardeau de réponse des petites entreprises. Au lieu de soumettre ces dernières aux procédures habituelles de collecte des données, il fut donc convenu d'utiliser les dossiers fiscaux de la Taxe sur les produits et services (TPS) afin de dériver leurs données (Thomas et Cook, 2005). Depuis le remaniement de 1999, le système de production des données a donc fait l'objet de plusieurs modifications et ajouts. Il est aussi important de noter que malgré l'utilisation d'un bon nombre de systèmes généralisés de Statistique Canada, d'autres systèmes pourraient être intégrés afin de remplacer certains programmes informatiques faits sur mesure.

L'EMCGD fournit des renseignements sur le rendement du secteur du commerce de gros et de détail, et constitue elle aussi un indicateur important de la santé de l'économie canadienne. L'Enquête fournit des estimations mensuelles des ventes et des stocks des marchands de gros, de même que des ventes et du nombre d'emplacements d'affaires pour le commerce de détail. Le dernier remaniement majeur de l'Enquête a été complété en 2004. Tout comme l'EMIM, un des changements importants a été d'incorporer l'utilisation des données fiscales de la TPS afin de réduire le fardeau de réponse des petites entreprises (voir Trépanier, 2004, pour plus de détails sur ce remaniement). En ce qui concerne l'Examen de l'assurance de la qualité, notons d'abord que le système en place pour l'EMCGD était beaucoup plus à jour que celui de l'EMIM, du moins relativement à la technologie utilisée. Il n'en demeurait pas moins que l'EMCGD avait un grand besoin d'outils plus conviviaux, par exemple pour analyser la cohérence entre les enquêtes annuelles et mensuelles (y compris la fonctionnalité d'effectuer les révisions nécessaires), de même que pour pouvoir examiner de plus près les séries chronologiques produites pour l'EMCGD.

Finalement, l'EMSR fournit des estimations des ventes pour les restaurants, traiteurs et débits de boissons. Au moment du projet d'intégration des systèmes, l'Enquête venait tout juste de compléter un remaniement complet, tant au niveau méthodologique qu'informatique. Le système en place utilisait des technologies courantes, et, en grande majorité, les systèmes généralisés de Statistique Canada. Malgré le fait que cette enquête n'ait pas nécessairement fait l'objet d'un examen d'assurance de la qualité, elle constituait l'une des trois enquêtes mensuelles de la Direction de la statistique de l'industrie et, afin d'assurer une harmonisation complète, fit partie du projet d'intégration des systèmes. L'expertise acquise à la suite de son remaniement récent a permis d'assurer un développement du projet adéquat.

3.3 Déroulement global du projet

Le projet a débuté en 2008 et devait initialement durer deux ans. Toutefois, devant la complexité et l'ampleur de la tâche, l'échéancier a du être repoussé de deux ans, et ainsi voir en 2011 une première enquête utiliser le système nouvellement développé. Le projet peut se décomposer en quatre grandes phases comportant des défis très distincts. La section qui suit décrit ces quatre phases en soulignant les défis à relever et les leçons apprises.

4. Phases du projet d'intégration avec leurs défis et leçons apprises

4.1 Compréhension et documentation des outils et méthodes en place

Avant même d'entreprendre toute modification majeure d'un processus de l'ampleur de ceux utilisés dans les enquêtes mensuelles, il est important de d'abord dresser un portrait complet de la situation. Il faut donc décomposer le processus en composantes et s'assurer de bien comprendre l'importance de chacune. Il est aussi crucial de bien comprendre comment ces composantes interagissent entre elles afin de développer un nouveau système qui reproduit le même produit final de qualité, mais de façon plus efficace.

Un examen de révision du processus a été mené et représentait en quelque sorte le lancement du projet. Les trois systèmes en place pour les enquêtes ciblées ont été examinés de près et le modèle fonctionnel pour le nouveau système à développer a été défini. Le tout incluait donc une description complète des processus, la liste des fichiers d'entrée et de sortie pour chaque processus, de même qu'un diagramme reflétant le cheminement des données à travers les processus et comment tous ces fichiers étaient liés entre eux. L'exercice de révision a également pris soin de bien définir les rôles et responsabilités de chaque équipe de travail affectée à ces tâches, de souligner l'importance d'utiliser un langage technique commun, et d'inclure quelques recommandations techniques en ce qui a trait à la technologie informatique à utiliser.

Parallèlement, un examen complet des méthodes statistiques en place a été effectué. Un regard détaillé a été jeté sur les étapes de calendarisation des données déclarées, de vérification et d'imputation, puis d'estimation. Le constat général fut que les trois enquêtes utilisent des méthodologies relativement semblables en surface, mais montrent plusieurs différences dans les paramètres utilisés à l'intérieur même de ces méthodes. Les outils informatiques diffèrent aussi d'une enquête à l'autre. Le tableau 4.1-1 présente un extrait des conclusions de cet examen des processus méthodologiques.

Pour cette première phase du projet d'intégration des sytèmes, la principale leçon vient du constat fait lors de la phase d'intégration (décrite à la section 4.3), alors que les différents processus devaient être liés entre eux pour ainsi former un système complet et fonctionnel. La leçon consiste à déployer beaucoup de temps et d'efforts pour cette première phase de documentation, de façon à obtenir un portrait complet et détaillé des processus en place. Le projet d'intégration des systèmes avait pris soin de décortiquer les systèmes en place, mais en se concentrant surtout sur les grandes étapes de traitement. Les étapes intermédiaires, souvent réalisées en arrière-plan, mais permettant de lier certaines de ces étapes entre elles, avaient été moins examinées. Celles-ci se sont avérées un casse-tête assez complexe et ont nécessité du temps de développement supplémentaire. Un exemple d'une telle étape intermédiaire est celui de l'extraction des données fiscales à partir des bases de données centralisées, puis leur stockage temporaire afin d'être utilisées et manipulées par la suite en cours de traitement des enquêtes. Il était donc important de dédier une grande portion du calendrier prévu à cette première phase puisque des oublis dès le début pouvaient mener à des embûches plus loin en cours de route lorsque les efforts seraient à ce moment consacrés à d'autres tâches.

Tableau 4.1-1
Sommaire des conclusions tirées de l'examen des processus méthodologiques pour certaines étapes d'enquête

Étapes	Conclusion
Calendarisation des données déclarées	Les trois enquêtes ont une approche semblable, à l'exception de l'EMCGD qui utilise aussi un poids journalier. Les outils utilisés diffèrent et n'incluent pas de systèmes généralisés.
Vérification	Chaque enquête possède des règles qui déterminent si une valeur est trop grande ou incohérente par rapport à l'information historique ou auxiliaire disponible. Les approches générales sont semblables, mais utilisent des paramètres différents (p. ex. les règles de tolérance). Les outils utilisés sont différents; une enquête utilise déjà le système BANFF, alors que les deux autres prévoient l'adopter bientôt. La détection de valeurs aberrantes est aussi faite pour les trois enquêtes. L'EMIM utilise une méthode différente (fondée sur la distance de Mahalanobis) des deux autres enquêtes (Hidiroglou-Berthelot) par l'entremise d'outils différents.
Imputation	Des sources de données semblables sont utilisées pour faire l'imputation : données historiques (avec tendance), données provenant de la base de sondage ou des données fiscales, ou les données recueillies auprès des répondants (imputation par classe, par exemple via la moyenne). Lorsque des classes sont utilisées, elles sont construites de façon semblable à l'aide des variables de classification industrielle, d'information géographique, et une mesure de la taille. Les outils utilisés sont différents.
Estimation	Les méthodes sont semblables pour les trois enquêtes à l'exception de l'estimation pour la portion à tirage nulle, pour laquelle les trois enquêtes utilisent des données fiscales au moyen de différentes approches. Le système généralisé (SGE) est utilisé par les trois enquêtes.

4.2 Harmonisation des outils et méthodes

Le travail effectué lors de la première phase a permis de diriger les efforts d'harmonisation entamés par la suite. Puisque le projet visait principalement à harmoniser les systèmes, les experts en technologie de l'informatique ont pu alors examiner les possibilités en matière de technologie et des systèmes à utiliser. Statistique Canada détient et soutien une large gamme de produits généralisés, et le mot d'ordre était évidemment d'utiliser ceux-ci lorsqu'ils pouvaient bien sûr répondre aux besoins. Bref, lorsqu'une étape de traitement était accomplie par deux enquêtes par l'entremise de systèmes différents, l'approche utilisant un système généralisé était favorisée.

Comme mentionné précédemment, un objectif secondaire de ce projet consistait à harmoniser les méthodes statistiques utilisées. Les deux principaux changements méthodologiques sont survenus au niveau de l'harmonisation de la méthode de calendarisation des données déclarées, de même que pour la détection de valeurs aberrantes à l'étape de vérification. Pour la calendarisation, une méthode inspirée de celle utilisée pour le traitement des données fiscales a été adoptée. Celle-ci était déjà en quelque sorte en place pour l'EMSR et a donc servi de point de départ. L'utilisation de « poids » mensuels et journaliers a entre autres été instaurée de façon plus cohérente pour les trois enquêtes. En ce qui concerne la détection de valeurs aberrantes, la méthode en place pour l'EMIM à été changée. Cette enquête utilisait une méthode fondée sur la distance de Malahanobis, alors que les deux autres enquêtes utilisaient plutôt la méthode Hidiroglou-Berthelot (HB) (voir Hidiroglou et Berthelot, 1986). Puisque la méthode HB était employée dans le système BANFF, et qu'elle menait à des résultats très semblables à ceux observés avec la méthode de Malahanobis, elle fût adoptée pour l'EMIM.

Le principal défi rencontré durant cette phase était d'instaurer un changement de mentalité. Les trois enquêtes en place sont effectuées et traitées par trois équipes différentes. Malgré le fait que ces équipes travaillent toutes plus ou moins à l'intérieur de divisions communes à Statistique Canada, il s'agit quand même d'une approche fondée principalement sur une structure traditionnelle dite « en silos ». Le projet visait à instaurer une approche plus « globale » qui répondrait mieux aux besoins communs des enquêtes et permettrait la collaboration horizontale, l'intégration et le partage de solutions entre les équipes. Durant cette période d'harmonisation, les efforts étaient essentiellement dirigés vers l'adoption de solutions et d'outils existants, ce qui pourrait dans certains cas être perçu comme une entrave à l'innovation, une dimension souvent plus attirante et motivante du travail sur les enquêtes.

Cette interruption à l'innovation n'allait toutefois être que temporaire, et devait alors à partir de ce moment se faire de façon « globale ».

4.3 Développement et intégration des composantes du nouveau système

La troisième phase visait à prendre les processus ou outils déterminés précédemment et développer ce qui allait être le nouveau système. Initialement, le plan était de développer ce système en utilisant une mécanique totalement nouvelle, tout en incorporant le bon outil lorsqu'il était requis de le faire. Par exemple, pour l'étape de vérification et d'imputation des données, BANFF était l'outil de prédilection. Il fallait donc maintenant s'assurer de fournir les spécifications propres à chaque enquête (les « paramètres ») mais à l'intérieur d'un même processus harmonisé, assurant ainsi que les mêmes fonctionnalités étaient disponibles pour les trois enquêtes. Pour la majorité des étapes d'enquête, ces outils ont été développés de cette façon. Toutefois, faute de ressources humaines et de temps, quelques processus déjà en place en production ont simplement été réutilisés tels quels et liés au nouveau système. Par exemple, les programmes d'estimation propres à chaque enquête ont été réutilisés, ne subissant que des modifications techniques permettant le raccordement au système intégré, assurant du même coup la bonne transmission des paramètres, le raccordement aux fichiers d'entrée, *etc*.

Comme mentionné plus haut, c'est aussi durant cette phase que certains processus intermédiaires importants ont été négligés lors de la première phase de documentation, et ils ont donc du être développés sur-le-champ. Ceci avait comme incidence d'augmenter la charge de travail déjà élevée des membres de l'équipe, et du même coup augmentait le risque d'erreurs.

4.4 Essais et mise en production

La dernière phase du projet consistait à faire de nombreux essais avant de passer éventuellement à l'étape de la mise en production. Une stratégie de mise à l'essai a été élaborée, passant d'abord à l'exécution de tests locaux (un processus à la fois, de façon indépendante) au test ultime qui constituait en un test en parallèle (production complète à la fois sur l'ancien et le nouveau système intégré). Ces essais permettaient d'assurer le bon déroulement de la production et de s'assurer que les mêmes résultats de qualité étaient obtenus. Il avait été convenu que les estimations produites en bout de ligne pourraient être différentes par rapport au système de production actuel puisque certaines modifications apportées lors de l'harmonisation des méthodes auraient un effet, non significatif, sur les chiffres produits. Les effets potentiels avaient toutefois été bien documentés, ce qui permettait ainsi de mieux comprendre et d'expliquer des différences étaient observées.

La mise en production officielle du nouveau système était prévue pour la fin 2011, début 2012. Au moment de l'écriture de cet article, l'EMSR, allait être la première enquête à effectuer la migration au nouveau système. Les deux autres enquêtes allaient en faire de même dans les mois suivants, tentant de coordonner le mieux possible cette transition avec d'autres étapes importantes du processus d'enquête (telle que la révision annuelle des estimations).

5. Sommaire du projet

En plus des défis à relever au cours des quatre phases d'intégration, certains aspects importants se sont ajoutés à la complexité de ce projet. Par exemple, le nouveau système a du être développé tout en maintenant la production mensuelle des enquêtes. Les échéanciers de production pour celles-ci sont très serrés, ce qui laissait très peu de temps pour mettre à l'essai le nouveau système en temps réel, afin de simuler une vraie production. De plus, la majorité du personnel affecté au développement du nouveau système était des personnes déjà affectées, de près ou de loin, à la production mensuelle. Cette tâche était donc d'un côté exigeante pour ces personnes en question, mais, d'un autre côté, nécessaire afin d'assurer une bonne compréhension des systèmes en place, et de bien déterminer les exigences du nouveau système.

En conclusion, le développement d'un système de production intégré permet d'aborder plusieurs risques en place et besoins des équipes de production. Après tout, malgré un échéancier plus long que prévu, et un produit final moins généralisé que souhaité, ce projet représente un pas dans la bonne direction. À la suite de cette intégration d'enquêtes mensuelles, Statistique Canada a entrepris un projet de plus grande envergure visant à intégrer ses enquêtes-

entreprises annuelles au sein du Programme intégré de la statistique des entreprises. Ce Programme représentera en fait une version remaniée et élargie de l'Enquête unifiée auprès des entreprises déjà en place. Les leçons apprises lors du projet d'intégration des enquêtes mensuelles pourront donc être mises à profit pour l'organisme.

Bibliographie

Deguire, Y., Reedman, L. et M. Wenzowski (2011), « Systèmes généralisés : l'expérience de Statistique Canada », *Recueil du Symposium international de 2011 sur les questions de méthodologie*, Statistique Canada.

Hidiroglou, M. et J.-M. Berthelot (1986), « Contrôle statistique et imputation dans les enquêtes entreprises périodiques », *Techniques d'enquête*, vol. 12, no 1, p. 79 à 89, Statistique Canada.

Thomas, S. et K. Cook (2005), « Combiner des données administratives et des données d'enquête dans l'Enquête mensuelle sur les industries manufacturières », *Recueil du Symposium international de 2005 sur les questions de méthodologie*, Statistique Canada.

Trépanier, J. (2004), « The Redesigned Candadian Monthly and Retail Trade Survey: A Postmortem of the Implementation », *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Les multiples facettes de la refonte des applications permettant de produire les statistiques conjoncturelles de l'INSEE : le programme Premice

Fabien Guggemos[1]

Résumé

Le programme Premice de l'Insee a pour vocation de coordonner la mise en œuvre et le développement des projets de refonte d'applications informatiques gérant des indicateurs de court terme relatifs aux entreprises (indicateurs conjoncturels d'activité, indices de prix, indicateurs des enquêtes de conjoncture). Par la coordination des différents projets d'une part, l'offre ou le développement de services informatiques voire de modules statistiques mutualisés d'autre part, le programme Premice vise à améliorer la qualité et la cohérence des indicateurs conjoncturels, à sécuriser et rationaliser la production de ces derniers ainsi qu'à dégager des gains de productivité en développement puis en maintenance..

Mots clés : GSBPM ; indicateurs conjoncturels ; mutualisation ; processus statistiques ; services informatiques communs.

1. Introduction : Genèse et mise en place du programme Premice

L'idée de rapprocher les modes de production des indicateurs conjoncturels de l'Insee (Institut national de la statistique et des études économiques, France) est apparue dès le milieu des années 90 ; à l'époque fut développé un ensemble - nommé Propice - d'outils statistiques organisés dans une application informatique et destinés à servir simultanément aux applications permettant de produire les indices de production industrielle, de commandes, de chiffres d'affaires et de gérer l'enquête sur les produits, les charges et les actifs. Propice partait du constat d'une grande similitude entre les objets manipulés et les fonctions à assurer. À l'époque, le choix a été fait d'utiliser le langage SAS et de créer des catalogues de « macro-programmes ». Si ces outils sont encore actifs aujourd'hui, ils ne sont pas exempts de défauts. Les chaînes ne sont ni modulaires, ni interruptibles ; elles proposent des traitements complets depuis le contrôle des données en entrée jusqu'à la mise en forme des résultats, mais avec une durée d'exécution comprise entre trois heures et une nuit. Le manque de souplesse de ces chaînes a finalement conduit au développement de programmes en libre-service qui doublonnent des parties complètes de l'application.

Plus récemment, le projet CRPI (Collecte et Retour Par Internet) de l'Insee a construit un outil de collecte par Internet largement utilisé dans le monde des indicateurs conjoncturels. Né de la fusion en 2001 de trois projets (collecte internet pour les enquêtes de conjoncture et les prix industriels ainsi qu'un portail entreprises), l'ouverture du site entreprises.insee.fr a été lancé fin 2003. L'application CRPI est opérationnelle et a permis de proposer aux entreprises la réponse en ligne à plusieurs enquêtes, pilotées par diverses directions de l'Institut. La maintenance comporte une large composante évolutive et les moyens sont donc pour une très large part consacrés à l'investissement sur de nouvelles enquêtes. Pour autant, le coût d'intégration des nouvelles enquêtes est souvent considéré comme trop important pour les applications potentiellement candidates. En particulier, il est nécessaire de retravailler à la fois l'enquête et l'outil CRPI pour construire le dispositif.

Ces différents projets ont permis d'acquérir de l'expérience tant sur la façon de mettre en œuvre avec succès des projets interdirectionnels que sur la proximité des indicateurs conjoncturels. Pour ce qui est du premier point, le CRPI était le premier projet de cette nature : la collaboration a toujours été satisfaisante… bien que la question du propriétaire de l'application n'ait jamais été tranchée. Quant au second point, il a notamment été confirmé par une étude d'urbanisation, menée début 2008 à la demande de la direction de l'Institut, sur l'ensemble du système de statistiques conjoncturelles. Celle-ci a ainsi permis de dresser le triple constat suivant :

[1]Fabien Guggemos, Institut national de la statistique et des études économiques (INSEE), 18, boulevard Adolphe Pinard, 75675 Paris cedex 14, France (fabien.guggemos@insee.fr).

- les applications permettant de gérer les indicateurs de court terme s'avèrent pour la plupart dépassées, car développées avec des technologies et des logiciels anciens voire obsolètes ;

- cette obsolescence engendre des risques non négligeables pour la production des indicateurs ainsi que des coûts élevés de fonctionnement, d'une part pour les équipes de maintenance, et, d'autre part, pour les utilisateurs, parfois contraints de réaliser en libre-service des « verrues » rendant les services non assurés par l'application ;

- les modes opératoires de ces applications au niveau macro se ressemblent fortement et les processus statistiques qui y sont à l'œuvre présentent d'importantes similitudes, si bien qu'une réflexion globale sur la rénovation de ces différents applicatifs semble opportune.

Dans un tel contexte, marqué par la concomitance du lancement des refontes des applications de production des indices de prix (projet Papaye) et de gestion des enquêtes de conjoncture (projet Conj2), tous les éléments sont réunis pour mettre en place une structure garantissant la coordination des projets et leur implication dans une démarche commune de mutualisation. C'est ainsi que le programme Premice, acronyme signifiant Programme de REfonte avec Mutualisation des Indicateurs Conjoncturels d'Entreprises, voit officiellement le jour au début de l'année 2009.

2. Les objectifs stratégiques du programme Premice

Le programme Premice a pour vocation de coordonner la mise en œuvre et le développement des projets de refonte d'applications informatiques gérant des indicateurs de court terme relatifs aux entreprises. La démarche de mutualisation qui le sous-tend est cependant envisagée « prudemment ». Cette dernière phrase est caractéristique de l'esprit du programme Premice. Conscient de la difficulté de conduire un unique projet et fort de l'expérience du projet Propice cité en introduction, Premice a fait le choix d'un positionnement d'appui, de facilitation des projets plus que de prescription et de normalisation. Il ne s'agit donc pas d'un cadre rigide pesant sur les projets, mais bien d'une démarche qui les accompagne. Aussi le succès de Premice se mesure-t-il à la qualité du service rendu aux utilisateurs plutôt qu'au nombre de modules développés en commun. En deux mots, le programme Premice concilie à la fois pragmatisme et ambition.

Le programme Premice vise la poursuite et l'atteinte des trois objectifs stratégiques suivants :

- améliorer la qualité et la cohérence des indicateurs conjoncturels ;
- sécuriser et rationaliser la production des chiffres ;
- dégager des gains de productivité tant en développement qu'en maintenance.

2.1 Améliorer la qualité et la cohérence des indicateurs conjoncturels

Les projets sont généralement l'occasion de s'interroger sur le processus d'élaboration des statistiques. En d'autres termes, au-delà de la création d'un nouvel outil informatique qui en constitue généralement le livrable le plus visible, un projet peut conduire à faire évoluer la méthodologie statistique ainsi que le processus de production des indices. Au sein du programme Premice, les équipes de projet statistiques (EPS) sont en contact permanent, ce qui favorise la communication et le partage d'expérience. Pour autant, il ne s'agit pas d'imposer des méthodes auxquelles les responsables des projets n'adhèrent pas.

L'une des critiques adressées à Propice, projet mentionné plus haut, était son caractère monolithique; la chaîne est lancée en une seule fois et prend un long temps avant de livrer ses chiffres. Or, dans le cadre de l'analyse conjoncturelle, une valeur ajoutée des responsables d'indices est de pouvoir proposer différents scénarios (par exemple, produire des indices, bruts ou désaisonnalisés, aux différents postes et niveaux de nomenclatures existants). L'une des ambitions de Premice est donc de maximiser le nombre de scénarios joués pour accroître la qualité des indicateurs produits, ce qui exige des traitements informatiques suffisamment souples et rapides. Le producteur doit disposer d'un certain nombre de leviers pour agir sur l'indicateur et ce rapidement afin de construire un nombre suffisant de scénarios. Il faut néanmoins prendre garde à maintenir un volume raisonnable de possibilités offertes et à sécuriser les traitements en offrant une forte traçabilité.

Le cadre de réflexion commun apporté par Premice vise également à améliorer la cohérence entre les différents indicateurs conjoncturels. Une telle cohérence est en effet nécessaire lors de la confrontation des divers indices, exercice pratiqué notamment lors de l'élaboration des comptes trimestriels. L'analyse comparée d'indicateurs conjoncturels désaisonnalisés selon des méthodes distinctes ou ne reposant pas sur les mêmes nomenclatures ne s'avèrerait guère pertinente. En explicitant les méthodes de production de chaque indicateur, chacune des équipes de projet peut se positionner et de ce fait adopter une stratégie de production de son propre indicateur compatible avec celle des autres indices.

La preuve la plus tangible de la cohérence des traitements est la mise en œuvre des mêmes outils. Si la mise en place de modules communs peut être difficile, il existe différents niveaux de mutualisation; cela va de la vision commune d'un traitement (utiliser le même vocabulaire, la même description pour une tâche donnée) au traitement commun (effectuer la même action). L'objet de Premice n'est pas de forcer l'adoption de tel ou tel outil, mais bien d'offrir aux producteurs la capacité technique de mettre en œuvre les meilleures pratiques.

Il s'agit finalement de passer d'une organisation en tuyaux d'orgue à une politique d'échanges nourris au sein d'une communauté. Outre la qualité des indices, on peut espérer pour les agents une mobilité plus aisée entre les différents indicateurs et l'acquisition d'une expertise forte sur ces domaines.

2.2 Sécuriser et rationaliser la production des chiffres

La production des indicateurs conjoncturels d'entreprises est une opération très fréquente dans l'échelle des opérations statistiques; le plus souvent, le rythme de parution des indicateurs est mensuel. L'enjeu est donc d'être en mesure de produire ces chiffres chaque mois au jour et à l'heure prévus. Avec les applications de Premice, les différents indicateurs devraient voir leur temps de traitement significativement diminuer. Plus que la réduction du délai de publication, c'est la qualité du chiffre qui devrait bénéficier de la rapidité des traitements, en laissant une plus grande part à l'analyse des indicateurs et de leurs révisions.

Par ailleurs, les délais de mise en œuvre du processus sont également courts; il faut donc faire preuve d'une grande efficacité et éviter autant que faire se peut toute action inutile. Dans le cadre de Premice, l'objectif est de mutualiser les expériences afin d'en retirer tous les éléments permettant de produire au plus vite les indicateurs avec une norme de qualité suffisante. Encore une fois, les gains attendus sont en termes de qualité de la phase d'analyse.

Enfin, une sécurisation complète implique de pouvoir réagir rapidement à des évènements divers non anticipés. L'un des objectifs de Premice est d'offrir la qualité de service d'une application avec la souplesse du libre-service. Parmi les éléments indispensables à l'ouverture des applicatifs aux statisticiens, la traçabilité est renforcée et facilitée. La production informatique des indices sera ainsi assurée par des outils actualisés et aisément maintenables, tandis que l'ensemble des chaînes de production sera documenté de façon détaillée.

2.3 Dégager des gains de productivité tant en développement qu'en maintenance

Le développement de modules communs nécessite initialement un investissement important pour s'assurer que les services rendus répondront bien aux besoins des différents projets. Ce potentiel surcoût est cependant contrebalancé par de nombreux avantages. En abordant de manière coordonnée un certain nombre de problématiques relatives à la production d'indicateurs conjoncturels et en multipliant les relations entre équipes, chaque projet profite des réflexions de l'ensemble des autres projets pour converger, le cas échéant, vers des solutions communes ou approchantes. De facto, les échanges entre projets contribuent à définir une même culture et un même langage dans le monde des indicateurs conjoncturels. Sur un plan plus concret, le développement d'un module commun permet d'éviter que des fonctionnalités identiques entre deux projets ou plus soient programmées à plusieurs reprises. Un projet peut ainsi profiter des fonctionnalités d'un module développé par un autre projet, ou les adapter au besoin en fonction de ses propres spécificités.

Le schéma modulaire que doivent adopter autant que faire se peut les projets du programme Premice contribue également à l'amélioration de la qualité des applicatifs. Il s'agit de réduire les coûts de maintenance de ces derniers et de leur offrir des possibilités de rénovation partielle. Un composant qui ne donne plus pleinement satisfaction pourra ainsi être modifié dans un laps de temps plus court et être rebranché en lieu et place de l'ancien selon un calendrier plus compatible avec les exigences des maîtrises d'ouvrage. Outre les gains liés au fait que la maintenance

d'un module commun, assurée par une seule équipe, profitera à l'ensemble des applications, le schéma modulaire facilite grandement la mise en œuvre de la mutualisation en évitant des chaînes de traitement linéaires et non interruptibles.

En définitive, le développement de modules communs (pour lesquels les maintenances seront centralisées) et la proximité forte entre les architectures informatiques (qui favorise une culture commune au sein des équipes de développement et diminue le coût de formation des nouveaux arrivants) constituent autant de gages de pérennité pour l'avenir des applications de Premice que de facteurs de gains de productivité, enjeu fort pour l'Institut.

3. Le périmètre du programme Premice

Le programme Premice est mis en place fin 2009 à l'occasion de la refonte des applications de production des indices de prix (projet Papaye) et de gestion des enquêtes de conjoncture (projet Conj2). Ce programme est susceptible d'accueillir ensuite, au fur et à mesure du lancement de leurs refontes respectives, l'ensemble des applications permettant la production d'indicateurs de court terme relatifs aux entreprises. C'est ainsi que la rénovation des systèmes de production des indices de chiffres d'affaires (projet Harmonica) et des indices de production industrielle (projet Ocapi) ont intégré le programme lors de leurs lancements respectifs fin 2010 et fin 2011. Au final, l'ensemble du périmètre concerné est le suivant :

- **indicateurs d'activité**
  - indices de production industrielle - *(Projet Ocapi, de fin 2011 à 2015)*
  - indices de chiffres d'affaires - *(Projet Harmonica, de fin 2010 à 2014)*
  - indices de commandes

- **indices de prix -** *(Projet Papaye, de fin 2009 à 2013)*
  - indices de prix à la production, prix à l'importation
  - indices de prix agricoles

- **enquêtes de conjoncture -** *(Projet Conj2, de fin 2009 à mi-2014)*

- **autres indicateurs**
  - indices du coût de la construction, de référence des loyers, des loyers commerciaux
  - enquête mensuelle auprès des grandes surfaces alimentaires

4. Quatre dimensions opérationnelles pour le programme Premice

Les trois objectifs stratégiques se déclinent selon quatre dimensions opérationnelles relatives à l'organisation générale du programme ainsi qu'à sa traduction concrète en termes informatiques et statistiques :

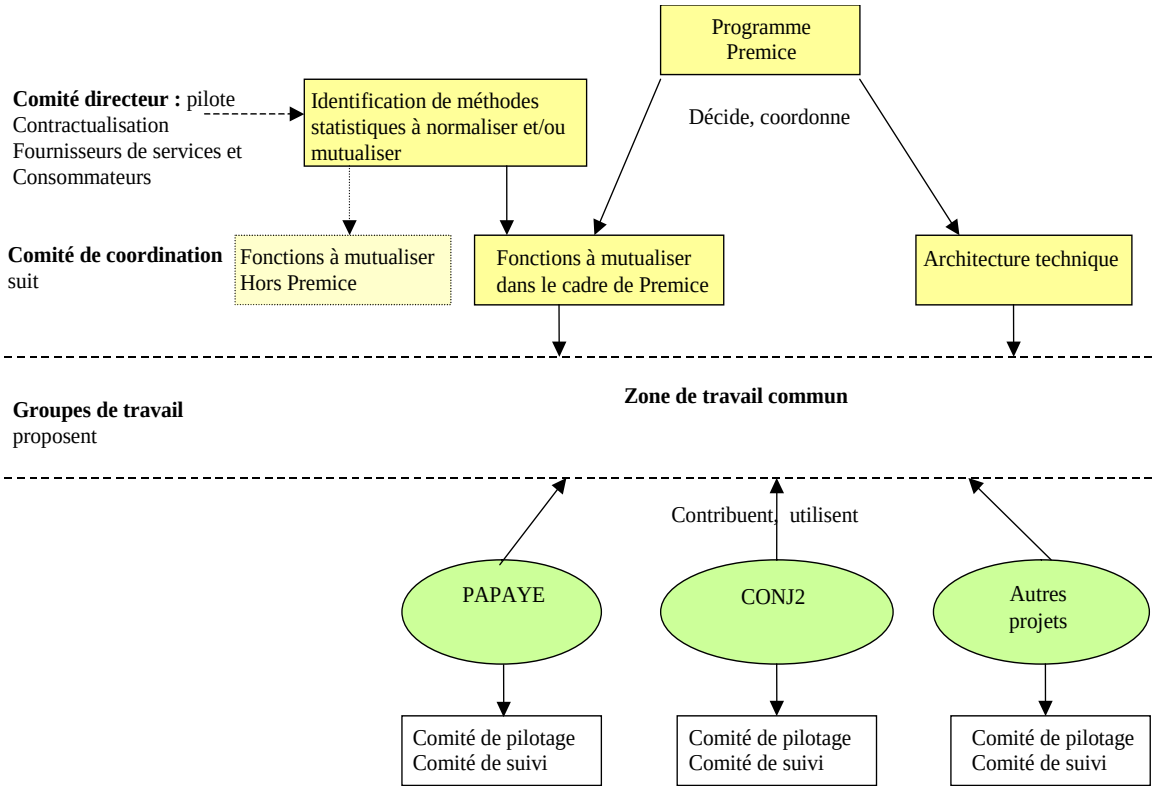
- définir une stratégie coordonnée pour l'ensemble du programme Premice ;
- articuler la conception et la réalisation des applicatifs du périmètre du programme ;
- mettre en place les services informatiques susceptibles d'être utilisés conjointement par les applications relevant du programme ;
- mettre en œuvre les processus statistiques susceptibles de faire l'objet d'une mutualisation voire d'une normalisation.

4.1 Fonctionnement du programme Premice

Définir une stratégie coordonnée pour l'ensemble du programme Premice

Le programme Premice vise en premier lieu à définir des principes de développement des projets concernés. Pour chacun des systèmes, l'identification des fonctions pérennes et stables dans le temps et des fonctions susceptibles d'évoluer régulièrement ainsi que la mise en évidence de caractéristiques communes (notamment la nécessité de disposer d'outils permettant des calculs fréquents et rapides) est un préalable à la mise en oeuvre d'une architecture et de composants techniques communs. Cette organisation informatique commune doit permettre de répondre aux besoins de réactivité et d'évolutivité des futures applications, de sécuriser leur fonctionnement et de diminuer les charges tant de développement que de maintenance. Au-delà de cette organisation sont identifiés les services informatiques susceptibles d'être utilisés conjointement ainsi que les modules statistiques pouvant faire l'objet d'une mutualisation. Il revient au comité directeur, véritable instance décisionnelle du programme, de décider des investissements mutualisés. Chargé plus généralement de veiller au bon déroulement du programme, il décide en effet des services et modules à mutualiser, de la répartition des travaux de développement de ces derniers entre les différentes équipes de projet et enfin de la désignation d'un propriétaire - responsable de la maintenance - pour chacun de ces services et modules communs. Le comité directeur prend les arbitrages stratégiques du programme en cas de désaccord entre les maîtrises d'ouvrage des divers projets. Ses décisions sont instruites et préparées par un comité de coordination. Lieu privilégié d'échanges entre les différentes maîtrises d'ouvrage, ce dernier pilote et coordonne les travaux opérationnels, en créant et suivant les groupes de travail mis en place pour identifier puis étudier en détail les possibilités de mutualisation entre les applications.

Figure 4.1-1
Schéma de fonctionnement du programme Premice



4.2 Articulation des différents projets

Articuler la conception et la réalisation des applicatifs du périmètre de Premice

La conception et la réalisation des services à mutualiser sont confiées à l'un ou l'autre des projets, selon la disponibilité des équipes. Les autres projets disposent ensuite directement de ces services et peuvent éventuellement les adapter à la marge selon leurs besoins particuliers. Par ailleurs, le déroulement concomitant des projets permet un dialogue constructif entre les équipes. Ces échanges se traduisent par une homogénéisation significative sur de nombreux aspects (macroanalyse, méthodes de travail - par exemple, l'adoption d'une grille standardisée d'analyse des processus, issue de modèles internationaux tels que le *Generic statistical Business Process Model* (GSBPM) - etc.). Pour autant, chaque projet participant au programme Premice conserve son identité et son autonomie; il possède par conséquent ses propres objectifs tout en utilisant l'architecture et la plate-forme de développement communes.

Aussi a-t-il été clairement posé qu'un projet participant à Premice pouvait ne pas utiliser des modules mutualisés dès lors que ceux-ci ne rendent pas le service attendu par le projet. Si Premice a la dimension d'un programme, il n'en a pas nécessairement toutes les caractéristiques. Les ressources dédiées exclusivement à Premice s'avèrent d'ailleurs particulièrement réduites (un unique coordinateur statistique et un demi équivalent temps plein sur les aspects techniques et informatiques). Ce sont les projets qui disposent des moyens, à la mesure de leur complexité respective; un projet participant à Premice peut donc refuser des demandes relatives à Premice susceptibles de mettre en danger son bon déroulement.

Malgré ses moyens dédiés réduits, le programme Premice produit plusieurs livrables, de deux natures distinctes :

- des livrables statistiques; Premice fournit un cadre de réflexion sur l'univers des indicateurs conjoncturels ; différents groupes de travail se succèdent sur des thèmes choisis pour leur potentialité de mutualisation ;

- des livrables informatiques, pouvant être le pendant des livrables statistiques (car certains groupes de travail ont des thèmes informatiques) ou bien concrétisant les livrables statistiques sous forme d'outils et de services construits puis utilisés par les projets.

Premice est donc un objet relativement nouveau dans le paysage des applications et des projets de l'Institut. Il ne répond pas à une problématique d'outillage d'une de ses opérations statistiques, mais il accompagne la rénovation d'un quartier de son système d'information.

4.3 Mise en place de services informatiques communs

Mettre en place les services informatiques susceptibles d'être utilisés conjointement par les applications relevant du programme Premice

L'étude d'urbanisation évoquée en introduction a conduit à préconiser une utilisation de services informatiques communs, notamment dans les phases « amont » et « aval ». Les services offerts peuvent être de différentes natures; ce peut être un module intégré dans l'application ou dans une application externe avec laquelle l'application va échanger des informations.

Ainsi le CRPI, pour la gestion de la collecte par Internet, met-il déjà en œuvre un service commun pour la collecte des données entre enquêtes sur les prix de production et les enquêtes de conjoncture, qui préserve une certaine étanchéité entre les processus. D'autres services ont été identifiés et la décision a été prise en comité directeur d'instruire la mise en œuvre de ces services. Il s'agit de la plate-forme unique d'accueil des données, des opérations de diffusion des résultats, notamment vers la banque de données macro-économiques de l'Insee, du dialogue avec le futur répertoire statistique d'entreprises de l'Institut (tirages d'échantillon, signaux sur la démographie des entreprises). Outre des gains de productivité évidents lors du développement, la mise en place de services communs doit permettre de disposer d'applicatifs plus homogènes, donc a priori de meilleure qualité et plus aisés à maintenir.

4.4 Mutualisation de processus statistiques

Mettre en œuvre les processus statistiques susceptibles de faire l'objet d'une mutualisation voire d'une normalisation

Le programme Premice s'articule autour de l'identification de modules, sous-parties du processus statistique, susceptibles d'être mutualisés. Les projets incorporent ces modules dans le cadre des développements informatiques ou prévoient au moins la possibilité de les prendre en compte lorsqu'ils seront prêts. Une liste de ces modules dits « mutualisables » a été établie grâce à la comparaison des processus de production des divers indices, chacun d'eux ayant fait ensuite l'objet d'études plus détaillées. Plusieurs domaines potentiellement mutualisables ont ainsi été identifiés : l'analyse des révisions des indicateurs, les méthodes de contrôle et de redressement des données, l'application des règles de confidentialité, les techniques de correction des variations saisonnières et des jours ouvrables, *etc*. Afin d'évaluer précisément les possibilités de mutualisation, des études ont ainsi été conduites dans le cadre de groupes de travail rassemblant des membres des différentes équipes de projet. Le partage de bonnes pratiques statistiques permet dès lors de renforcer la cohérence des indicateurs produits et d'améliorer la qualité globale du diagnostic conjoncturel.

5. Conclusion

Le programme Premice - qui s'étend ainsi jusqu'en 2015 - est ambitieux et mobilisateur pour les équipes de projet. Impliquant plusieurs directions de l'Institut, il avance dans un climat de confiance mutuelle entre les différents et nombreux acteurs. L'organisation même du programme, évoquée dans la partie 4, est un gage de bonne conduite; en effet, le succès de Premice passe d'abord par la réussite des projets qui le composent. La visibilité sur le programme est donc de l'ordre des projets. À ce titre, la démarche de mutualisation demande un investissement et un accompagnement constants. Pour les aspects statistiques, les groupes de travail jouent ce rôle, tandis que pour les aspects informatiques, la mutualisation se fait à différents niveaux, depuis la boîte à outils du développeur jusqu'à la réponse mutualisée aux problèmes spécifiques du domaine des indicateurs conjoncturels. Les premières réalisations concrètes commencent d'ailleurs à voir le jour avec des prototypes d'intégration de progiciels d'ajustement saisonnier et de gestion de la confidentialité. Si le chemin doit être plus facile en passant par Premice, il faut aider les projets à franchir certains obstacles et rentabiliser par la suite les investissements. En permanence, la question est posée de savoir quelle est la meilleure réponse collective ; c'est bien là le souci premier des instances de gouvernance de Premice.

Bibliographie

METIS Steering Group (2009), « Generic Statistical Business Process Model », version 4.0, www.unece.org/stats/gsbpm.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Normalisation des enquêtes-entreprises infra-annuelles au Royaume Uni

Salah Merad et Pete Brodie[1]

Résumé

Comme d'autres pays membres de l'Union européenne, le Royaume-Uni a dû adopter une version mise à jour de sa classification type des industries (SIC pour Standard Industrial Classification, équivalente à la NACE), appelée SIC 2007, pour ses enquêtes-entreprises infra-annuelles réalisées à partir de 2010. L'organisme a profité de cette occasion pour examiner le processus d'enquête complet, des exigences des clients à la publication des données. En particulier, il a été décidé de fusionner toutes les enquêtes mensuelles portant sur des secteurs particuliers en une enquête-entreprise mensuelle unique (MBS pour Monthly Business Survey). À la première étape, deux enquêtes seulement, portant sur le secteur de la production et celui de la distribution et des services, ont été harmonisées, normalisées et traitées ensemble. D'autres enquêtes, dont celles sur les secteurs du commerce de détail et de la construction, ont adopté le nom MBS, mais leur harmonisation avec les deux premières sera entreprise à une date ultérieure.

Dans le présent article, nous décrivons les divers aspects de l'harmonisation et de la normalisation, y compris la conception du questionnaire, la conception et la répartition de l'échantillon, les stratégies d'obtention des réponses, les règles de validation des données, les classes d'imputation et les méthodes d'estimation. Nous discutons également des problèmes opérationnels qu'il a fallu surmonter durant le processus de mise en œuvre et examinons les leçons tirées de notre expérience.

Mots clés : Méthodes normalisées ; plan de sondage ; intégration ; fardeau de réponse.

1. Introduction

1.1 Contexte

Jusqu'en 2010, l'Office for National Statistics (ci-après nommé l'organisme) réalisait un ensemble d'enquêtes mensuelles couvrant divers secteurs économiques, à savoir la production, la distribution et les services; le commerce de détail; et la construction. Les enquêtes sur la production et sur la distribution et les services recueillaient des données sur le chiffre d'affaires, l'enquête sur le commerce de détail recueillait la valeur des ventes, tandis que l'enquête sur la construction recueillait des données sur la valeur de la production pour différents types de travaux. Des données étaient également obtenues pour d'autres variables. Des données sur l'emploi étaient recueillies mensuellement dans certaines enquêtes et trimestriellement dans d'autres, et des données sur les nouvelles commandes et les exportations étaient recueillies pour le secteur de la production.

Les quatre enquêtes, dont les antécédents étaient différents, avaient été élaborées indépendamment; par conséquent, différentes méthodes étaient utilisées, même quand l'information recueillie était très semblable, comme l'emploi et le chiffre d'affaires dans les enquêtes sur la production et sur la distribution et les services. Il existait un certain degré d'intégration en ce qui a trait à l'usage d'une base de sondage commune et à la sélection d'échantillons coordonnés, mais l'organisme considérait qu'il était possible d'intégrer et de normaliser davantage les enquêtes. Un document en vue d'établir la portée du projet, décrivant les objectifs et les avantages a été rédigé en 2004, mais les travaux n'ont pas été entrepris en raison d'autres priorités à l'époque.

[1]Salah Merad, Office for National Statistics, Government Buildings, Cardiff Road, Newport, UK, NP10 8XG (salah.merad@ons.gsi.gov.uk); Pete Brodie, même affiliation et adresse en tant que premier auteur (pete.brodie@ons.gsi.gov.uk).

1.2. Motifs de la normalisation

Comme d'autres pays membres de l'Union européenne, le Royaume Uni a dû adopter une version mise à jour de sa classification type des industries (SIC pour Standard Industrial Classification, équivalente à la NACE), appelée SIC 2007, pour ses enquêtes-entreprises infra annuelles réalisées à partir de 2010. Certaines activités inscrites dans le secteur de la production dans l'ancien système de classification ont été classées dans celui des services dans le nouveau système, et inversement, certaines activités sont sorties du champ d'observation des enquêtes tandis que de nouvelles y ont été introduites. Dans le cadre du processus d'adaptation à la nouvelle classification, l'organisme a décidé d'examiner de nombreux aspects des opérations d'enquête et de profiter de cet examen pour intégrer davantage les enquêtes conjoncturelles et de normaliser leurs méthodes et processus. Cette intégration devrait réduire au minimum les problèmes opérationnels causés par le déplacement de certaines entreprises d'un secteur à un autre et réduire l'impression d'une augmentation du fardeau de réponse. En particulier, l'organisme a décidé de fusionner toutes les enquêtes mensuelles propres à un secteur en une enquête mensuelle auprès des entreprises (MBS pour Monthly Business Survey) unique. À la première étape, deux enquêtes seulement, couvrant le secteur de la production et celui de la distribution et des services, ont été entièrement intégrées, quoique la normalisation de certaines méthodes ait été étendue aux enquêtes sur le commerce de détail et sur la construction; l'harmonisation et l'intégration complètes de ces enquêtes seront envisagées à une date ultérieure. Le remaniement de la série statistique sur l'emploi (Workforce Jobs) avait lieu simultanément; comme les données sur l'emploi recueillies au moyen des enquêtes conjoncturelles représentent la source principale pour la production de ces statistiques, la normalisation des méthodes utilisées dans ces enquêtes devrait produire des statistiques plus comparables entre industries.

En plus des considérations de qualité, la normalisation des méthodes et des processus devrait permettre de réaliser des économies : l'exécution d'un plus petit nombre d'enquêtes devrait réduire les coûts des opérations et de la maintenance, tandis que l'adoption de processus normalisés devrait permettre aux membres du personnel de s'occuper de divers secteurs en recevant un minimum de formation supplémentaire.

Dans le présent article, nous décrivons les travaux de remaniement exécutés tout au long du processus d'enquête, de la définition de la portée de l'enquête jusqu'à la publication, auxquels nous donnons à l'organisme le nom de « chaîne de valeur statistique ». L'article s'inspire fortement des articles de James (2010) et de Taylor et coll. (2011). À la section 2, nous décrivons les travaux de remaniement tout le long de la chaîne de valeur statistique, et à la section 3, nous discutons des difficultés qu'il a fallu résoudre à l'étape de la mise en œuvre. À la section 4, nous énumérons les principales leçons tirées de nos efforts de normalisation et discutons des travaux courants et à venir.

2. Remaniement tout au long de la chaîne de valeur statistique

2.1 Principes et contraintes de remaniement

Un objectif important du travail de remaniement était de présenter aux répondants une seule enquête visant à recueillir des données économiques mensuelles, de sorte que l'organisme a décidé de donner à toutes les enquêtes mensuelles un nouveau nom uniformisé. Le choix s'est porté sur Monthly Business Survey (enquête-entreprise mensuelle). Parallèlement, l'Annual Business Inquiry, qui est l'enquête structurelle auprès des entreprises de l'organisme, a été renommée Annual Business Survey (enquête-entreprise annuelle). Celui-ci a également décidé d'éliminer toutes les différences inutiles entre les méthodes et les processus.

Un examen distinct des statistiques de l'emploi assorti d'une consultation a mené à la conclusion que les données sur l'emploi n'étaient nécessaires que trimestriellement; d'où la décision de normaliser la fréquence de la collecte des données sur l'emploi dans tous les secteurs. De surcroît, comme les données sur l'emploi sont moins instables que celles sur le chiffre d'affaires, l'organisme a décidé qu'un sous échantillon seulement des entreprises recevrait les questions sur l'emploi tous les trimestres, ce qui devrait réduire le fardeau de réponse. La mise en œuvre du sous échantillonnage a été relativement facile, car il était déjà utilisé dans l'enquête sur les services. L'organisme a également décidé d'utiliser des méthodes communes de vérification, d'imputation et d'estimation dans la mesure du possible et, au besoin, d'établir les valeurs des paramètres en se servant des mêmes critères pour tous les secteurs.

Nous étions soumis à un certain nombre de contraintes : la taille totale d'échantillon devait rester la même et nous ne pouvions exécuter une répartition conjointe de l'échantillon que pour les secteurs de la production et des services. Les exigences de qualité des programmes d'enquête ne pouvaient pas être toutes satisfaites au moyen de la taille d'échantillon disponible de sorte qu'il a fallu faire des compromis. Enfin, tous les travaux devaient être achevés et tous les changements, mis en œuvre en janvier 2010, au moment où aurait lieu le changement de classification des industries, ce qui limitait le nombre de méthodes pouvant être étudiées et le temps consacré à leur évaluation en vue de faire un choix.

2.2 Descriptions brèves des travaux de remaniement

2.2.1 Portée de l'enquête

En raison de l'adoption du nouveau système de classification (SIC 2007), la portée de l'enquête a été étendue. Par exemple, l'« architecture paysagère » a été transférée du secteur de l'agriculture (hors du champ des enquêtes conjoncturelles) au secteur des services. En outre, l'organisme souhaitait depuis longtemps procéder à la collecte directe du chiffre d'affaires dans certaines industries pour lesquelles les données provenaient d'une autre source, un exemple étant les « activités vétérinaires ». Certaines de ces industries ont été introduites dans la MBS en 2010, tandis que d'autres le seront plus tard, afin d'avoir plus de temps pour mettre le questionnaire à l'essai. Dans les industries pour lesquelles seules les données sur l'emploi sont nécessaires, parce que celles sur le chiffre d'affaires proviennent d'une source externe, une autre enquête, appelée Quarterly Business Survey (enquête-entreprise trimestrielle), a été lancée.

L'une des très longues tâches du remaniement a consisté à se mettre d'accord sur les groupes d'industries de la SIC pour lesquels des données de sortie seraient publiées. Les gestionnaires des données de sortie sur le chiffre d'affaires et sur l'emploi, qui représentaient les clients externes, ainsi que des représentants des principaux clients internes (par exemple, Comptes nationaux) et de la Méthodologie ont participé aux discussions en vue d'établir les groupes de publication. Sur le plan méthodologique, une question importante était de réduire le niveau de détail, car les groupes de publication constituent le fondement de la stratification; or, un trop grand nombre de strates entraîne une répartition inefficace de l'échantillon. Pour le secteur du commerce de détail, le nombre de groupes de publication est resté le même, soit 27, mais pour les secteurs de la production et des services, il a été réduit, pour passer d'un peu plus de 300 à 150. Pour les industries pour lesquelles l'organisme ne recueille que des données sur l'emploi, le nombre de groupes est passé d'environ 40 à environ 30 dans l'enquête-entreprise trimestrielle.

Une autre consultation a été menée en vue de déterminer quelles variables n'étaient plus nécessaires et quelles composantes il fallait inclure dans la définition du chiffre d'affaires. Cet exercice s'est soldé par l'abandon d'un certain nombre de questions particulières à des industries et par la réduction du nombre de types de questionnaires qui est passé de 65 à 26.

2.2.2 Conception du questionnaire

Comme nous l'avons mentionné plus haut, un nom unique, enquête-entreprise mensuelle (Monthly Business Survey ou MBS) a été adopté pour toutes les enquêtes mensuelles sur le « chiffre d'affaires ». Cependant, pour le secteur du commerce de détail, une brève description supplémentaire a été ajoutée, à savoir enquête-entreprise mensuelle – Indice des ventes au détail (Monthly Business Survey - Retail Sales Index). Le programme de l'enquête jugeait cet ajout important pour assurer la continuité avec l'ancien nom (Services - Retail Sales Index). Arriver à s'entendre sur le nom de l'enquête n'a pas été facile et des discussions approfondies ont eu lieu entre les programmes d'enquête et la Méthodologie au sujet des répercussions éventuelles sur la réponse. En fin de compte, il a été convenu d'ajouter des notes d'enquête particulières au secteur à la page 1 du questionnaire et une description de l'objet de l'enquête à la page 2. Par exemple, les notes pour le secteur de la production sont les suivantes [traduction] :

L'information que vous fournissez contribue à l'établissement de l'indice de la production (IP), qui montre les variations de la production du secteur de la fabrication et est une mesure clé de la contribution des entreprises manufacturières à l'économie. La Banque d'Angleterre et le Trésor de Sa Majesté utilisent le PIB et le PI comme indicateurs clés pour surveiller et prévoir la croissance économique et pour permettre de prendre des décisions

stratégiques éclairées. Ces mesures sont également utilisées par diverses associations commerciales pour faire des comparaisons internationales.

Un examen général des questionnaires a été effectué pour s'assurer que chaque industrie recevait un questionnaire approprié. Dans de nombreux cas, le questionnaire existant sous l'ancienne classification pouvait être utilisé, mais dans d'autres, la liste des composantes à inclure et à exclure dans le chiffre d'affaires devait être modifiée ou certaines questions devaient étaient supprimées. Ce dernier élément offrait un avantage supplémentaire : pour les types de questionnaires dans lesquels le nombre de questions était faible, l'entrée de données par téléphone (EDT) est devenue le mode de collecte par défaut. En outre, le petit nombre de types de questionnaires a facilité l'expansion de la portée de l'EDT.

Les questionnaires pour les industries se trouvant nouvellement dans le champ de l'enquête ont été mis à l'essai par une méthode cognitive auprès d'un petit échantillon d'entreprises obtenu par choix raisonné. Un aspect important qui n'a pas été pris en considération dans le cadre de ces travaux est l'harmonisation de la question sur le chiffre d'affaires entre les enquêtes mensuelle et annuelle. Cet exercice est envisagé dans le cadre d'un projet d'harmonisation plus vaste en cours au sein de l'organisme.

2.2.3 Plan d'échantillonnage et répartition de l'échantillon

Pour chaque enquête, l'échantillon a été stratifié selon le groupe d'industries et la taille de l'effectif, le groupe d'industries ayant tendance à correspondre au niveau le plus détaillé du système de classification des industries. Dans la MBS, il a été convenu d'utiliser, d'une façon générale, les groupements employés par les Comptes nationaux, ce qui a réduit considérablement les groupes d'industries dans les secteurs de la production et des services. Dans ces deux secteurs et dans celui du commerce de détail, quatre intervalles de taille d'effectif ont été utilisés dans chaque groupe d'industries, et la strate contenant les entreprises les plus grandes a été dénombrée entièrement. Différents jeux de bornes d'intervalle de taille ont été utilisés dans chaque enquête et, dans certaines d'entre elles, plus d'un jeu a été employé. Il serait certes commode d'un point de vue opérationnel de n'utiliser qu'un seul jeu de bornes d'intervalle de taille pour tous les volets de la MBS, mais cela ne serait pas approprié; par exemple, les entreprises les plus grandes du secteur de la production ont tendance à être plus petites que celles du secteur des services.

Pour des raisons pratiques, le nombre total de jeux d'intervalles de taille a été limité à huit; par conséquent, nous avons dû déterminer les huit meilleures combinaisons qui couvriraient toutes les industries dans la MBS. Diverses règles, dont celles de la « racine carrée cumulée de f » et de la « racine carrée cumulée de x », ont été utilisées pour obtenir les bornes optimales dans chaque industrie. Après avoir examiné les divers jeux de bornes, un compromis a été réalisé en ce qui concerne les huit meilleurs à utiliser. Pour les secteurs du commerce de détail et des services, en plus des quatre intervalles de taille, il existe une strate entièrement dénombrée pour les entreprises ayant un effectif petit ou moyen et un chiffre d'affaires élevé. Cet intervalle a été introduit pour les secteurs de la production et de la construction dans le cadre de la MBS.

La répartition de l'échantillon entre les strates a été effectuée conjointement pour les secteurs de la production et des services, mais séparément pour ceux du commerce de détail et de la construction. Antérieurement, on utilisait la répartition optimale de l'échantillon de Neyman, l'objectif étant de réduire au minimum la variance totale sous une taille d'échantillon fixe [voir, par exemple, Cochran (1977)]. Dans la MBS, l'organisme a décidé de répartir l'échantillon de manière que les coefficients de variation (c.v.) à certains niveaux spécifiés d'agrégation ne dépassent pas les cibles fixées. Ce problème peut s'exprimer sous forme d'un problème de répartition multivariée, où l'objectif est de minimiser la taille totale d'échantillon sous des contraintes de précision [voir, par exemple, Särndal et coll. (1992, p. 470)]. Le problème peut être formulé comme la minimisation d'une fonction convexe sous un certain nombre de contraintes linéaires; l'obtention de la solution optimale nécessite parfois des calculs très intensifs, mais il existe des heuristiques rapides donnant de bonnes solutions sous optimales. Nous avons utilisé la solution élaborée par l'Australian Bureau of Statistics (voir Preston, 2004) et l'avons mise en œuvre dans une macro SAS.

Spécifier les c.v. cibles n'a pas été une tâche facile, car les programmes d'enquête n'étaient pas capables d'indiquer quels niveaux seraient satisfaisants pour leurs utilisateurs. Donc, à titre de guide, nous avons estimé les c.v. au niveau des groupes de publication qui résulteraient d'une répartition de Neyman au niveau global et avons examiné leur distribution. Il s'est avéré que la majorité des c.v. au niveau le plus faible d'agrégation dans les groupes de publication étaient assez faibles (inférieurs à 5 %); donc, nous avons utilisé 5 % comme c.v. cible à ce niveau. Un

autre élément pris en compte était que le c.v. global devrait s'approcher du c.v. sous la répartition de Neyman. L'application de la macro multivariée avec ces cibles, et d'autres cibles à d'autres niveaux d'agrégation, a produit une taille d'échantillon plus grande que le maximum disponible; par conséquent, nous avons augmenté certaines cibles. Après quelques itérations, nous avons obtenu une répartition avec une taille totale d'échantillon très proche de la taille disponible et produisant des c.v. que les programmes d'enquête ont jugé satisfaisants. Le rééquilibrage de l'échantillon comparativement à celui obtenu en utilisant la répartition de Neyman a été faible dans la plupart des strates, mais important dans quelques unes.

2.2.4 Vérification et imputation

Les règles de vérification pour les variables similaires ont été harmonisées pour les diverses industries dans les secteurs des services et de la production, et toutes les règles propres à une industrie ont été examinées et modifiées pour les adapter à la nouvelle classification des industries. Une méthode de vérification sélective, en vue de repérer les enregistrements susceptibles de contenir des erreurs pouvant avoir une incidence importante sur les estimations si elles ne sont pas corrigées, était utilisée dans ces secteurs, mais les paramètres de la méthode n'avaient pas été mis à jour depuis un certain nombre d'années. Le remaniement de l'enquête nous a donné l'occasion d'examiner les paramètres de la vérification sélective en nous appuyant sur un critère commun pour toutes les industries, c'est à dire le biais maximal en pourcentage qui a été fixé à 1 % [voir Hooper et coll. (2011)]. La vérification sélective a également été adoptée pour le secteur du commerce de détail, ce qui a réduit de 20 points de pourcentage la proportion d'enregistrements désignée pour une reprise de contact en vue de corriger des erreurs possibles. L'introduction de la vérification sélective pour le secteur de la construction sera envisagée à une date ultérieure.

L'imputation est utilisée dans tous les secteurs pour traiter la non réponse totale et la non réponse partielle, mais les classes d'imputation ont été définies différemment dans chaque enquête. En raison des limites de notre système de traitement, il fallait que la définition des classes d'imputation soit la même pour tous les volets de l'enquête unique. Dans le cadre de l'intégration, les secteurs de la production et des services seront traités ensemble, si bien que nous avons dû choisir une définition unique des classes d'imputation. Les méthodes appliquées aux secteurs de la production et des services, et quelques autres variantes, ont été testées en se servant de données antérieures, ce qui a permis de constater que la méthode utilisée pour le secteur de la production offrait le meilleur compromis et elle a donc été adoptée.

2.2.5 Estimation

Dans tous les secteurs, l'estimation par le ratio était utilisée pour le chiffre d'affaires ainsi que pour l'emploi : le chiffre d'affaires provenant du registre des entreprises était utilisé comme variable auxiliaire pour les variables financières, tandis que le registre de l'emploi était utilisé pour obtenir les variables d'emploi (nombre total d'employés, employés à temps plein et à temps partiel et répartition selon le sexe). Pour d'autres variables financières, comme les exportations et les nouvelles commandes, les estimations étaient produites par la même méthode que pour le chiffre d'affaires. Pour le secteur de la distribution et des services, un calage était effectué dans chaque strate d'échantillonnage, ce qui signifie que des estimations par le ratio individuelles étaient utilisées, tandis que pour le secteur de la production, on utilisait l'estimation par le ratio combinée sur les strates échantillonnées dans chaque groupe d'industries employé pour la stratification. Pour le secteur du commerce de détail, l'estimateur par le ratio individuel était utilisé pour certaines industries et l'estimateur par le ratio combiné, pour d'autres.

Pour la MBS, l'organisme a décidé que la position par défaut consisterait à utiliser l'estimateur par le ratio individuel et que l'estimateur par le ratio combiné ne devrait être utilisé que si une strate contient un échantillon de petite taille et qu'il existe une strate « similaire » à laquelle on peut la combiner. Des études empiriques ont été effectuées pour décider dans quelle circonstance l'estimation par le ratio combinée pouvait être utilisée.

2.2.6 Traitement et publication des résultats

Comme nous l'avons mentionné plus haut, les secteurs de la production et des services sont maintenant entièrement intégrés dans la MBS, y compris leurs données, qui sont stockées dans une base de données unique et traitées ensemble. En outre, le traitement combiné des résultats pour les deux secteurs est devenu la responsabilité d'une même équipe, ce qui a nécessité un certain nombre de changements opérationnels. Ceux ci nous ont donné l'occasion

d'éliminer les incohérences en ce qui concerne les échéances et les méthodes de livraison des données sur le chiffre d'affaires aux Comptes nationaux. Une approche harmonisée a maintenant été adoptée pour produire des données pour expliquer les premières estimations du PIB (ce qui n'était fait auparavant que pour le secteur des services) et pour réviser les résultats dans toutes les industries.

Comme les données de sortie de l'enquête sont utilisées principalement comme données d'entrée dans l'indice de la production et l'indice des services, seul un ensemble limité de données de sortie des enquêtes était publié, mais elles figuraient dans deux publications distinctes, l'une couvrant le secteur de l'ingénierie, qui fait partie du secteur de la production, et l'autre, le secteur des services. Cependant, en janvier 2010, l'organisme a lancé une nouvelle publication commune dont la couverture de la production est plus vaste, intitulée Turnover, Orders in Production and Services Industries. En outre, une nouvelle stratégie de points de presse englobant la croissance ainsi que les révisions a été adoptée pour la MBS afin d'assurer un traitement uniforme des industries des secteurs de la production et des services.

En raison des contraintes de temps, l'intégration des enquêtes sur les secteurs du commerce de détail et de la construction à celles sur les secteurs de la production et des services sera envisagée à une date ultérieure. Nous nous attendons à ce que cette intégration pose plus de difficultés que celle des secteurs de la production et des services. En effet, les données sur le commerce de détail sont recueillies pour différentes périodes (cinq semaines, quatre semaines, quatre semaines) par opposition au mois civil et nécessitent des ajustements calendaires plus complets, tandis que pour le secteur de la construction, nous recueillons des données pour des variables très différentes de celles utilisées pour les autres secteurs. Nous devons examiner minutieusement les compromis entre les avantages du traitement combiné des données et les coûts éventuels sous forme de perte de flexibilité et d'accroissement de la complexité des systèmes.

3. Mise en œuvre du remaniement

Une version d'essai, hors ligne, a été configurée en se fondant sur le nouveau plan d'échantillonnage, le nouveau type de questionnaire, ainsi que les règles de vérification et les méthodes d'imputation et d'estimation révisées. Nous avons donc pu procéder à une mise à l'essai complète du système avant de le mettre en ligne en janvier 2010. En outre, un essai de volume de la base de données combinée a été couronné de succès.

Le changement de classification et de plan d'échantillonnage signifiait que l'on introduirait dans l'échantillon un beaucoup plus grand nombre d'entreprises que cela n'aurait été le cas en appliquant les taux existants de renouvellement de l'échantillon – l'organisme utilise un échantillonnage rotatif en utilisant des numéros aléatoires permanents (NAP) pour contrôler le chevauchement des échantillons. Pour s'assurer que le chevauchement de l'échantillon soit adéquat, on s'est efforcé de spécifier un NAP de départ approprié dans chaque strate. Malgré ces efforts, le nombre d'entreprises n'ayant pas rempli de questionnaire auparavant était plus élevé que d'habitude, ce qui a rendu la validation des données difficile – la règle de vérification qui consiste à comparer la déclaration courante à la déclaration précédente n'a pas pu être appliquée dans un nombre de cas beaucoup plus élevé qu'à l'ordinaire.

Les entreprises ayant changé de secteur ont également posé des difficultés, car elles avaient reçu un type de questionnaire différent auparavant. Par exemple, une entreprise transférée du secteur du commerce de détail, pour lequel elle déclarait ses « ventes au détail », au secteur des services devra déclarer son « chiffre d'affaires total », de sorte qu'il faudra plus de temps pour valider les données.

La MBS a adopté la nouvelle classification à compter de janvier 2010, mais les Comptes nationaux avaient besoin de données conformément à l'ancienne classification pour une période de presque deux ans. Il a donc fallu écrire un nouveau programme de traitement et le mettre à l'essai, ce qui a accentué la pression ressentie par tous les participants au projet. En outre, les programmes d'enquête ont trouvé que l'exécution en parallèle des deux systèmes était difficile et gourmande en ressources; bien qu'une formation ait été offerte, un soutien important a néanmoins été nécessaire au cours des quelques premiers mois.

La MBS a été lancée en janvier 2010 et des lettres d'avertissement ont été envoyées tôt aux répondants. De cette façon, la transition s'est faite sans heurts et sans qu'aucun problème important ne soit signalé. Le principal problème

s'est posé quand les répondants ont communiqué avec l'organisme pour obtenir des éclaircissements : une certaine confusion a eu lieu quand les répondants faisaient référence à la MBS alors que nos employés continuaient de penser aux anciens noms d'enquête. Bien que les modifications apportées aux enquêtes aient été communiquées aux employés, la formation offerte à certains programmes n'a pas été suffisante. Une plus grande importance a été accordée à la production des données de sortie et à la communication avec les utilisateurs externes au sujet des changements dans les séries de données dus à la modification de la classification des industries.

Le projet visant à créer la MBS s'est déroulé parallèlement aux projets de mise en œuvre du changement de classification des industries et de remaniement des séries de données sur l'emploi. Chaque projet a été géré par une équipe distincte, mais la Méthodologie était représentée par la même personne dans les trois projets. Cela, et le fait que l'équipe de projet de la MBS comprenait des représentants de toutes les parties intéressées (collecte des données, gestionnaires d'enquête pour toutes les données de sortie pertinentes, méthodologie et gestion de l'information) ont rendu la communication entre les divers volets plus efficace. Les participants ont cherché à obtenir des commentaires au sujet de l'efficacité du groupe et des modifications opérationnelles ont été apportées.

Des plans ont été établis très tôt et les parties intéressées ont été consultées; cette mesure s'est avérée utile, quoiqu'il ait été difficile d'obtenir la pleine participation de ceux qui n'étaient pas directement concernés à l'époque. Par conséquent, des demandes de changement au sujet de détails de l'enquête ont été soumises à la dernière minute, ce qui a accru la pression exercée sur les membres de l'équipe de projet qui devaient déjà respecter un échéancier très exigeant.

4. Conclusion et travaux à venir

Le remaniement de tous les processus et de toutes les méthodes de production des statistiques conjoncturelles devrait mener à des estimations plus pertinentes et de meilleure qualité. En outre, la normalisation des processus et des méthodes, et l'intégration complète de deux secteurs – production et services – devraient permettre de réaliser des économies. La vérification méticuleuse des spécifications et l'assurance rigoureuse de la qualité des systèmes, le soutien offert à l'équipe chargée des résultats, ainsi qu'une bonne gestion et une bonne coordination du projet ont contribué au succès de la mise en œuvre des changements.

Les principales leçons tirées de l'expérience sont que des compromis sont souvent nécessaires, que la participation des parties intéressées est importante, mais difficile à obtenir et que du temps pour les changements de dernière minute devrait être intégré dans la planification. En outre, l'évaluation prévue des changements après un an s'est avérée utile : nous apportons certaines modifications aux plans d'échantillonnage maintenant que nous disposons de meilleurs renseignements.

Les travaux concordaient bien avec d'autres projets de l'organisme et ont ouvert la voie à de futures améliorations. L'un des domaines visés est l'harmonisation de la poursuite des réponses dans les divers volets de la MBS : à l'heure actuelle, nous appliquons de simples règles d'établissement des priorités, mais elles ne sont pas les mêmes pour tous les secteurs et ne sont pas nécessairement optimales. L'organisme a étudié ce problème au cours des dernières années : des travaux effectués à la Southampton University (voir Berger, 2009) portaient sur plusieurs méthodes d'attribution d'un score qui ont été testées dans une simulation fondée sur des données provenant du secteur de la distribution et des services. Les travaux ont montré qu'il existe une interaction entre les classes d'imputation, la méthode d'imputation et la méthode d'attribution du score. Par conséquent, l'organisme a décidé d'attendre jusqu'à ce que des données obtenues en utilisant les nouveaux plans d'échantillonnage soient disponibles pour l'analyse avant de décider quelle stratégie de poursuite de la réponse il convient d'adopter. L'analyse des profils de réponse est en cours. Un autre domaine d'étude important est celui de l'utilisation de données administratives pour produire des statistiques conjoncturelles. Ces travaux sont exécutés dans le cadre d'un projet européen de modernisation des statistiques européennes sur les entreprises et sur le commerce (ESSnet sur l'amélioration de l'utilisation des données administratives et comptables pour la production de statistiques sur les entreprises). Le principal défi posé par ces travaux consiste à résoudre la question du mélange de périodicités des données administratives et celle de leur actualité (voir Orchard et coll. 2011). Un troisième domaine est celui de l'utilisation de la collecte électronique des données, qui couvre deux aspects : le premier est l'utilisation de questionnaires en ligne et le deuxième, l'obtention de flux de données (data feed) provenant des entreprises. L'organisme a entrepris d'examiner en premier lieu la faisabilité de l'établissement de flux de données pour les données sur la paye.

Le regroupement des divers volets des travaux, ainsi que l'élaboration de plans cohérents pour apporter les changements requis à nos systèmes et processus en vue de mettre en œuvre les méthodes qui découleront des travaux courants et à venir qui continuent de poser des défis.

Bibliographie

Berger, I.G. (2009), « Priority Response Chasing », rapport technique non publié, University of Southampton, Royaume-Uni.

Cochran, W. G. (1977), *Sampling Techniques*: 3e édition.Wiley series in Probability and Mathematical Statistics, New York.

Hooper, E., Lewis, D. et C. Dobbins (2011), « The application of Selective Editing in the ONS Monthly Business Survey », *Survey Methodology Bulletin*, Office for National Statistics, Royaume-Uni, no 68, p. 1 à 11.

Gareth, J. (2010), « Improving the Design of UK Business Surveys », Article présenté lors de la European Conference on Quality in Official Statistics, Helsinki, Finlande, disponible à http://q2010.stat.fi/media//presentations/session-14/james_paper.pdf.

Orchard, C., Moore, K. et A. Langford (2011), « Practices for Using VAT Turnover Data Within the UK to Produce Estimates of Growth and Monthly Turnover », rapport technique, produit 4.2 pour Work Package 4, ESSnet sur l'amélioration de l'utilisation des données administratives et comptables pour la production de statistiques sur les entreprises, disponible à http://essnet.admindata.eu/Document/GetFile?objectId=5278.

Preston, J. (2004), « Optimal Sample Allocation in Multivariate Surveys: An Integer Solution », document non publié, Australian Bureau of Statistics.

Särndal, C.-E., Swensson, B. et J. Wretman (1992), *Model-Assisted Survey Sampling*, New York:Springer-Verlag.

Taylor, C., James, G. et P. Pring (2011), « The Development of the Monthly Business Survey », *Economic and Labour Market Review*, Office for National Statistics, Royaume-Uni, vol. 5, no 2, p. 95 à 103.

SÉANCE 3B

DIFFUSION ET ACCÈS AUX DONNÉES

SICON d'Eurostat : Infrastructure de sécurité pour l'accès aux données confidentielles et le partage de ces données

Dario Buono[1]

Résumé

Le but global de ce projet est d'élaborer et de mettre en œuvre un projet pilote d'infrastructure, de services et de documentation pour l'accès par des partenaires externes aux ensembles de données confidentielles de l'Union européenne (UE) détenus par Eurostat, principalement des Instituts nationaux de statistique (INS), en vue d'intégrer les processus des États membres (EM) et d'Eurostat. L'infrastructure est destinée à être utilisée par l'unité de production d'Eurostat (les principaux utilisateurs) intéressée à développer des processus de production de statistiques plus intégrées pour l'UE avec des partenaires clés. À la première étape, seules une consultation des données et des mesures à distance sont envisagées; la fonction de transfert de fichier ne sera pas mise en œuvre. Les solutions proposées devraient assurer la sécurité des données, tirer le meilleur parti possible de la solution déjà élaborée au niveau de la Commission (environnement DIGIT-RACHEL pour CITRIX et WEBGATE/INTRAGATE) et être compatibles avec l'infrastructure des TI des INS. Les lignes directrices et les protocoles pour l'exécution et la gestion du système doivent être établis et inclus dans le manuel de traitement des données confidentielles à Eurostat. Des essais pilotes sont prévus pour tenir compte des besoins des deux projets connexes, Euro Group Register et Decentralised Access for Scientific Purposes. À la fin du projet, des recommandations seront faites en vue d'adapter l'infrastructure aux besoins réels et serviront de modèle opérationnel pour l'exécution et le maintien de cette infrastructure.

[1]Dario Buono, Eurostat, Luxembourg.

Aperçu et description technique du système amélioré de production de tableaux : Normalisation de la production de tableaux personnalisés pour accroître la qualité des données et l'efficacité

Peter Timusk, Mamdouh Mansour, Éric Pelletier et Eric Turgeon[1]

Résumé

La présente communication fournit un aperçu et une description technique d'un nouveau système de totalisation amélioré, qui intègre des outils conformes à l'Architecture opérationnelle du Bureau et permet une meilleure assurance de la qualité, ainsi que des gains d'efficacité, grâce à la normalisation de la production de tableaux personnalisés. Elle aborde les objectifs du système amélioré, souligne les économies de programmation qui ont été réalisées et examine le rôle des méthodologistes pour l'adoption de stratégies d'uniformisation.

Mots clés : Totalisation; tableaux personnalisés; stratégies de normalisation; économies de programmation; assurance de la qualité.

1. Introduction

1.1 Introduction

Un système amélioré de production de tableaux a été mis en œuvre et tire profit des avantages des composantes de systèmes généralisés conformes à l'Architecture opérationnelle du Bureau (AOB), dans un environnement de logiciel d'analyse statistique (SAS), en plus de faire intervenir un système de production « Output Delivery System » (ODS) en SAS dans une nouvelle application de pointe (qui est aussi décrite). Ce système a été élaboré et a fait ses preuves au cours de la dernière année à la Division des enquêtes-entreprises spéciales et de la statistique de la technologie de Statistique Canada, où les spécialistes de la production de tableaux sont chargés du maintien de la capacité des systèmes de produire des totalisations fiables, pertinentes et descriptives pour une gamme variée d'enquêtes personnalisées à frais recouvrables. Les spécialistes sont chargés de la production de tableaux personnalisés qui se rapportent à des sujets, mesures, concepts et normes en constante évolution. C'est pourquoi un système très souple et adaptable était nécessaire pour produire les structures des tableaux, les spécifications d'entrée ou les exigences de sortie, qui changent avec chaque projet. La seule constante est l'utilisation du Système généralisé d'estimation (SGE) et de Confid2/G-Confid, ainsi que des processus opérationnels normalisés dont se sert la section pour appuyer les gestionnaires d'enquête.

1.2 Aperçu du processus de production

Pour produire des tableaux personnalisés de statistiques, nous utilisons deux outils conformes à l'Architecture opérationnelle du Bureau (AOB) de Statistique Canada. L'un d'eux, le Système généralisé d'estimation (SGE), permet le calcul d'estimations de ratios, de moyennes, de totaux et de chiffres pour les résultats d'enquête. Ce programme permet en outre de calculer un coefficient de variation ou, au besoin, une erreur-type, en tenant compte de l'échantillonnage et du plan de sondage. L'autre outil de l'AOB que nous utilisons, G-CONFID (anciennement CONFID2), est un système généralisé de Statistique Canada qui offre une méthodologie conçue pour prévenir la diffusion de données confidentielles. Il permet le calcul d'un schéma de suppression de X pour un tableau, grâce auquel l'estimation sera remplacée par la lettre « X ». On procède ainsi afin que le tableau produit ne puisse pas servir pour identifier les réponses d'une entreprise individuelle ou d'une personne et que seuls des résultats agrégés soient fournis.

[1]Peter Timusk, Mamdouh Mansour, Éric Pelletier et Eric Turgeon, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6.

En résumé, nous utilisons notre système de production de tableaux pour :

- produire des tableaux personnalisés d'estimations d'enquête
- produire des indicateurs de fiabilité (coefficients de variation ou erreurs-types)
- assurer la confidentialité (suppression et non-publication de certaines estimations).

2. Problèmes posés par les anciens systèmes de production de tableaux

Notre ancien système de production de tableaux comportait un certain nombre de problèmes techniques. Nous utilisions avec le SGE une lourde interface graphique , qui exigeait un contrôle fastidieux d'un grand nombre de variables. Le système comportait en outre des limites de taille, du fait que les programmes traitaient un nombre important d'estimations en même temps. Dans les totalisations, les estimations figuraient sous forme de variables alphabétiques dans des tableaux Excel, plutôt que sous forme de chiffres, comme il aurait fallu. Si une estimation comporte un coefficient de variation ou une erreur-type trop élevé, on lui attribue un indicateur de qualité « F » et elle n'est pas publiée. La lettre « F » remplace l'estimation dans le tableau publié. Dans le cas des estimations remplacées par des X parce qu'elles étaient supprimées ou par des F, pour des raisons de qualité, une lettre apparaît à la place du chiffre. C'est pourquoi les chiffres ont été publiés sous forme de variables caractères.

Une fois les tableaux créés, un spécialiste en la matière a été chargé de vérifier leur qualité. Il s'agissait d'un processus manuel de contrôle de la qualité (CQ) qui est susceptible à entraîné certaines erreurs.

La section de production de tableaux traitait des tableaux personnalisés découlant d'un plan de sondage ponctuel pour des enquêtes ponctuelles, et aucun système de traitement normalisé simple n'avait été conçu pour donner suite à ces demandes. Plus souvent qu'autrement, les tableaux prenaient la forme de totalisations croisées complexes, qui nécessitaient une programmation complexe pour que des estimations puissent être créées et que ces totalisations complexes puissent être produites.

Par le passé, on dépendait d'un programmeur principal pour exécuter ces travaux de programmation complexes. Cette structure, et la dépendance qui en découlait, rendaient la section vulnérable.

Par ailleurs, la documentation relative à ces tableaux complexes se limitait à quelques lignes de commentaires dans les programmes ou n'était pas accessible à un endroit ou dans un format lisible normalisé. Cela signifiait que lorsqu'un tableau avait besoin de données d'une enquête antérieure, beaucoup d'efforts devaient être déployés pour déterminer comment l'ancienne enquête avait été totalisée et quelles en étaient les spécifications exactes.

3. Remaniement du système de production de tableaux

Il a été décidé qu'on devait améliorer le système de production de tableaux. Tout d'abord, un bon modèle du système de production en place était nécessaire. Des analystes de systèmes opérationnels ont été recrutés pour documenter les systèmes, les processus et les procédures de production par la voie formelle de la modélisation des processus opérationnels (MPO). Pour apporter ces changements au système, l'équipe responsable a collaboré étroitement avec le personnel de production pour documenter ce qui était fait et de quelle façon, dans un cadre formel de la MPO.

Un développeur de systèmes a été recruté pour élaborer un ensemble de macros en SAS, afin de permettre une approche simplifiée à l'égard du SGE et de CONFID2.

4. Nouveau système amélioré de production de tableaux (SAPT) – Sommaire des changements, aperçu du système et exemple

4.1 Sommaire des changements

Dans le nouveau système amélioré de production de tableaux (SAPT), l'approche a consisté à utiliser une programmation simplifiée (un programme par tableau). Pour simplifier la conception des tableaux, on a demandé aux clients de les concevoir avec un moins grand nombre de variables. La conception a été confiée aux clients, plutôt qu'au personnel de production. Le transfert de la conception des tableaux à l'analyste d'enquête a permis la production d'ententes écrites formelles sur les spécifications de tableaux. Cela a mené à des spécifications de confidentialité et des spécifications de fiabilité par écrit.

Dans le contexte de la conception des tableaux et de la simplification de la programmation, on utilise maintenant une approche plus modulaire. Le système est plus souple pour donner suite aux multiples demandes personnalisées de tableaux. Le nouveau système utilise un processus et une procédure pouvant être répétés.

Par ailleurs, un certain nombre de normalisations ont été possibles pour la mise en œuvre du nouveau système, y compris des règles d'attribution des noms. Celles-ci ont été appliquées avec succès à la désignation des fichiers et des variables du questionnaire.

Un système de totalisation plus à jour peut maintenant être utilisé avec ODS en SAS, qui sert aussi d'autres façons pour calculer des ensembles de données pour la suppression selon les dénombrements lorsque CONFID2 n'est pas utilisé. L'utilisation de formats ODS en SAS et SAS a permis le recours à des « valeurs manquantes spéciales » en SAS pour coder les lettres désignant le niveau de qualité et les X comme variables numériques. Cela fait en sorte que les estimations qui prennent la forme de chiffres peuvent demeurer en format numérique lorsqu'elles sont totalisées.

Les programmes en SAS sont maintenant documentés de façon régulière au moyen de « commentaires », qui permettent aux autres programmeurs de comprendre clairement le code et de le localiser facilement selon les sections du programme. Cela a permis l'intervention de plusieurs programmeurs ayant des niveaux différents d'ensembles de compétences. En général, un programmeur subalterne et un programmeur principal ont collaboré.

Ainsi, un plus grand nombre de programmeurs ont pu rédiger et vérifier les programmes, ce qui a permis de rehausser de façon générale les compétences en programmation SAS de la section. Les programmeurs de la section ont suivi beaucoup de cours et une formation exhaustive. Par ailleurs, la section a connu un afflux de programmeurs en SAS expérimentés. La dépendance à l'égard d'un nombre limité de programmeurs ne représente plus un risque.

Un outil a été adopté pour permettre de vérifier la qualité des tableaux finaux de production en Excel par rapport aux mêmes estimations produites de façon indépendantes. Grâce à cet outil, le contrôle de la qualité est maintenant pleinement automatisé, et une comparaison à 100 % peut être effectuée pour toutes les estimations, pour des milliers de tableaux et de cellules.

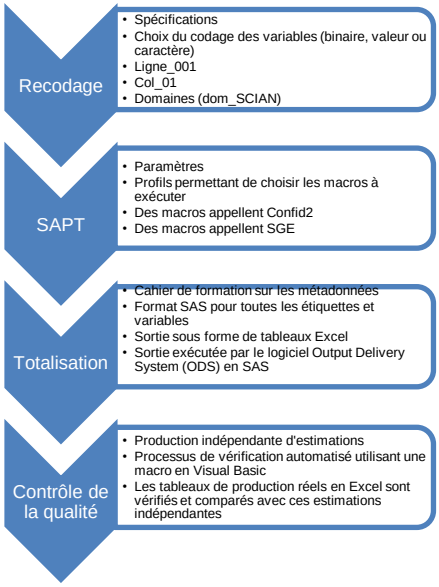
4.2 Aperçu du SAPT

Les programmeurs sont maintenant responsables du recodage des données d'enquête dans des ensembles de données devant être entrées dans les processus opérationnels de l'AOB, et cela est toujours du ressort de l'équipe de production de tableaux. Ces tâches peuvent être accomplies par des programmeurs de l'équipe à divers niveaux.

- Les spécifications sont détaillées et fournies par le client sous forme de « tableaux de spécifications ».
- Les paramètres sont utilisés dans le système de base, ainsi que dans le système de totalisation du SAPT.
- On a pu réaliser des gains réels d'efficacité grâce à de la documentation appropriée.

Le diagramme qui suit montre les étapes qui sont actuellement suivies :

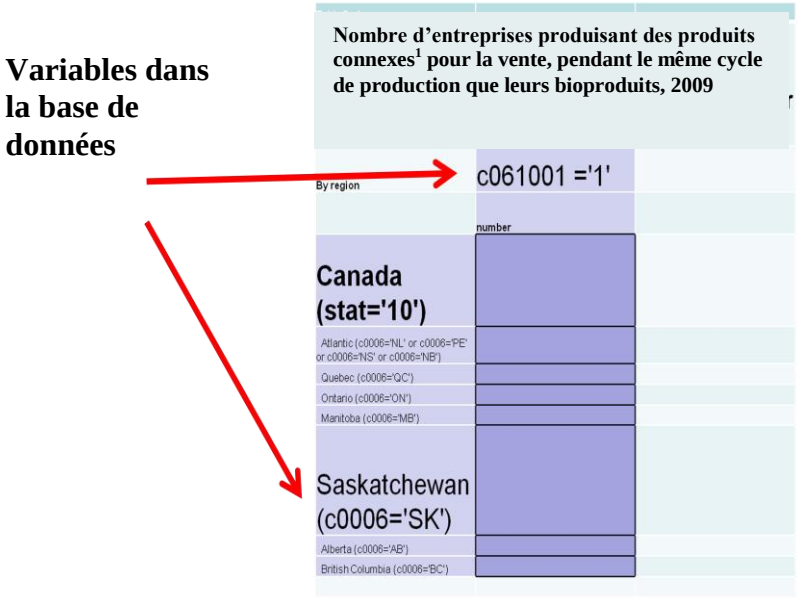
Figure 4.2-1
Aperçu du processus du SAPT



4.3 Exemple de SAPT

Pour le « tableau de spécifications », le client précise la conception du tableau et détermine les variables dans la base de données (souvent un fichier de résultats d'enquête). Le client précise alors les valeurs des variables pour toutes les lignes et colonnes, ainsi que les populations et les sous-populations. Les trois diagrammes qui suivent donnent un aperçu de la façon dont le client s'acquitte de cette tâche. Même si les tableaux semblent être les mêmes dans chaque diagramme, chacun a trait à une spécification différente du point de vue de la programmation, qui sera effectuée selon cette spécification.

Figure 4.3-1
Exemple de spécification – Variables



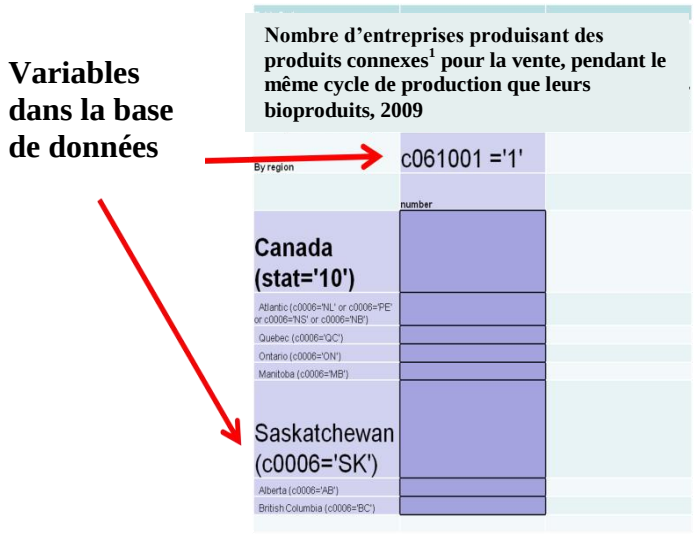


Figure 4.3-2
Exemple de spécification – Lignes et colonnes

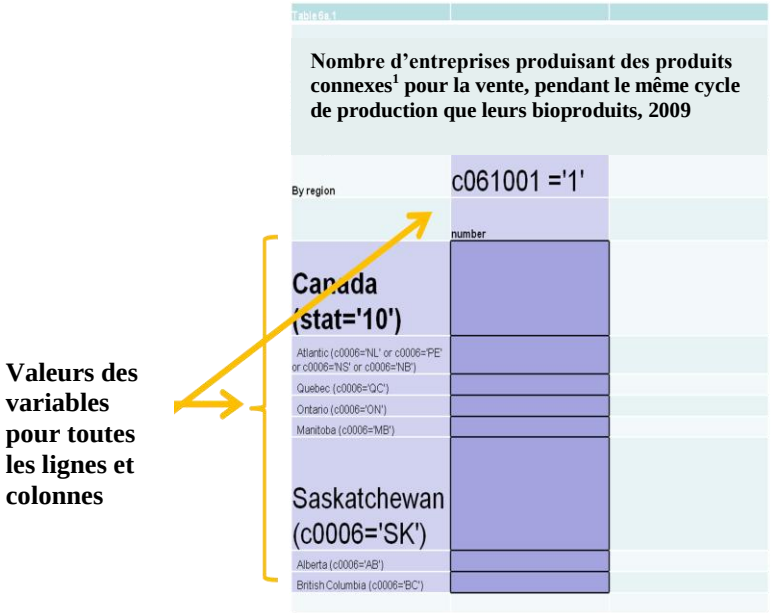
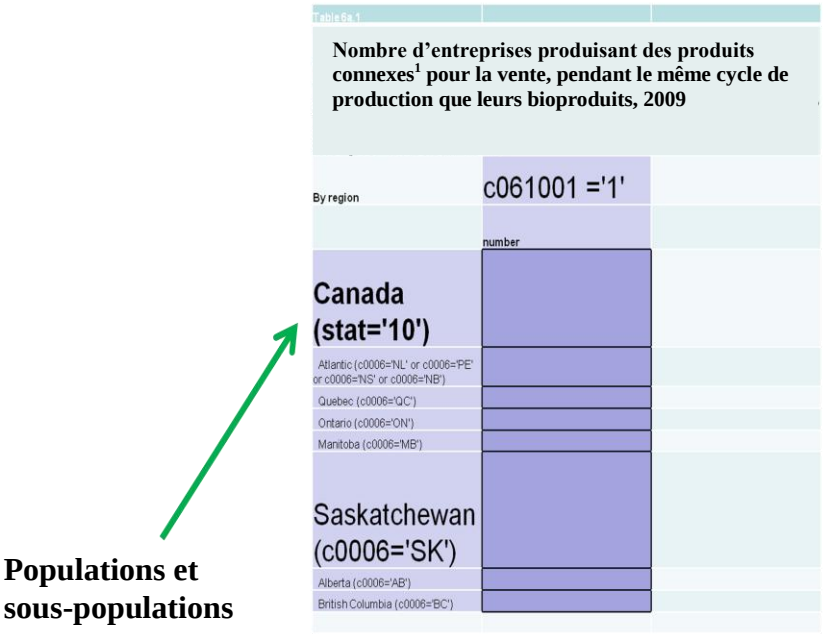


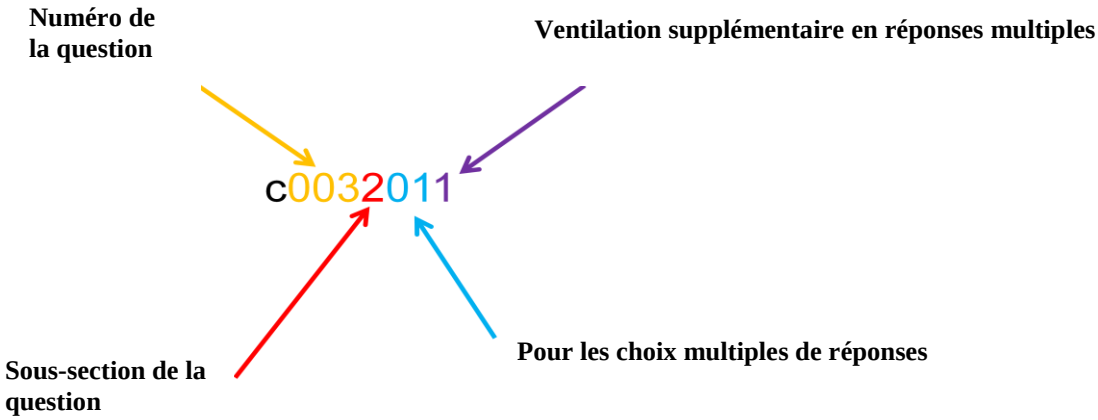
Figure 4.3-3
Exemple de spécification – Populations



5. Normalisation

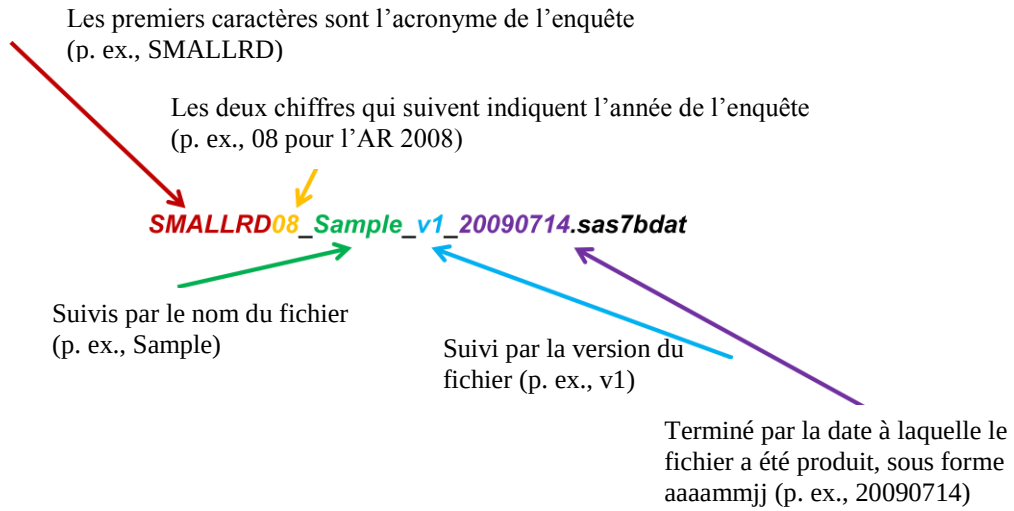
Les deux diagrammes qui suivent montrent des exemples de normalisation des règles d'attribution des noms qui ont été adoptées dans le SAPT. Le premier montre la normalisation de la numérotation des questions, qui sont utilisées comme noms de variable dans le fichier de données des résultats de l'enquête qui comprend les réponses à ces questions.

Figure 5.1-1
Normalisation de la numérotation des questions



Le deuxième diagramme montre la règle d'attribution des noms des fichiers de données utilisée dans le système amélioré de production de tableaux. Celle-ci a aussi été normalisée par les méthodologistes et adoptée par la section de production de tableaux. Lorsqu'un ensemble de données change souvent et que le fichier est mentionné ailleurs à de nombreux endroits, la date ne sera pas utilisée dans le nom, la modification du nom du fichier nécessitant de trop nombreux contrôles en amont et la date de production d'un fichier changeant souvent dans un projet de production.

Figure 5.1-2
Normalisation du nom de fichier



6. Qualité des données

Des efforts ont été déployés pour automatiser et documenter le contrôle de la qualité. Dans ce contexte, le contrôle de la qualité signifie un calcul indépendant et une comparaison des estimations. Statistique Canada a trouvé une macro en Visual Basic qui permet de comparer automatiquement cellule par cellule deux tableaux de sortie Excel. La macro produit un fichier journal de sortie des différences dans les estimations et les valeurs de fiabilité. Un développeur de système a aidé à transformer la macro en une application robuste répondant aux besoins du Système amélioré de production de tableaux. Le contrôle de la qualité est maintenant complètement automatisé et une comparaison à 100 % peut être effectuée pour toutes les estimations pour des milliers de tableaux et de cellules.

7. Efficacité, efficience et avantages concernant la répartition du travail

Le travail de programmation est maintenant affecté à différents programmeurs selon la complexité du tableau.

Nous procédons maintenant à une vérification entièrement automatisée du produit programmé, grâce à des totalisations indépendantes : une par le programmeur dans la production des tableaux, et une deuxième totalisation indépendante par l'agent spécialisé. Un programme par tableau réduit la complexité de la programmation et la charge imposée aux systèmes de traitement, comme le SGE.

Par ailleurs, l'utilisation de règles normalisées d'attribution des noms a augmenté l'efficacité.

8. Défis qui subsistent

Des processus opérationnels additionnels n'ont pas encore été intégrés dans le SAPT. Par exemple, « SEVANI », un système généralisé pour calculer les variances dues à l'imputation, n'est pas encore intégré dans les macros et les profils disponibles dans le SAPT. Les caractéristiques de protection de la confidentialité des données les plus récentes de G-CONFID ne sont pas encore intégrées dans le SAPT. Le système a été élaboré pour la version CONFID2 de CONFID.

Le processus de contrôle de la qualité du SAPT est maintenant amélioré, grâce à la vérification des estimations et des variations des coefficients de variance, en cas d'exécutions indépendantes du SGE. Nous pouvons aller plus loin, grâce à l'utilisation plus fréquente de processus automatisés pour vérifier les profils de suppression selon les dénombrements. Nous commençons à utiliser la comparaison indépendante des nombres de contributeurs aux estimations produits par SAS Proc Freq à ceux produits par SAS Proc Tabulate, pour vérifier les nombres de contributeurs pour des raisons de suppressions.

L'intégration du SAPT aux outils conformes à l'AOB permet d'assurer le soutien intégré des systèmes utilisés pour produire les tableaux. À long terme, des projets de base et permanents à frais recouvrables profiteront de cette initiative.

9. Conclusions

Notre nouveau SAPT simplifie la production parce qu'il permet à plusieurs programmeurs de collaborer efficacement. Il permet d'améliorer le contrôle de la qualité, grâce à l'automatisation du processus, et nous permet d'assurer un CQ à 100 %, en dépit du volume de produits programmés.

Nous avons maintenant restructuré l'utilisation du SGE et de CONFID2, afin que nos travaux puissent être axés sur la création d'ensembles de données d'entrée. Nous avons aussi utilisé le SAPT pour l'adoption de règles normalisées d'attribution des noms, qui permettent l'identification efficace et sans erreur des fichiers, ainsi que la détermination et la résolution des problèmes.

10. Remerciements

Les auteurs souhaitent remercier tous ceux qui ont participé à l'équipe de production améliorée de tableaux, pour les efforts consacrés au projet. Par ailleurs, Frances Anderson a fourni aux auteurs certaines directives clés pour la structuration de la présentation et la description du projet de production améliorée de tableaux. Par ailleurs, Paula Thompson et Greg Peterson ont passé en revue les ébauches finales des diapositives de présentation, et George Sciadas, la communication destinée au recueil. Les auteurs souhaitent remercier ces réviseurs pour leur aide.

SÉANCE 4A

VÉRIFICATION SÉLECTIVE

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

La vérification sélective des données et sa mise en œuvre à Statistics Sweden

Pär Brundell[1]

Résumé

La vérification sélective des données permet de réduire le processus de vérification des données des enquêtes-entreprises souvent gourmand en ressources. Par conséquent, Statistics Sweden a développé un logiciel générique de vérification sélective (de microdonnées) appelé Selekt. Le logiciel existe aujourd'hui dans sa version 1.2 et il est prévu, maintenant et dans un avenir proche, de le mettre en œuvre dans une demi-douzaine d'enquêtes-entreprises afin d'acquérir une expérience et des connaissances utiles sur lesquelles s'appuyer par la suite. Le présent article donne un bref aperçu de Selekt et de certains de ses fondements théoriques, ainsi que des commentaires sur les travaux de mise en œuvre dans certaines enquêtes-entreprises.

Mots clés : Vérification sélective des données ; macro Selekt en SAS.

1. Contexte

1.1 Erreurs

Les erreurs dans les données d'enquête peuvent être présentes dans les données brutes fournies par les répondants à l'organisme statistique ou se produire durant la transmission des données. Le processus de production de statistiques est une combinaison de nombreuses activités présentant des risques d'introduction d'erreurs. Les types d'erreurs comprennent, par exemple, la non-réponse partielle, les valeurs non valides, les erreurs de modélisation ou les contradictions. Les valeurs suspectes des données peuvent être réparties en deux catégories, à savoir les erreurs d'écart (valeurs aberrantes) et les erreurs de définition (valeurs non aberrantes) suspectes lorsque de nombreux répondants comprennent mal une question ou qu'ils vont chercher les données dans des systèmes d'information utilisant d'autres définitions que celles souhaitées. Les erreurs d'écart suspectes requièrent souvent un suivi manuel qui demande du temps et coûte cher. Dans de nombreux cas, les contrôles appliqués pour déceler ce genre d'erreurs ont un faible taux de succès et bon nombre des changements apportés aux données ont très peu d'effet sur les statistiques finales. Les erreurs de définition pourraient être difficiles à déceler et certains moyens de les découvrir pourraient consister à combiner la vérification pour plusieurs enquêtes, à procéder à des interviews en profondeur auprès de groupes de discussion, à rechercher les proportions élevées de non-réponses partielles ou à utiliser des contrôles graphiques.

1.2 Vérification

La vérification est une activité qui englobe la détection, la compréhension et la résolution des erreurs dans les données et dans les statistiques produites. Les activités de vérification comprennent généralement au moins certaines des opérations suivantes :

A. Vérification auprès du répondant
B. Vérification manuelle avant l'enregistrement des données
C. Vérification lors de l'enregistrement des données
D. Vérification en production/microvérification
   1 Vérification « classique »
   2 Vérification sélective
E. Analyse de cohérence
F. Vérification des données de sortie/macrovérification

[1]Pär Brundell, Statistics Sweden, Klostergatan 23, Suède, SE-70189.

G. Évaluation
H. Contrôle de la livraison

Les activités sont nombreuses, mais beaucoup d'efforts sont souvent consacrés à la vérification en production (D. ci-dessus) et c'est à ce stade que la vérification sélective peut, l'espère-t-on, réduire la charge de travail de l'organisme statistique. Contrairement à la vérification sélective, la vérification classique consiste souvent à établir des limites de tolérance pour les contrôles. Ces limites peuvent être déterminées en se fondant sur des cycles antérieurs des enquêtes. La vérification sélective peut s'appuyer sur ce genre de contrôle classique et sur une fonction de score afin de sélectionner uniquement les erreurs suspectes qui, en principe, auront l'effet le plus important. Dans le but d'appliquer à grande échelle la vérification sélective des données, Statistics Sweden a développé un outil logiciel appelé Selekt. Il est constitué de macros SAS et d'un « tableau de bord » qui permettent à l'utilisateur de fixer la valeur des paramètres nécessaires et facultatifs pour concevoir la vérification sélective des données pour un ensemble de données défini par l'utilisateur.

2. Vérification sélective des données

2.1 Concepts de base

Le but de la vérification sélective des données est de réduire le coût pour l'organisme statistique ainsi que pour les répondants, sans diminuer de manière significative la qualité des statistiques produites. Une approche de vérification sélective pourrait consister à cibler certains enregistrements de microdonnées (observations). Chaque enregistrement pourrait d'abord être classé comme étant susceptible ou non de contenir des erreurs. Ensuite, l'effet potentiel de chaque enregistrement suspect sur toutes les statistiques serait calculé et seuls les enregistrements ayant l'effet potentiel le plus important sur les statistiques finales seraient signalés pour le suivi manuel.

Toutefois, l'approche de vérification sélective adoptée à Statistics Sweden consiste à pousser encore plus loin l'approche décrite à la figure 1. Au lieu d'être dichotomique, la valeur du degré de soupçon peut être continue, allant de zéro à un. L'effet potentiel sur les statistiques produites est calculé au départ pour chaque enregistrement (observations) et pour chaque variable. L'effet prévu de l'enregistrement et de la variable est le produit du degré de soupçon et de l'effet potentiel. Les observations signalées deviennent alors celles possédant les valeurs agrégées les plus élevées des effets prévus.

L'approche de vérification sélective des données consiste à construire une fonction de score pour classer les variables et les enregistrements par ordre de priorité. À la première étape, le score est calculé en multipliant le degré de soupçon par l'effet. Le degré de soupçon varie de manière continue de 0 à 1 et la valeur de l'effet est toujours absolue, puisque la direction de l'effet ne présente aucun intérêt dans cette approche.

Figure 1. Une première approche de vérification sélective des données pourrait consister à signaler les observations suspectes ayant l'effet potentiel le plus important (cadran supérieur droit de la figure)

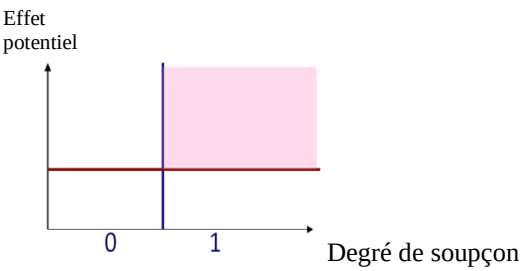
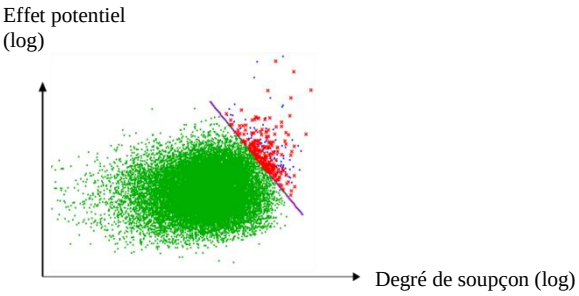


Figure 2. L'approche de vérification sélective des données de Statistics Sweden s'appuie sur les valeurs continues de l'effet et du degré de soupçon



2.2 Degré de soupçon

Le degré de soupçon peut être fixé individuellement pour chaque enregistrement (c'est-à-dire unité observée) et pour chaque variable. Le degré de soupçon varie de manière continue de zéro (aucun soupçon) à un (entièrement suspect). Le calcul du degré de soupçon repose sur la comparaison entre la valeur non vérifiée qui, par exemple, est transmise par le répondant et sa valeur et sa variation prédites. Dans Selekt, l'utilisateur a le choix entre deux grandes options pour ce qui est de la méthode d'obtention de la valeur prédite. Cette dernière peut être obtenue a) en analysant la série de données chronologiques sur les variables étudiées pour chaque unité observée ou b) en calculant la moyenne ou la médiane pour chaque variable étudiée dans les groupes de vérification définis par l'utilisateur de Selekt. Ce dernier comprend une fonction intégrée qui, d'après les groupes de vérification définis par l'utilisateur, donne la moyenne ou la médiane au niveau le plus faible possible en ce qui concerne le nombre d'observations. Le calcul repose sur une méthode d'analyse arborescente.

Dans Selekt, le degré de soupçon est défini comme étant Soupçon = R/(Tau+R) où R est défini comme étant

$$R = \begin{cases} \left(\tilde{z}_{j,k,l} - KAPPA\cdot\left(\tilde{z}_{j,k,l} - \tilde{z}^{L}_{j,k,l}\right) - z_{j,k,l}\right) / \left(\tilde{z}^{U}_{j,k,l} - \tilde{z}^{L}_{j,k,l}\right) & si\ z_{j,k,l} < \tilde{z}_{j,k,l} - KAPPA\cdot\left(\tilde{z}_{j,k,l} - \tilde{z}^{L}_{j,k,l}\right) \\ 0 & si\ \tilde{z}_{j,k,l} - KAPPA\cdot\left(\tilde{z}_{j,k,l} - \tilde{z}^{L}_{j,k,l}\right) < z_{j,k,l} < \tilde{z}_{j,k,l} + KAPPA\cdot\left(\tilde{z}^{U}_{j,k,l} - \tilde{z}_{j,k,l}\right) \\ \left(z_{j,k,l} - \tilde{z}_{j,k,l} - KAPPA\cdot\left(\tilde{z}^{U}_{j,k,l} - \tilde{z}_{j,k,l}\right)\right) / \left(\tilde{z}^{U}_{j,k,l} - \tilde{z}^{L}_{j,k,l}\right) & si\ z_{j,k,l} > \tilde{z}_{j,k,l} + KAPPA\cdot\left(\tilde{z}^{U}_{j,k,l} - \tilde{z}_{j,k,l}\right). \end{cases}$$

La formule qui précède figure dans Norberg, A. et coll. (2010, p. 22 sous la définition 4.1). Les valeurs des paramètres Kappa et Tau sont fixées par l'utilisateur. On pourrait dire que Kappa définit l'« étendue» de l'intervalle d'acceptation. Une faible valeur de Kappa donne un degré de soupçon déjà supérieur à zéro pour un petit écart par rapport à la valeur prédite. Tau est utilisé pour ajuster le degré de soupçon. Par exemple, si Tau est égal à 0,001, le degré de soupçon devient zéro ou un. Si Tau est plus grand, par exemple égal à 10, le degré de soupçon devient proportionnel à la distance par rapport au point médian de la distribution. Afin de calculer le degré de soupçon dans Selekt, les quantités qui suivent sont nécessaires :

$\tilde{z}^{L}_{j,k,l}$ = Limite inférieure

$\tilde{z}^{U}_{j,k,l}$ = Limite supérieure

$\tilde{z}_{j,k,l}$ = Valeur prévue

$z_{j,k,l}$ = Valeur non vérifiée

Voici certains graphiques du degré de soupçon pour diverses valeurs des paramètres Kappa et Tau. Ces exemples sont fondés sur un ensemble d'observations historiques, mais le principe des caractéristiques du soupçon en fonction des valeurs fixées des paramètres est général.

Figure 3. Quand Kappa est égal à zéro, le degré de soupçon est presque toujours plus grand que zéro.

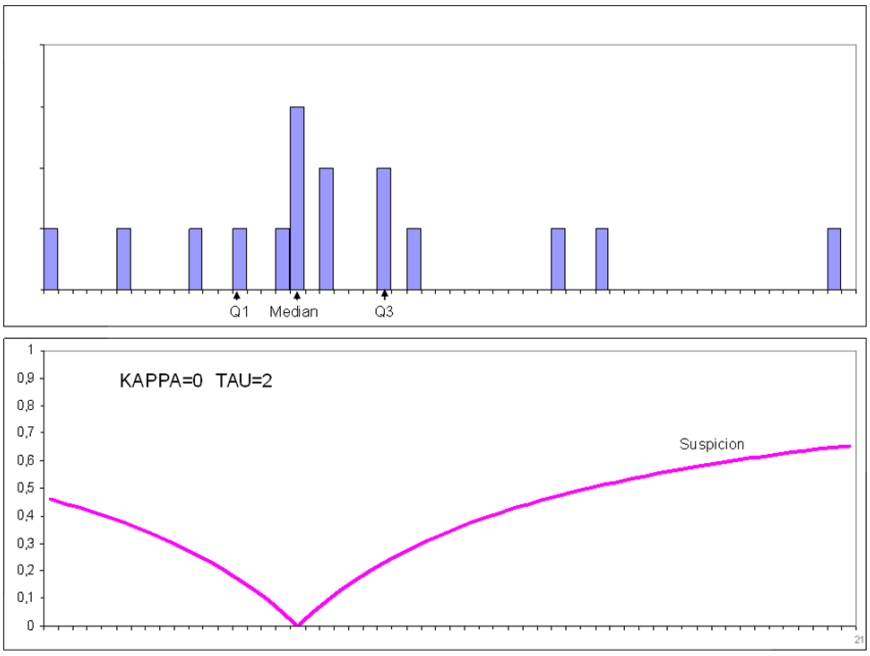


Figure 4. Ici, Kappa est égal à un au lieu de zéro comme ci-dessus.

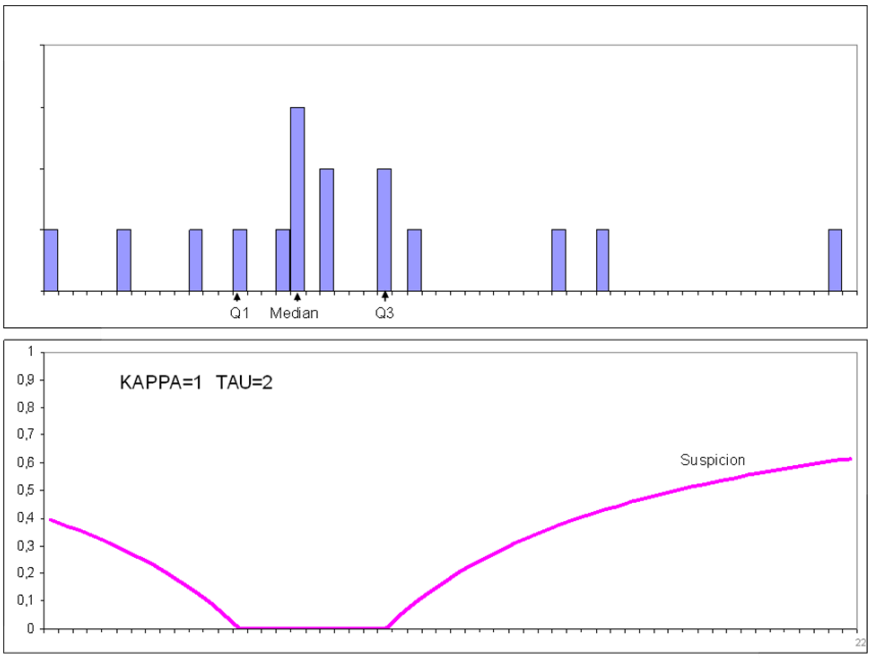


Figure 5. Kappa est égal à un et Tau est égal à cinq.

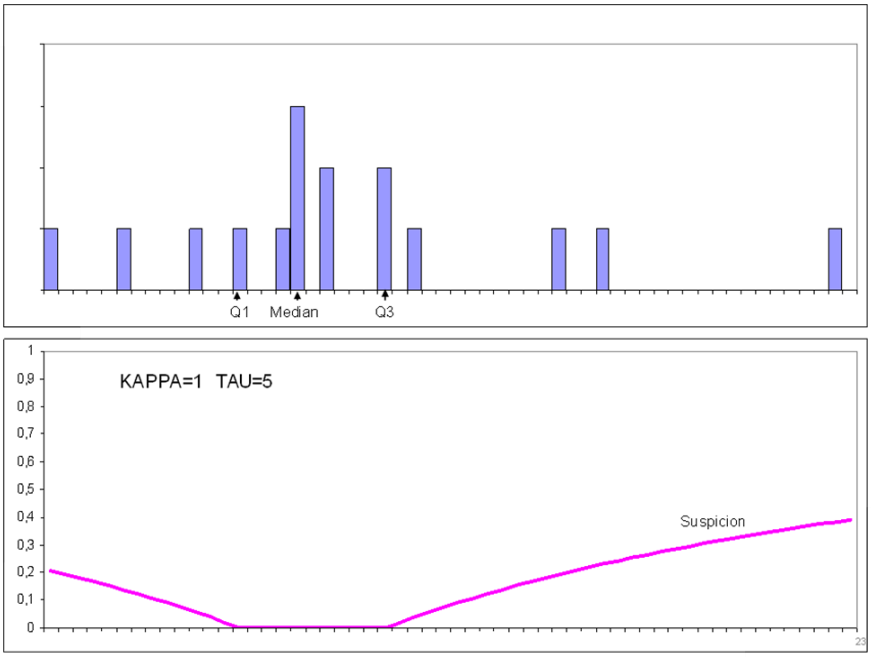
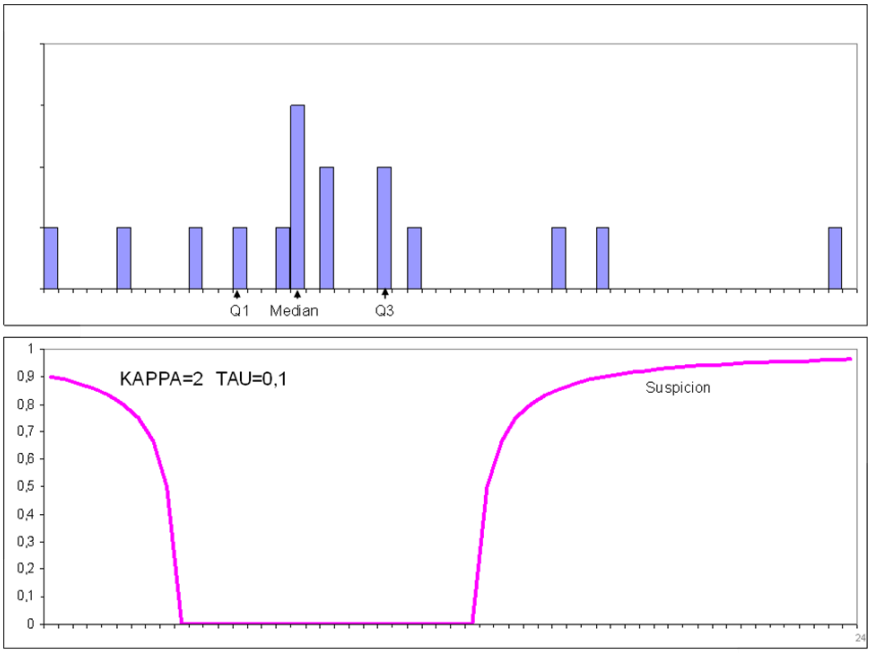


Figure 6. Quand Tau est faible, le degré de soupçon augmente abruptement.



2.3 Effet

Le concept d'effet a pour but de déterminer la mesure dans laquelle l'utilisation de la valeur non vérifiée au lieu de la valeur « corrigée » (ou du moins ce qui serait considéré comme la valeur correcte) aurait une incidence sur les statistiques produites. S'il n'y a aucune raison de penser que la valeur fournie contient une erreur, l'effet sera nul. Si, au contraire, la valeur non vérifiée diffère de la valeur correcte, la valeur absolue de l'effet sur les statistiques produites sera plus grande que zéro. Dans notre terminologie, il existe en fait plusieurs types d'effets. L'indice *nvér*

désigne la valeur non vérifiée, c'est-à-dire la valeur fournie par le répondant avant toute vérification. L'indice *vér* désigne la valeur vérifiée, c'est-à-dire la valeur corrigée. Très souvent, la valeur non vérifiée est la même que la valeur vérifiée. L'effet réel est défini comme étant w ($y_{nvér} - y_{vér}$) qui, pour une observation, est l'effet sur le total de domaine estimé de la variable y si l'on garde $y_{nvér}$ au lieu de procéder à un examen pour trouver $y_{vér}$. La mesure de l'effet contient le poids de sondage w, car l'on s'intéresse ici à la statistique ou au résultat produit. L'effet potentiel est défini comme w ($y_{nvér} - y_{préd}$) et est une approximation de l'effet réel destiné à être utilisé en pratique. La valeur $y_{préd}$ est une prédiction, c'est-à-dire la valeur prédite de $y_{vér}$, puisque cette dernière est inconnue tant qu'un examen n'a pas été effectué. L'effet prévu (par domaine, variable, observation) est le produit du degré de soupçon et de l'effet potentiel.

2.4 Fonction de score

Le score local produit en multipliant le degré de soupçon par l'effet potentiel peut être amélioré en ajoutant des paramètres concernant l'importance des variables ou des domaines. Le paramètre $Violon_j$ (valeur par défaut=1) permet d'ajuster l'importance de la variable j en multipliant le score par ce facteur. Le paramètre $Clarinette_{c(d)}$ (valeur par défaut=1) permet d'ajuster l'importance de la classification c, qui définit les domaines d, en multipliant de manière similaire le score par ce facteur. Le paramètre $Hautbois_j$ est un facteur d'ajustement pour la taille du total estimé ou son erreur-type pour la variable j. Le paramètre Violoncelle_alfa établit le lien entre l'effet et l'estimation du total ou l'erreur-type de ce total estimé. Les paramètres Violon, Clarinette, Hautbois et Violoncelle_alfa font tous partie du facteur d'ajustement complet appelé Violoncelle. Le produit du degré de soupçon et de l'effet potentiel est donc multiplié par le facteur Violoncelle de sorte que le score soit égal à $soupçon_{j,k,l} \times |effet\ potentiel_{d,j,k,l}| \times Violoncelle_{d(c),j}$ (pour le domaine *d*, la variable *j*, l'unité primaire d'échantillonnage *k* et l'unité secondaire d'échantillonnage *l*). La formule du paramètre Violoncelle est

$$VIOLONCELLE_{c(d),j} = \frac{VIOLON_j \times CLARINETTE_{c(d)}}{\left(\max imum\left\{ VIOLONCELLE\_ALFA_j \times \hat{T}_{d,j,t0}, SE\left(\hat{T}_{d,j,t0}\right)\right\}\right)^{HAUTBOIS_j}} .$$

Les scores locaux par domaine, variable, unité secondaire d'échantillonnage (s'il y a lieu) et unité primaire d'échantillonnage sont agrégés pour produire des scores globaux pour l'unité primaire d'échantillonnage et, finalement, pour l'unité répondante au besoin. Selekt offre un choix de méthodes d'agrégation. À l'heure actuelle, les options sont l'utilisation de la somme, de la somme des carrés ou du maximum. À chaque niveau d'agrégation (domaine, variable, *etc*.), Selekt offre à l'utilisateur un ensemble de seuils de manière qu'à chaque niveau d'agrégation, le score soit en fait le maximum de zéro ou du score moins le seuil.

3. Mise en œuvre de la vérification sélective des données à Statistics Sweden

L'objectif de la mise en œuvre de la vérification sélective à Statistics Sweden est de réduire le travail de vérification sans diminuer de manière significative la qualité des statistiques produites. L'objectif est également, dans le long terme, de réduire les coûts, même si un certain investissement doit être fait pour mettre en œuvre la vérification sélective dans une enquête. La liste qui suit donne une brève idée des enquêtes dans lesquelles la vérification sélective a été mise en œuvre jusqu'à présent à Statistics Sweden.

Enquête	Étape de mise en œuvre
Structures des traitements et salaires dans le secteur privé (SLP)	La vérification sélective en se servant de Selekt a été mise en œuvre.
Indicateurs d'activité des entreprises (Kortind)	La vérification sélective en se servant de Selekt a été mise en œuvre.
Statistiques conjoncturelles, traitements et salaires, secteur privé (KLP)	La vérification sélective en se servant de Selekt a été mise en œuvre.
Enquête sur les flux de biens et services (VFU) 2009	La vérification sélective en se servant de Selekt a été utilisée dans l'enquête.
Commerce extérieur – exportations et importations de services (UHT)	La vérification sélective en se servant de Selekt a été mise en œuvre.
Loyers des logements (HIB)	La vérification sélective en se servant de Selekt a été mise en œuvre.
Emploi à court terme (KS)	La vérification sélective en se servant de Selekt doit, en principe, être mise en œuvre au début de 2012.
Statistiques sur le chiffre d'affaires et les stocks pour les services	La vérification sélective en se servant de Selekt doit, en principe, être mise en œuvre au début de 2012.
Revenus et dépenses des immeubles à logements multiples (IKU)	La vérification sélective en se servant de Selekt doit, en principe, être mise en œuvre au début de 2012.

Bibliographie

Granquist, L. (1995), « Improving the Traditional Editing Process », dans Cox et coll. (éds), *Business Survey Methods*, Wiley.

Jäder, A. et A. Norberg (2006), « A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics », *Background facts on Economic Statistics 2006:3*, Statistics Sweden.

Norberg, A. et coll. (2010), « A General Methodology for Selective Data Editing », version 1.0 2010-02-04, rapport non publié, Statistics Sweden.

Norberg, A. et coll. (2011), « User's Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing », rapport non publié, Statistics Sweden.

SeleMix : Un progiciel R pour la vérification sélective au moyen de modèles de contamination

Marco Di Zio et Ugo Guarnera[1]

Résumé

Le but de la *vérification sélective* est de repérer dans les données les erreurs susceptibles d'avoir un effet important sur les estimations cibles afin de les corriger selon une procédure de vérification interactive précise. Cette tâche est habituellement accomplie en se servant de *fonctions de score* qui expriment l'importance des erreurs qui affectent les observations.

Les fonctions de score sont généralement définies en se fondant sur l'analyse des résidus par rapport à certaines prédictions découlant des données. Les modèles utilisés dans les méthodes classiques ne tiennent habituellement pas compte explicitement de la nature « intermittente » du processus d'erreur. Par conséquent, il est difficile de faire la distinction entre la composante de la variabilité observée associée aux écarts naturels par rapport à un comportement moyen et la composante due à la présence d'erreurs de mesure.

Afin de surmonter cette difficulté, une approche axée sur la modélisation explicite des données ainsi que du processus d'erreur a été proposée récemment. Elle repose sur l'utilisation de modèles de contamination, c'est-à-dire des modèles à classes latentes dans lesquels la variable latente doit être interprétée comme une variable indicatrice de l'occurrence des erreurs. Cette formalisation permet de relier les valeurs de la fonction de score aux erreurs prévues dans les données.

Afin de faciliter l'utilisation des modèles de contamination, un progiciel R dénommé *SeleMix* a été mis en œuvre récemment. L'article présente le modèle et illustre les principales fonctionnalités de *SeleMix*.

Mots clés : Erreurs influentes ; fonction de score ; modèles à mélange de lois.

1. Introduction

La vérification sélective a pour principe de rechercher les unités présentant des erreurs importantes afin de pouvoir limiter les procédures de vérification précises à ce sous-ensemble d'unités. Le but est de réduire le coût de la phase de vérification et d'imputation tout en maintenant un niveau acceptable de qualité des estimations (Lawrence et McKenzie, 2000; Lawrence et McDavitt, 1994). En pratique, l'ordre de priorité des observations est établi en fonction des valeurs d'une fonction de score qui mesure l'importance des erreurs que ces observations peuvent contenir (Latouche et Berthelot, 1992; Hedlin, 2003). Les unités dont le score est supérieur à un seuil donné sont sélectionnées pour être vérifiées avec précision.

Les méthodes employées le plus souvent pour déterminer les scores reposent sur la comparaison des valeurs observée et prédite (Hedlin, 2008). Ce résidu est composé de l'erreur possible et de la variabilité naturelle de la quantité analysée. Dans les conditions habituelles, il n'est pas possible de distinguer ces deux éléments, de sorte que le score d'une observation n'est pas relié directement à l'erreur prévue pour l'unité en question. Par conséquent, la valeur du seuil ne peut pas être interprétée directement comme étant le niveau d'exactitude des estimations finales et une règle d'arrêt pour déterminer le sous-ensemble d'unités qu'il faut sélectionner ne sera disponible que dans un contexte de simulation, quand les données vérifiées (considérées comme exactes) et les données brutes d'une ancienne enquête sont disponibles (de Waal et coll., 2011).

Di Zio et coll. (2008) ont proposé d'utiliser un modèle à variables latentes qui permet, sous certaines hypothèses, d'estimer l'erreur prévue associée à chaque unité. Dans ces conditions, la valeur du seuil est directement interprétée comme étant le niveau d'exactitude des données vérifiées permettant d'établir une règle d'arrêt reliée à la quantité

[1]Marco Di Zio, Istat – Istituto Nazionale di Statistica, Rome - Via Cesare Balbo 16, Italie, 00164 (courriel : dizio@istat.it); Ugo Guarnera, Istat – Istituto Nazionale di Statistica, Rome - Via Cesare Balbo 16, Italie, 00164 (courriel : guarnera@istat.it).

d'erreurs encore présentes dans les données. La méthode est fondée sur l'utilisation de modèles normaux de contamination dans lesquels les données erronées sont supposées suivre la même loi que les données dépourvues d'erreurs, mais en ayant une plus grande variance (voir Ghosh-Dastidar et Schafer, 2006).

Les résultats d'expériences portant sur des données simulées ainsi que réelles montrent que la vérification sélective fondée sur des modèles de contamination est utile dans de nombreux contextes (Bellisai et coll., 2009; Buglielli et coll., 2010).

Afin de faciliter l'usage de la méthode, un progiciel R nommé *SeleMix* a été mis en œuvre. Il est disponible en libre accès sur le site Web du projet R. Les fonctions incluses dans le progiciel permettent également d'appliquer la procédure à des données présentant des valeurs manquantes. Dans ce cas, pour chaque enregistrement incomplet, les éléments de données manquants sont remplacés par leurs valeurs prévues conditionnellement aux valeurs observées. Les espérances sont calculées d'après le modèle de contamination, de sorte que les prédictions tiennent compte de l'existence possible d'erreurs dans les éléments de données observés. En ce sens, l'approche du modèle de contamination peut aussi être utilisée comme une méthode d'imputation « robuste ».

La présentation de l'article est la suivante. Aux sections 2 et 3, nous décrivons brièvement le modèle de contamination et l'approche de vérification sélective, respectivement; des renseignements plus détaillés peuvent être consultés dans Buglielli et coll., 2011. La section 4 est consacrée à l'illustration de *SeleMix* et de ses principales fonctionnalités.

2. Le modèle

Les éléments essentiels de l'approche présentée ici sont 1) la spécification d'un modèle paramétrique pour les données exactes (non contaminées) et 2) la spécification d'un modèle d'erreur. Cela nous permet de déterminer la distribution des données exactes conditionnellement aux données observées. Cette distribution joue un rôle central dans la méthode de vérification sélective. L'un des aspects importants est le fait que la spécification du modèle reflète la nature intermittente du processus d'erreur. Autrement dit, on suppose que les erreurs n'affectent qu'un sous-ensemble des données, c'est-à-dire que chaque unité de l'ensemble de données possède une certaine probabilité a priori (inconnue) de présenter une erreur. L'hypothèse d'une erreur intermittente, qui est très fréquente dans le contexte du traitement des données d'enquête, mène naturellement à la spécification du modèle d'erreur fondé sur un mélange de lois de probabilité. Par conséquent, la distribution des données observées est également un mélange dont les composantes correspondent aux données dépourvues d'erreur et aux données contaminées, respectivement. Les modèles de ce genre, souvent appelés modèles de contamination, sont appliqués fréquemment dans le contexte de la détection des valeurs aberrantes. Nous allons maintenant décrire le modèle plus en détail.

2.1 Modèle pour les données exactes

Nous supposons que deux ensembles de variables sont observés : les variables du premier groupe, disons les variables X, que nous supposons être mesurées correctement, et les variables du second groupe, disons les variables Z, qui correspondent aux éléments susceptibles d'être affectés par des erreurs de mesure. Dans ces conditions, qui peuvent être utiles quand les données sur certaines variables proviennent de sources administratives ou sont mesurées avec une grande précision, il est assez naturel de traiter les variables observées avec une erreur comme des variables réponse et les variables fiables, comme des covariables. Ce cadre englobe le cas particulier où les covariables fiables X ne sont pas disponibles de sorte que l'on doit modéliser la distribution conjointe des variables Z.

Dans la suite, nous modélisons les données exactes au moyen d'une loi de probabilité log-normale. Cela semble une hypothèse raisonnable dans de nombreux cas où des données économiques sont analysées.

Étant donné les hypothèses qui précèdent, les données exactes correspondant aux éléments contaminés sont représentées comme une matrice Z^* de dimensions $n \times p$ de n réalisations indépendantes d'un vecteur aléatoire de taille p que l'on suppose suivre une loi log-normale dont les paramètres peuvent dépendre d'un certain ensemble de q covariables non affectées d'une erreur. Donc, si $Y^*=\ln Z^*$, nous avons le modèle de régression :

$$Y^* = XB + U \quad (1)$$

où X est une matrice de dimensions $n \times q$ dont les lignes sont les mesures des q covariables sur les n unités, B est la matrice de dimensions $q \times p$ des coefficients, et U est la matrice de dimensions $n \times p$ des résidus normaux dont la i^e ligne U_i suit une loi normale de moyenne nulle et de matrice de covariance Σ :

$$U_i \sim N(0, \Sigma), \quad i=1,..,n. \quad (2)$$

2.2 Modèle d'erreur

Afin de modéliser la nature intermittente du processus d'erreur, nous introduisons une variable aléatoire bernoullienne I avec le paramètre π, où $I=1$ si une erreur a lieu et $I=0$ autrement. Dans la suite, Z et Y désigneront une variable contaminée possible sur l'échelle originale et sur l'échelle logarithmique, respectivement. Donc, si $I=0$, l'expression $Z=Z^*$ ($Y=Y^*$) doit être vérifiée. En outre, si $I=1$, les erreurs affectent les données selon un processus additif représenté par une variable aléatoire gaussienne de moyenne nulle et de matrice de covariances Σ_ε proportionnelle à Σ, c'est-à-dire étant donné $\{I=1\}$:

$$Y = Y^* + \varepsilon, \qquad \varepsilon \sim N(0, \Sigma_\varepsilon), \quad \Sigma_\varepsilon = \lambda\Sigma, \quad \lambda > 0.$$

Il est commode de représenter le modèle d'erreur au moyen de la distribution conditionnelle :

$$f_{Y|Y^*}(y \mid y^*) = (1-\pi)\delta(y - y^*) + \pi N(y; y^*, \Sigma_\varepsilon) \quad (3)$$

où π (*poids de mélange*) est la probabilité « a priori » de contamination et $\delta(t'-t)$ est la fonction delta avec la masse en t.

Dans le cas où l'ensemble de variables X est vide, les variables Y_i ($i=1,\ldots,n$) suivent une loi normale ayant un vecteur de moyennes commun μ. Il convient de souligner qu'étant donné l'hypothèse d'erreur intermittente, il est conceptuellement possible d'imaginer les données comme étant partitionnées en données exactes et erronées, et d'estimer, pour chaque observation, la probabilité qu'elles soient exactes ou corrompues. La distribution des données observées s'obtient facilement en multipliant la densité de probabilité normale des données exactes impliquées par (1) et (2) et la densité de probabilité des erreurs (3), puis en intégrant sur Y^* **:**

$$f_Y(y) = (1-\pi)N(y; B'x, \Sigma) + \pi N(y; B'x, (\lambda+1)\Sigma). \quad (4)$$

La distribution (4) a trait aux données observées et peut être estimée facilement en maximisant la vraisemblance fondée sur n unités de l'échantillon en se servant de l'algorithme ECM.

3. Vérification sélective

Afin d'utiliser le modèle de contamination pour la vérification sélective, nous devons déterminer la distribution des données dépourvues d'erreur Y^* conditionnellement aux données observées (y compris les covariables X).

Une application directe de la formule de Bayes donne :

$$f_{Y^*|X,Y}(y^* \mid x, y) = \tau_1(x, y)\delta(y^* - y) + \tau_2(x, y)N(y^*; \tilde{\mu}_{x,y}, \tilde{\Sigma}) \quad (5)$$

où τ_1 et τ_2 sont les probabilités a posteriori de faire partie des données exactes et des données erronées, respectivement :

$\tau_1(x_i, y_i) = \Pr(y_i = y_i^* \mid x_i, y_i)$

$\tau_2(x_i, y_i) = \Pr(y_i \neq y_i^* \mid x_i, y_i) = 1 - \tau_1(x_i, y_i), \quad i=1,\ldots,n$

et

$$\tilde{\mu}_{x,y} = \frac{(y + \lambda B' x)}{\lambda + 1}; \quad \tilde{\Sigma} = (\frac{\lambda}{\lambda + 1})\Sigma.$$

La détermination de la distribution conditionnelle correspondante sur l'échelle originale est immédiate :

$$f_{Z^*|Z}(z^* \mid z) = \tau_1(\ln(z))\delta(z^* - z) + \tau_2(\ln(z))LN(z^*; \tilde{\mu}_{x,\ln z}, \tilde{\Sigma}) \qquad (6)$$

où $LN(\cdot, \mu, \Sigma)$ désigne la densité de probabilité log-normale ayant les paramètres (μ, Σ) et, pour simplifier, nous avons supprimé les variables X de la notation chaque fois qu'elles figurent comme variables de conditionnement. L'estimation de la distribution (6) s'obtient en substituant les estimations de $(\mu, \Sigma, \pi, \lambda)$ résultant de l'algorithme ECM aux paramètres correspondants.

Une fois que la distribution cible (6) est estimée, les « prédictions » des valeurs « exactes » z_i^*, conditionnellement aux valeurs observées z_i, peuvent être obtenus pour toutes les observations $i=1,\ldots,n$ selon :

$$\hat{z}_i = E(z_i^* \mid z_i) = \int z_i^* f_{Z^*|Z}(z^* \mid z) dz_i^*$$

Donc, pour $i = 1,\ldots n$, l'*erreur prévue* ε_i peut également être définie comme étant :

$$\varepsilon_i = (\hat{z}_i - z_i) = \tau_2(\ln(z_i))(z_i - \tilde{\mu}_{x_i,\ln z_i}). \qquad (7)$$

Dans le contexte de la statistique officielle, les estimations de certaines quantités (telles que les totaux et les moyennes) d'une population finie U présentent habituellement un intérêt. L'approche du modèle de contamination peut être combinée à l'inférence sous randomisation pour obtenir des estimations robustes. Concrètement, supposons que l'estimation cible est donnée par le total T_z de la variable Z, c'est-à-dire $T_z = \sum_{i \in U} z_i$ et qu'on utilise un estimateur $\hat{T}_z = \sum_{i \in S} w_i z_i$, où w_i désigne les poids d'échantillonnage associés à chaque unité d'un échantillon S de taille n. Une version robuste de $\hat{T}_z$ est donnée par $\hat{T}_z^* = \sum_{i \in S} w_i \hat{z}_i$, où le dernier estimateur est obtenu à partir du précédent en remplaçant les valeurs observées z_i par les prédictions $\hat{z}_i$.

Utiliser le modèle de contamination directement pour l'estimation est une proposition séduisante, mais les résultats pourraient être trop sensibles aux écarts par rapport aux hypothèses de modélisation. Dans le présent contexte, nous nous intéressons à la vérification sélective. Donc, nous définissons une fonction de score en nous fondant sur la valeur absolue de l'erreur prévue estimée selon le modèle de contamination.

Cette définition est particulièrement utile en ce sens qu'elle permet d'estimer l'erreur résiduelle restantes dans les données après la vérification des unités pour lesquelles les erreurs prévues sont les plus grandes. Il s'ensuit que le nombre d'unités à examiner peut être choisi de manière que l'erreur résiduelle soit inférieure à un seuil prédéterminé. En particulier, le seuil peut être défini en s'appuyant sur le ratio entre l'erreur résiduelle prévue et une estimation de référence (robuste) telle que celle fournie par le modèle de contamination proprement dit. Afin de définir la fonction de score, introduisons l'erreur individuelle relative r_i comme étant le ratio entre l'erreur prévue (pondérée) et l'estimation de référence T_z^*

$$r_i = \frac{w_i(\hat{z}_i - z_i)}{\hat{T}_z^*}.$$

Notons que l'erreur prévue qui figure dans la formule susmentionnée est le produit, selon la formule (7), de $\tau_2(\ln(z_i))$ et $(z_i - \tilde{\mu}_{x_i,\ln z_i})$, qui peuvent être interprétés comme la composante de risque et la composante d'influence, respectivement (Jäder et Norberg, 2005).

La fonction de score est définie comme $SF_i = | r_i |$. En outre, soit R_M la valeur absolue du pourcentage résiduel prévu d'erreurs dans les données après avoir éliminé les erreurs dans les unités appartenant à l'ensemble *M* (le pseudo-biais absolu, Latouche et Berthelot, 1992) :
$R_M = | \sum_{i \in \bar{M}} r_i |$, où $\bar{M}$ désigne le complément de *M* dans *S*.
Lorsqu'un seuil d'« exactitude » η est choisi, la procédure de vérification sélective consiste à :

1. trier les observations par ordre décroissant en fonction de la valeur de SF_i;
2. choisir les $\bar{k}$ premières unités pour l'examen, où :
$\bar{k} = \min\{k \in (1,..,n) \mid R_{M_j} < \eta, \forall j > k\}$ et M_m dans l'ensemble composé des *m* premières unités.

Ces deux étapes garantissent une borne supérieure non seulement pour la quantité totale d'erreur résiduelle dans les données (à savoir le seuil η), mais aussi pour l'erreur affectant chaque observation individuelle non vérifiée. En fait, il est facile de confirmer que SF_i est inférieure à 2η pour $i > \bar{k}$.

L'algorithme décrit jusqu'à présent peut être étendu facilement au cas multivarié en définissant une fonction de score globale GS_i comme étant $max_p\ SF_{i,p}$, où $SF_{i,p}$ est la fonction de score locale pour la i^e unité et la p^e variable. Ce score global fait en sorte que les propriétés mentionnées antérieurement sont encore vérifiées pour chaque variable d'intérêt.

4. Le progiciel R *SeleMix*

Afin de mettre en œuvre la méthode de vérification sélective fondée sur un modèle de contamination, nous avons développé des fonctions R et les avons incluses dans un progiciel. Ce dernier peut aussi être utilisé quand la variable contaminée multivariée *Y* contient des valeurs manquantes. Dans ces cas, il fournit des prédictions robustes pour les valeurs manquantes. Le logiciel permet également d'inclure dans le modèle un ensemble de variables « épurées » *X* comme variables explicatives. Cette caractéristique est utile, entre autres, quand de l'information auxiliaire (par exemple, des données administratives ou historiques) est disponible.

L'élément de base du progiciel est composé de trois fonctions `ml.est`, `pred.y`, `sel.edit`. Des outils graphiques sont également disponibles.

La principale sortie du progiciel est l'identification des unités critiques correspondant aux erreurs les plus influentes étant donné un seuil d'exactitude prédéterminé. Nous allons maintenant décrire plus en détail les fonctions du progiciel *SeleMix*.

`ml.est`. Cette fonction estime les paramètres $\theta = (B, \Sigma, \pi, \lambda)$ sur les données observées en utilisant l'algorithme ECM. Elle donne aussi les valeurs « anticipées » (prédictions) pour les variables *Y* pour toutes les observations.

L'entrée de la fonction `ml.est` est la matrice de données observées et, facultativement, la matrice des covariables *X*.

Par défaut, l'algorithme débute avec les valeurs $\lambda = 3$ et $\pi = 0{,}05$, mais il est possible de définir d'autres points de départ. L'utilisateur doit spécifier si les données exactes sont censées suivre une loi normale ou log-normale. Dans le second cas, les zéros dans les données sont remplacés par une faible valeur (10E-8) et un avertissement est produit.

Les valeurs de départ des coefficients de régression *B* et de la matrice de covariance Σ sont calculées sur les données d'entrée *Y* et *X* par la méthode des moindres carrés ordinaire (c'est-à-dire comme si elles étaient exemptes d'erreur). L'algorithme EM consiste à appliquer de manière répétée les étapes de calcul de l'espérance et de maximisation jusqu'à la convergence ou jusqu'à ce que soit atteint un nombre maximal d'itérations spécifié par l'utilisateur.

La fonction calcule, pour chaque unité, la probabilité a posteriori τ qu'elle appartienne à la composante du mélange de lois correspondant aux données contaminées. Cette probabilité est utilisée pour définir un indicateur de valeur aberrante qui est égal à 1 si τ est plus grande qu'un seuil spécifié (par défaut égal à 0,5), et 0 autrement.

La fonction produit les valeurs du critère d'information bayésien BIC (pour *Bayesian Information Criterion*) et du critère d'information d'Akaike AIC (pour *Akaike Information Criterion*) afin d'évaluer la qualité de l'ajustement du modèle de mélange de lois par opposition au modèle normal centré réduit.

Cette information aide l'utilisateur à évaluer la validité de l'utilisation d'un modèle de mélange de lois.

La sortie de la fonction `ml.est` est fournie sous forme d'une liste dont les éléments sont les paramètres du modèle θ, les valeurs prévues, les scores BIC et AIC, les indicateurs de valeur aberrante et les probabilités a posteriori τ.

La fonction `ml.est` comprend un appel de la fonction `pred.y` qui calcule les prédictions pour les variables *Y*.

`pred.y`. Cette fonction a pour objectif d'estimer la distribution des données exactes, conditionnellement aux variables réponses observées et aux covariables. Les entrées nécessaires sont les paramètres $\theta = (B, \Sigma, \pi, \lambda)$ et un ensemble de données observées. Il convient de souligner que des valeurs manquantes ne sont pas permises pour les variables *X*. La fonction produit, pour chaque unité, une « prédiction » pour les données observées ainsi que manquantes pour chaque variable *Y*, l'indicateur de valeur aberrante et la probabilité a posteriori τ.

`sel.edit`. Cette fonction classe les observations par ordre de priorité d'après les valeurs de la fonction de score et signale les unités qui doivent être vérifiées, de sorte que l'erreur résiduelle prévue soit inférieure à un seuil d'exactitude prédéterminé.

Il convient de mentionner que `sel.edit` peut être utilisée indépendamment des autres fonctions de *SeleMix*. En fait, la détection des unités influentes peut être exécutée quel que soit le modèle utilisé pour la prédiction.

Comme données d'entrée, la fonction reçoit la matrice de données observées et la matrice de valeurs prévues correspondante, le total de référence estimé de chaque variable *Y*, les poids d'échantillonnage et le seuil prédéterminé d'exactitude.

Le total de référence de *Y* est facultatif; s'il est omis, il est calculé comme étant la somme pondérée des valeurs prédites. Les poids sont supposés égaux à 1 sauf indication contraire, et le seuil d'exactitude par défaut est 0,01.

Les unités influentes sont sélectionnées en fonction des valeurs d'un score global calculé comme il suit. Premièrement, un score local pour une variable donnée est défini comme étant l'écart absolu pondéré entre la valeur observée et la valeur prévue, en normalisant par rapport au total de référence estimé. Ensuite, le score global est obtenu en calculant le maximum des scores locaux, puis les observations sont classées en fonction des valeurs décroissantes du score global.

La dernière étape consiste à trouver les *k* premières unités, telles que, pour toutes les variables, l'erreur résiduelle totale (prévue) dans les (*n*-*k*) autres unités soit inférieure au seuil prédéterminé.

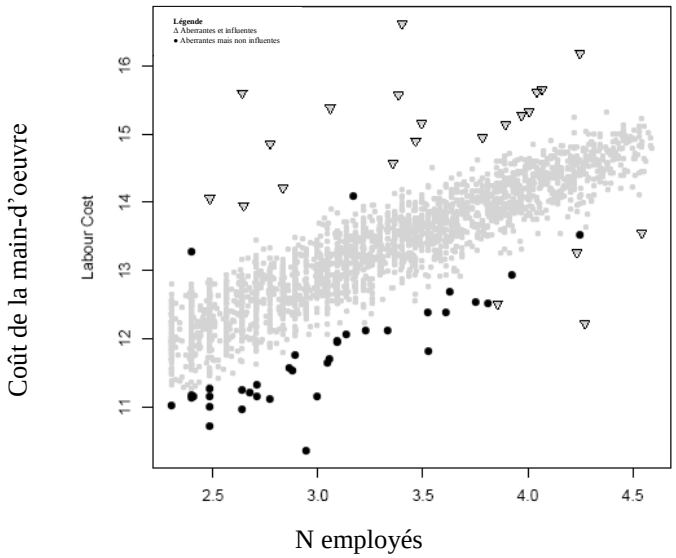
La sortie de `sel.edit` est une matrice contenant l'indicateur des unités influentes, le rang conformément au score global, les scores globaux et locaux, ainsi que l'erreur résiduelle cumulée pour chaque variable.

Les deux figures qui suivent, obtenues au moyen d'un outil graphique disponible dans *SeleMix*, montrent les valeurs aberrantes par opposition aux erreurs influentes, et les erreurs résiduelles estimées par opposition à réelles pour une expérience exécutée sur des données provenant d'une enquête économique d'Istat. Les détails de l'expérience peuvent être consultés dans Buglielli et coll., 2011.

À la figure 1, les observations représentées par des triangles gris sont celles classées comme des erreurs influentes, tandis que les points noirs sont les observations classées comme des valeurs aberrantes, mais non influentes. La sélection est faite en se basant sur un seuil η égal à 0,005. Dans cette application, toutes les erreurs influentes sont

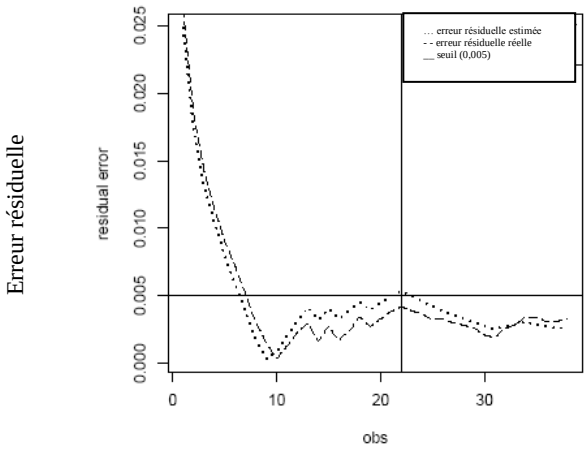
des valeurs aberrantes, mais nous constatons également que certaines valeurs aberrantes ne sont pas des erreurs influentes. Il s'agit d'une particularité importante de la vérification sélective qui permet d'économiser des ressources pour la révision des données. En fait, même si les observations sont classées comme des erreurs, leur incidence sur les estimations est négligeable en ce qui a trait au niveau choisi d'exactitude.

Figure 1. Valeurs aberrantes et erreurs influentes pour un seuil η égal à 0,005, sur une échelle logarithmique.



À la figure 2, la courbe en trait interrompu montre l'erreur résiduelle réelle, tandis que la courbe en pointillé donne l'erreur résiduelle estimée pour la variable *Coût de la main-d'œuvre* sur le sous-ensemble des 40 premières observations. Toutes les unités situées à gauche de la droite verticale sont les observations influentes. Nous constatons que, pour certaines unités, les courbes de l'erreur résiduelle cumulée réelle ainsi qu'estimée se situent en dessous du seuil prédéterminé avant d'arriver à la dernière observation considérée comme influente. Cela tient au fait que l'erreur cumulée est calculée sur la différence entre les valeurs observées et prévues, et que les valeurs peuvent s'annuler les unes les autres. Nous rappelons que le critère d'arrêt fait en sorte que l'erreur résiduelle soit inférieure à un certain seuil d'exactitude d'après la dernière unité sélectionnée.

Figure 2. Erreur résiduelle estimée (trait pointillé) et réelle (trait interrompu) pour un seuil η égal à 0,005.



Bibliographie

Bellisai, D., Di Zio, M., Guarnera, U. et O. Luzi (2009), « A Selective Editing approach based on contamination models: An application to an Istat business survey », *UNECE Work Session on Statistical Data Editing.* Neuchatel, 5-7 octobre 2009.

Buglielli, M.T., Di Zio, M. et U. Guarnera (2010), « Use of Contamination Models for Selective Editing », *Q2010, European Conference on Quality in Survey Statistics*, 4-6 mai 2010, *Helsinki*.

Buglielli, M.T., Di Zio, M., Guarnera, U. et F.R. Pogelli (2011), « Selective Editing of Business Survey Data Based on Contamination Models: An Experimental Application », *NTTS 2011 New Techniques and Technologies for Statistics*, Bruxelles, 22-24 février 2011.

de Waal, T., Pannekoek, J. et S.Scholtus S. (2011), *Handbook of Statistical Data Editing and Imputation*, New Jersey, Wiley.

Di Zio, M., Guarnera, U. et O. Luzi (2008), « Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data », *UNECE Work Session on Statistical Data Editing*, Vienna.

Ghosh-Dastidar, B. et J.L. Schafer (2006), « Outlier Detection and Editing Procedures for Continuous Multivariate Data », *Journal of Official Statistics*, Vol. 22, n°. 3, 2006, pp. 487-506.

Hedlin, D. (2003), « Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics », *Journal of Official Statistics*, Vol. 19, n° 2, pp. 177-199.

Hedlin, D. (2008). « Local and global score functions in selective editing », *UNECE Work Session on Statistical Data Editing*, Vienna.

Jäder, A. et A. Norberg (2005), « A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics », *UNECE Work Session on Statistical Data Editing*, Ottawa.

Latouche, M. et J.M. Berthelot (1992), « Use of a score function to prioritize and limit recontacts in editing business surveys », *Journal of Official Statistics*, 8, n° 3, pp. 389- 400.

Lawrence, D. et C. McDavitt (1994), Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics*, Vol. 10, n° 4, pp. 437-447.

Lawrence, D. et R. McKenzie (2000), "The General Application of Significance Editing", *Journal of Official Statistics*, 16, n° 3, pp. 243-253.

Méthodes et outils de vérification sélective – Perspective de l'Australian Bureau of Statistics

Eden Brinkley, Keith Farwell et Frank Yu[1]

Résumé

Au début de la dernière décennie, l'Australian Bureau of Statistics a lancé un important programme de refonte en vue de transformer ses processus, méthodologies et technologies de production des statistiques économiques. La vérification des données faisait partie des premiers processus examinés et l'une des nouvelles méthodologies adoptées a été la vérification sélective des microdonnées à l'aide d'un outil créé à l'interne appelé Significance Editing Engine (moteur de vérification selon l'importance). L'article décrit la méthodologie qui sous-tend notre approche de vérification sélective des microdonnées et le moteur de vérification selon l'importance, et se termine par un résumé des principaux enseignements que nous avons tirés de l'expérience jusqu'à présent et de l'évolution de notre processus de réflexion.

Mots clés : Vérification sélective ; vérification selon l'importance ; moteur de vérification selon l'importance ; refondre ; transformer ; messages.

1. Introduction

Au début de la dernière décennie, l'Australian Bureau of Statistics (ABS) a lancé un important programme de refonte appelé Business Statistics Innovation Program (programme d'innovation de la statistique des entreprises). L'objectif du programme était de transformer les processus, les méthodologies et les technologies employés pour produire les statistiques économiques de l'ABS. D'importants travaux ont été entrepris en vue d'examiner les processus opérationnels dans le contexte d'un cadre complet et cohérent de haut niveau. En outre, ces travaux visaient à remplacer les processus disparates par un plus petit nombre d'approches communes et axées sur des meilleures pratiques. De nouvelles méthodologies ont également été adoptées pour réduire les coûts ainsi que le fardeau imposé aux fournisseurs de données et pour améliorer la qualité des données.

L'un des premiers processus examinés était la vérification des données et l'une des nouvelles méthodologies adoptées était la vérification sélective des microdonnées à l'aide d'un outil créé à l'interne appelé Significance Editing Engine (SigEE) (moteur de vérification selon l'importance). Le présent article décrit la méthodologie de l'ABS qui sous tend l'approche de vérification des microdonnées selon l'importance et le SigEE. Sont aussi exposées en détail les fonctionnalités offertes par le SigEE et la façon dont il répond à diverses exigences de vérification des microdonnées. Un nouvel outil de vérification des macrodonnées selon l'importance, qui s'appuie sur des scores d'importance pour détecter les estimations anormales, est également décrit. L'article se termine par un résumé des principaux enseignements que nous avons tirés de l'expérience jusqu'à présent et de l'évolution du processus de réflexion de l'ABS au cours des dix dernières années.

2. Qu'est-ce que la vérification selon l'importance?

La vérification selon l'importance est une forme de vérification sélective fondée sur le principe voulant que, s'il est possible de prédire l'effet des interventions de vérification sur les résultats que l'on tente d'obtenir, on peut alors décider de ce qu'il faut vérifier et de la portée que doit avoir la vérification (Farwell et Raine, 2000). Concrètement, il s'agit de cerner et de classer par ordre de priorité les différences importantes entre ce que l'on observe dans les

[1]Eden Brinkley, Australian Bureau of Statistics, Keith Farwell, Australian Bureau of Statistics, Frank Yu, Australian Bureau of Statistics, Locked Bag 10, Belconnen, ACT, Australie, 2616.
Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas forcément celles de l'Australian Bureau of Statistics.

données et ce que l'on attend de celles-ci. L'importance de la différence est évaluée en fonction de son incidence sur les données de sortie prévues. Des scores sont créés en se fondant sur des mesures de l'effet prévu de la vérification. Ces scores peuvent être utilisés pour donner un ordre de priorité aux éléments de données, aux enregistrements fournis ou aux estimations considérés comme étant anormaux. Les données sont classées en fonction de la valeur du score et des seuils sont appliqués pour sélectionner les données anormales. Le choix de la valeur des seuils repose sur une analyse coûts-avantages permettant de trouver un compromis entre les coûts et les avantages de la vérification. Une fois que les anomalies importantes dans les données ont été décelées, des processus doivent être mis en place pour déterminer leur nature et mettre en œuvre une méthode de traitement.

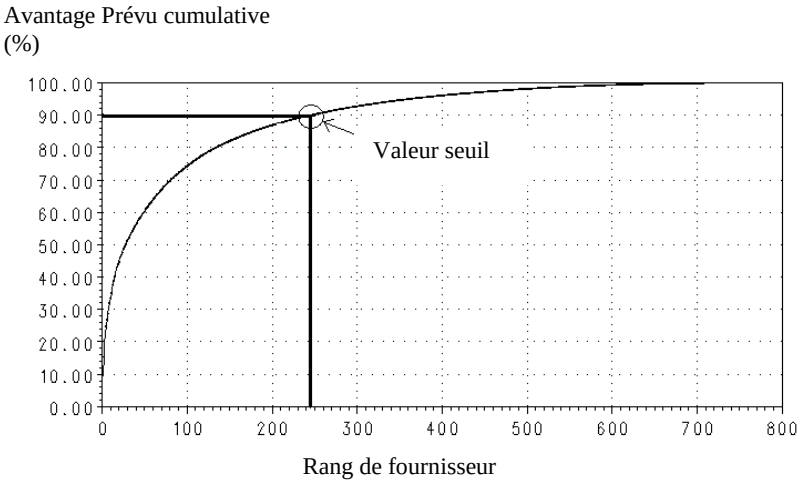
La vérification selon l'importance a été élaborée au départ pour la vérification des microdonnées, mais le cadre général a maintenant été élargi afin d'inclure la vérification sélective des macrodonnées. L'ABS a entrepris l'élaboration d'un outil complémentaire de vérification des macrodonnées selon l'importance à ajouter au SigEE (Farwell, 2009). Il sera utilisé pour détecter les estimations anormales.

3. Concepts de la vérification selon l'importance

L'idée qui sous-tend la vérification des microdonnées selon l'importance consiste à sélectionner un ensemble de variables clés (c'est-à-dire des items clés) qui entrent dans un ensemble d'estimations clés à un niveau particulier (appelé *niveau cible* ou *d'importance*). À chaque valeur d'un item clé dans l'enregistrement d'une unité est attribué un *score d'item*. Les scores d'item sont calculés en partant des espérances des items clés et des estimations cibles (appelées *valeurs prévues* et *estimations prévues*, respectivement). La valeur de la réponse à l'item, la valeur prévue et l'estimation prévue (et d'autres attributs tels que les poids de sondage) sont utilisées pour prédire l'effet de la vérification. L'effet normalisé de la vérification est également appelé *avantage* et le score est une estimation de l'*avantage prévu* (les termes « score » et « avantage prévu » sont utilisés indifféremment dans le présent article).

Le cadre de vérification selon l'importance permet de combiner les scores locaux (tels que les scores d'item) en utilisant une métrique pour créer des scores globaux. Par exemple, il est souvent préférable d'attribuer un score aux fournisseurs de données plutôt qu'aux éléments de données. Chaque enregistrement fournisseur possédera plusieurs scores d'item (un pour chaque item clé) qui peuvent être combinés en un score global pour le fournisseur en question (*score de fournisseur*). Les tailles des scores locaux et globaux peuvent être utilisées pour créer des classements locaux et globaux. Les scores de fournisseur sont utilisés pour classer les fournisseurs de données plutôt que les éléments de données. Par exemple, si l'on se sert d'un questionnaire pour recueillir les données, le score de fournisseur peut être utilisé pour classer les questionnaires remplis. Des seuils sont ensuite utilisés pour sélectionner les enregistrements anormaux. Un exemple de seuil élémentaire est celui où les données sont considérées anormales si le score qui y est associé est supérieur à la valeur seuil. En outre, des graphiques représentant l'avantage prévu cumulatif en fonction du rang peuvent être utilisés pour choisir les seuils « au fur et à mesure ». Les graphiques sont appelés *courbes coûts-avantages*. Ces dernières, qui sont une forme de courbe de Lorenz, peuvent être classées à l'aide d'un coefficient GINI. Elles illustrent le compromis entre le coût et l'avantage de la vérification. Dans la figure 3-1 qui suit, environ 90 % de l'avantage prévu est dû à environ 250 des 800 fournisseurs soumis au SigEE.

Figure 3-1
Exemple de courbe coûts-avantages



4. Scores d'importance

4.1 Score d'importance générique

Le score d'importance générique (Farwell, 2009), qui peut être utilisé pour créer des scores d'importance pour les microdonnées ou les macrodonnées, prend la forme suivante :

$$\text{Score générique} = 100 * \left| \frac{\text{Effet prévu de la vérification}}{\text{Valeur de normalisation}} \right| \qquad (1)$$

Effet de la vérification = Estimation cible prévue corrigé – Estimation cible prévue (2)

On peut considérer que le score générique (1) « cible » une estimation particulière. La mesure de l'effet de la vérification (2) est définie en fonction des différences entre les données prévues et observées. Dans le cas de la vérification de microdonnées, il faut aussi tenir compte de la probabilité que l'enregistrement soit incorrect pour mesurer l'effet prévu (ce qui n'est pas nécessaire pour la vérification des macrodonnées). Consulter la section 4.2 pour un exemple et une brève discussion.

Dans le cas de la vérification des microdonnées, l'estimation cible prévue est, techniquement, égale à la somme des valeurs prévues pondérées. L'estimation cible prévue corrigée est un recalcul de l'estimation cible prévue en remplaçant la valeur prévue par la valeur déclarée (toutes les autres valeurs prévues demeurant inchangées). Elle est calculée pour une valeur à la fois, afin de créer une estimation cible prévue corrigée (et une mesure de l'effet de la vérification) pour chaque valeur déclarée. Le processus de création du score est le même pour la vérification des macrodonnées, excepté que les valeurs d'item sont remplacées par les estimations (provenant d'un ensemble d'estimations étudié). Pour les estimations d'un total, l'estimation cible prévue est la somme des estimations étudiées prévues, et la somme des estimations comprises dans l'ensemble étudié concorde avec les estimations au niveau cible. Consulter Farwell (2009) pour des renseignements plus détaillés.

Dans le cadre de la vérification selon l'importance, la valeur de normalisation mentionnée dans (1), qui peut être une estimation cible prévue ou un multiple de l'erreur type prévue pour l'estimation cible (options dans le SigEE), permet de comparer et de combiner les scores de différentes variables.

Bien que le score générique s'applique aux scores de macrodonnées, le SigEE ne comprend aucune fonctionnalité pour la vérification des macrodonnées selon l'importance. La suite de l'exposé traite par conséquent de l'application de la méthode aux microdonnées.

4.2 Score d'importance générique

En se servant de (1) et (2), et en prenant comme estimation cible l'estimation d'Horvitz-Thompson d'un total, le score d'item pour l'item *j* du fournisseur *i* est donné par :

$$s_{ij} = 100q_j \left| \frac{Y^*_{adj,ij} - Y^*_j}{Y^*_j} \right| = 100q_j \left| \frac{w^*_i (y_{ij} - y^*_{ij})}{Y^*_j} \right|. \qquad (3)$$

où w^*_i est le poids d'estimation prévu pour le fournisseur i, y_{ij} est la valeur d'item déclarée pour l'item j du fournisseur i, y^*_{ij} est la valeur d'item prévue de l'item j du fournisseur i, Y^*_j est l'estimation cible prévue pour l'item j, q_j est la probabilité de déclarer incorrectement l'item j (c'est-à-dire la probabilité que y_{ij} soit incorrecte), et $Y^*_{adj,ij} = Y^*_j - w^*_i y^*_{ij} + w^*_i y_{ij}$.

Plus techniquement, l'effet de la microvérification d'un élément de donnée peut être défini comme étant la réduction absolue du biais induit dans l'estimation cible par la déclaration incorrecte (Farwell, Poole et Carlton, 2002). L'effet doit être prédit puisqu'il ne peut être connu qu'après la vérification. Pour (3), cela se résume à estimer l'espérance de $\left| w^*_i (y_{ij} - y^*_{ij}) \right|$ que nous avons approximée pour le moment par $q_j \left| w^*_i (y_{ij} - y^*_{ij}) \right|$. Cette approximation peut être justifiée si q_j et $\left| w^*_i (y_{ij} - y^*_{ij}) \right|$ sont indépendants. Malheureusement, cela n'est pas strictement vrai et les recherches dans ce domaine doivent se poursuivre. À l'ABS, il est fréquent que l'on en sache peu au sujet de q_j et la tendance consiste à fixer $q_j = 1$.

Pour calculer ce score, nous avons donc besoin d'une valeur d'item déclarée, d'une valeur d'item prévue, d'une estimation cible prévue, d'un poids d'estimation prévu et d'une valeur de q_j. Si l'on dispose de ces termes avant de recevoir les réponses, on peut calculer un score aussitôt que l'on obtient une réponse. Ensuite, on peut comparer le score à une valeur seuil prédéterminée et décider immédiatement s'il convient de modifier la valeur.

Alors que l'expression $w^*_i (y_{ij} - y^*_{ij})$ peut être utilisée pour mesurer l'effet de la vérification dans le cas de l'estimation d'un total, la définition générique (1) est également utile pour calculer des scores plus complexes, dont ceux qui s'appliquent à des estimations de taux et à des erreurs types, et pour la macrovérification. Par exemple, nous pouvons calculer un score d'item de microvérification pour le ratio de deux estimations d'Horvitz-Thompson d'un total ($R_{jk} = Y_j / Y_{k \neq j}$, pour les items *j* et *k*) comme il suit. Nous utilisons le ratio des estimations cibles prévues du numérateur et du dénominateur (Y^*_j / Y^*_k) comme estimation cible prévue (R^*_{jk}). L'estimation cible prévue corrigée devient le ratio des estimations cibles prévues corrigées du numérateur et du dénominateur, ce qui donne $R^*_{adj,ijk} = Y^*_{adj,ij} / Y^*_{adj,ik}$, où $Y^*_{adj,ij} = Y^*_j - w^*_i y^*_{ij} + w^*_i y_{ij}$ et $Y^*_{adj,ik} = Y^*_k - w^*_i y^*_{ik} + w^*_i y_{ik}$.

5. Scénarios de vérification de l'ABS couvert par le SigEE

L'ABS a mis en œuvre la vérification selon l'importance au début des années 1990 pour l'Australian Survey of Average Weekly Earnings (AWE) (Lawrence et McDavitt, 1994). La vérification s'appuyait sur des scores d'item établis conformément à (1) pour des estimations de taux. L'AWE était une enquête pour laquelle le nombre de scores d'item clé était faible et qui offrait une quantité suffisante de données prévérifiées et postvérifiées, de données historiques et d'estimations stables. Les valeurs prévérifiées et postvérifiées ont permis de procéder à une analyse en

vue de déterminer les seuils effectifs. Les valeurs et les estimations prévues ont été obtenues facilement d'après les valeurs et les estimations historiques disponibles, et les estimations cibles ont été faciles à prédire.

Après la première application à l'AWE, diverses applications du cadre de vérification selon l'importance ont été mises en œuvre pour d'autres programmes de collecte de données, particulièrement ceux destinés à recueillir des données agricoles. Comme l'ABS s'était donné pour objectif d'établir des processus plus génériques et une infrastructure commune de traitement des données économiques, nous avons décidé de tester la versatilité du cadre en l'appliquant à plusieurs enquêtes, chacune posant une série différente de difficultés de vérification. Les essais ont été réalisés en 2002 et en 2003 en se servant d'une suite de programmes SAS qui ont « évolués » pour finalement devenir le SigEE. Les essais comprenaient des collectes pour lesquelles il n'existait pas de données prévérifiées et postvérifiées, des collectes nouvelles ou ponctuelles pour lesquelles il n'existait aucune données historiques, des collectes comportant des valeurs de réponse très erratiques, des collectes pour lesquelles les estimations étaient très difficiles à prédire, et des collectes comprenant de nombreuses données de sortie clés et de nombreux éléments de données clés. Consulter Farwell (2004) et Farwell (2005) pour des renseignements plus détaillés. Ces essais ont abouti à l'élaboration de seuils interactifs, de courbes coûts-avantages, de scores d'item de rechange et d'options pour les scores de fournisseur.

6. Fonctionnalités du SigEE

Le SigEE possède des fonctionnalités qui permettent de procéder à la vérification selon l'importance avec ou sans valeurs prévues, avec ou sans estimations prévues, et avec ou sans seuils prédéterminés. L'une des caractéristiques des fonctionnalités du SigEE est la possibilité d'utiliser quatre types de scores d'item. Dans le SigEE, les quatre approches axées sur les scores d'item sont appelées *chemins*, les scores d'item sont appelés *scores d'item du chemin A, du chemin B, du chemin C* et *du chemin D*.

6.1. Choix des scores d'item : Chemins A, B, C et D

Pour créer les scores d'item du *chemin A*, on se sert du score générique (1). Le calcul de ces scores nécessite des valeurs prévues, des estimations prévues et des poids prévus (le SigEE utilise $q_j = 1$ comme valeur par défaut de la probabilité de commettre une erreur de déclaration, ce qui équivaut à supposer que toutes les valeurs déclarées ont la même probabilité d'être erronées).

Les scores du *chemin B* ont été élaborés pour traiter la situation où les estimations prévues ne sont pas disponibles et que l'on ne possède que des valeurs prévues et des poids prévus. Il s'agit essentiellement de scores du chemin A normalisés, obtenus en divisant chaque score du chemin A par la somme des scores du chemin A et en exprimant le résultat en pourcentage. En procédant ainsi, les estimations prévues ne sont pas nécessaires. Un score du chemin B pour l'item j du fournisseur i prend la forme $s_{ij}^{*} = 100 * \frac{s_{ij}}{\Sigma_i s_{ij}}$, où la somme est calculée sur l'ensemble des enregistrements fournisseurs qui requièrent une vérification pour l'item j. Par exemple, le score du chemin B d'une estimation du total est $s_{ij}^{*} = 100 \frac{\left| w_i^{*} (y_{ij} - y_{ij}^{*}) \right|}{\Sigma_i \left| w_i^{*} (y_{ij} - y_{ij}^{*}) \right|}$.

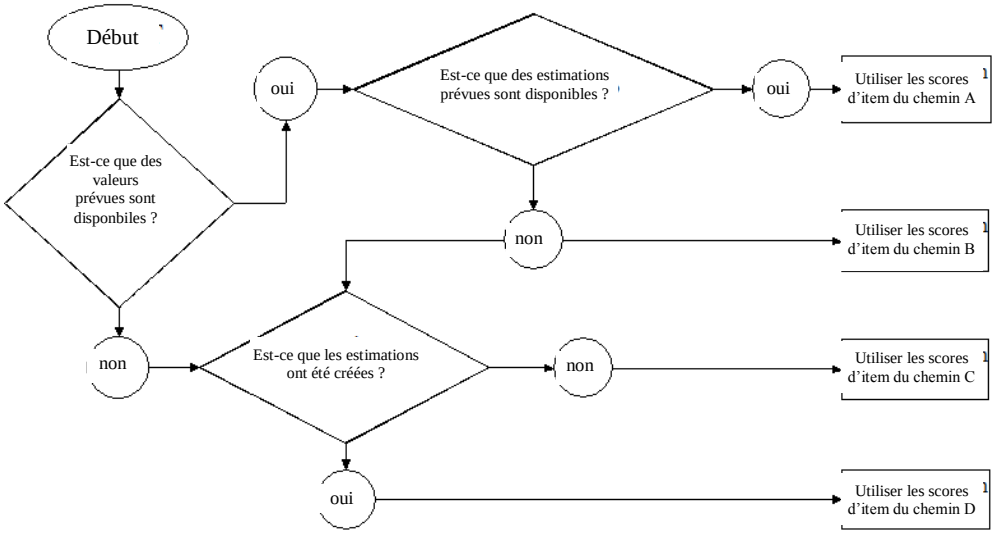
Les scores du chemin B peuvent être affectés par des valeurs extrêmes de s_{ij}, et une fonction de correction des scores extrêmes est intégrée dans le SigEE pour gérer cette situation. Les éléments de données à l'origine des scores extrêmes sont supprimés du processus de calcul des scores et placés dans le flux d'items critiques en vue de leur vérification, puis les nouveaux scores du chemin B sont calculés.

Les scores d'item des *chemins C et D* sont fondés sur les contributions normalisées combinées au niveau, aux fluctuations et à l'erreur type d'une estimation cible. Les scores *initiaux* pour le niveau, les fluctuations et l'erreur type sont combinés en utilisant une mesure euclidienne pondérée pour créer un score d'item. Ces scores d'item ont été créés pour contourner la situation où il n'existe aucune valeur prévue et aucune estimation prévue. Les scores

d'item du chemin C estiment les contributions au niveau, aux fluctuations et à l'erreur type des estimations (puisque les réponses de bonne qualité disponibles ne seront pas suffisantes pour appuyer la création des estimations). Le calcul des scores du chemin D requiert l'information auxiliaire provenant de la création des estimations pour calculer les contributions. Par conséquent, ces scores ne peuvent être créés qu'après avoir produit les estimations.

Le diagramme 6.1-1 qui suit illustre comment est choisi le chemin du SigEE.

Diagramme 6.1-1
Choix du chemin approprié



6.2 Scores de contributeur

Les scores initiaux utilisés pour produire les scores du chemin D peuvent servir à classer les contributeurs aux estimations, aux fluctuations des estimations et aux erreurs types des estimations; les estimations de la contribution d'une unité sont nettement plus exactes que celles employées pour créer les scores du chemin C. Cette fonctionnalité peut servir à produire des listes ordonnées des contributeurs unitaires qui sont utilisées fréquemment durant la macrovérification. En ce sens, cette dernière comporte souvent une composante de microvérification. Afin de séparer les tâches de macrovérification des tâches de microvérification, nous donnons aux scores initiaux du chemin D le nom de *scores de contributeur* lorsqu'ils sont utilisés durant la macrovérification.

6.3 Production des valeurs prévues

Le SigEE a la capacité de produire les valeurs prévues. Le système peut contenir des valeurs historiques utilisables comme valeurs prévues, si elles sont disponibles. Facultativement, des facteurs de correction peuvent être appliqués aux valeurs historiques. Le SigEE peut également produire des valeurs prévues en utilisant des moyennes pondérées ou effectuer une imputation par le ratio en utilisant le fichier de données d'une enquête antérieure ou courante. Le SigEE permet aussi à l'utilisateur de spécifier les valeurs prévues dans un fichier.

6.4 Choix des scores de fournisseur

Pour les scores de fournisseur s'appuyant sur des scores d'item pondérés, le SigEE offre les trois options qui suivent :

i) le maximum des scores d'item pondérés;
ii) la norme euclidienne des scores d'item pondérés;
iii) la moyenne quadratique (RMS pour *root mean square*) des scores d'item pondérés.

Par exemple, le score RMS pondéré de fournisseur est défini comme étant $s_i = \sqrt{\frac{\Sigma a_j s_{ij}^2}{n}}$, où s_i est le score du fournisseur i; s_{ij} est le score de l'item *j* du fournisseur i; a_j est un poids d'item défini par l'utilisateur pour l'item *j*, et *n* est le nombre d'items clés non manquants.

Le score euclidien pondéré est égal au score RMS en prenant $n = 1$. Le poids d'item (a_j) permet aux utilisateurs de rendre un item plus important qu'un autre lors de la création d'un score de fournisseur. Le poids d'item peut être utilisé pour tenir compte de la probabilité de commettre une erreur de déclaration quand q_j est fixé à 1 dans le SigEE. Les scores euclidiens et RMS (option par défaut du SigEE) donnent de bons résultats quand le nombre d'éléments clés est grand ou que des seuils interactifs sont nécessaires. Le score produit des courbes coûts-avantages significatives pour la sélection des seuils interactifs.

6.5 Détection multivariée des valeurs aberrantes inattendues

Les scores de contributeur peuvent également être utilisés pour détecter les valeurs aberrantes inattendues. Les scores de contributeur pour le niveau peuvent être combinés en utilisant la norme RMS pour les items clés. Le score combiné peut ensuite être employé pour sélectionner des valeurs aberrantes inattendues multivariées. Il s'agit de valeurs aberrantes qui ont un effet prononcé sur l'ensemble d'items clés.

6.6 Options concernant les seuils

Le SigEE offre l'option d'utiliser des seuils prédéterminés (créés avant d'obtenir les réponses) ou de choisir des seuils de manière interactive. Il est préférable d'utiliser, dans la mesure du possible, des seuils prédéterminés. Cependant, ces derniers deviennent parfois inefficaces au cours du temps (donnant lieu à la sélection d'un plus grand nombre d'éléments de données ou de fournisseurs pour la vérification) et doivent donc être mis à jour. Les seuils peuvent être déterminés de manière interactive en se référant à des listes et à des courbes coûts-avantages.

Le SigEE fournit quatre méthodes de détermination des seuils applicables à la vérification fondée sur les scores d'item ou sur les scores de fournisseur. Ces méthodes sont celles du seuil de *score*, du seuil de *score de rang égal ou supérieur à k (top-k)*, du seuil de *score cumulatif (%)* et du seuil *de score cumulatif itératif (%)*. Le seuil de score est la valeur des scores d'item ou de fournisseur au-delà de laquelle l'item ou l'enregistrement fournisseur est sélectionné pour la vérification. L'application du score *top-k* donne lieu à la sélection pour la vérification des enregistrements dont le rang est égal ou supérieur à k. Pour appliquer le seuil de score cumulatif (%), les scores normalisés sont cumulés par ordre de rang (en partant du rang 1). Tous les enregistrements nécessaires pour atteindre le score cumulatif total en pourcentage sont sélectionnés pour la vérification. Le seuil de score cumulatif itératif (%) est une application plus compliquée du seuil de score cumulatif (%). La méthode consiste à sélectionner itérativement les enregistrements sur plusieurs exécutions de la vérification de façon à atteindre le seuil prédéterminé de score cumulatif en pourcentage. Le SigEE fait le suivi des totaux des avantages et ajuste le seuil cible pour chaque exécution de la vérification. Le SigEE est également doté d'une fonction de détection des scores extrêmes destinée à être utilisée avec des seuils fondés sur des scores normalisés (c'est-à-dire des seuils faisant intervenir les pourcentages de score cumulatif). Les enregistrements dont les scores sont extrêmes sont supprimés, puis les scores normalisés sont recalculés. Le seuil est appliqué aux scores révisés. Les enregistrements dont les scores sont extrêmes sont placés dans le flux d'enregistrements critiques pour la vérification.

6.7 Vérification fondée sur les seuils établis pour les fournisseurs et pour les items

Pour la plupart des enquêtes économiques, on sélectionne des enregistrements fournisseurs complets pour la vérification, parce que les données ont tendance à avoir une structure de bilan. Une erreur dans un élément de donnée peut être reliée à des erreurs éventuelles ailleurs dans l'enregistrement. Par conséquent, les seuils utilisés principalement sont les seuils de fournisseur, car l'objectif est de déceler les enregistrements fournisseurs présentant des anomalies. Les vérificateurs peuvent utiliser des scores et des rangs d'item, et d'autres renseignements pour résoudre et traiter les erreurs dans l'enregistrement complet.

Les données recueillies dans le cadre de certaines enquêtes économiques n'ont pas cette structure de bilan (par exemple les données d'enquête sur les activités ou sur les biens et services). Pour ces enquêtes, les erreurs dans une partie de l'enregistrement fournisseur ont tendance à ne pas être reliées à des erreurs ailleurs dans l'enregistrement, de sorte que les éléments de données individuels sont sélectionnés pour la vérification plutôt que les enregistrements fournisseurs. Pour ces types d'enregistrements fournisseurs, les vérificateurs ne devraient pas consacrer de temps à des recherches inutiles dans l'enregistrement complet pour corriger les éléments de données particuliers. Par conséquent, les principaux seuils utilisés sont les seuils d'item, car le but est de déceler les valeurs d'item anormales. Il demeure nécessaire de classer les enregistrements fournisseurs par ordre de priorité afin de gérer la charge de travail, mais l'accent est mis surtout sur des éléments de données particuliers dans les enregistrements fournisseurs.

Pour chaque collecte de données, les principaux facteurs qui déterminent l'ensemble de fonctionnalités utilisées dans le SigEE sont le type de seuil principal (c'est-à-dire seuil d'item ou de fournisseur), la méthode de création des seuils (c'est-à-dire seuils prédéterminés ou interactifs), ainsi que le type de score d'item utilisé (c'est-à-dire score d'item de chemin A, B, C ou D).

7. Dix ans plus tard ...

7.1 L'évolution de l'environnement

La vérification selon l'importance et le SigEE ont été lancés il y a dix ans. Cependant, comme tous les organismes statistiques, l'ABS a connu d'importants changements au cours de cette période. Les clients demandent de plus en plus de données « meilleures, produites plus rapidement, moins onéreuses » et nous nous concentrons maintenant davantage sur l'emploi de données administratives pour livrer une solution statistique. Le rythme du progrès technologique s'est également accéléré, et de nombreux nouveaux outils et systèmes ont été installés pour faire face à l'accroissement de la demande auquel est soumise notre infrastructure statistique. En raison de compressions budgétaires, moins d'employés sont disponibles pour effectuer le travail, problème qu'aggrave le roulement plus important du personnel à mesure que la population active devient plus mobile. Par conséquent, nous avons été obligés de mieux gérer les connaissances et d'accroître plus rapidement les capacités du personnel.

Nos efforts concernant la vérification selon l'importance et le SigEE ont été couronnés d'un certain succès. La méthodologie est maintenant utilisée ordinairement par un certain nombre de programmes de collecte des données de l'ABS, ce qui a permis de réaliser des économies importantes. Ces succès sont certes gratifiants, mais l'ABS n'a pas réussi à ce que la vérification selon l'importance et (ou) le SigEE soient appliqués universellement à tous les programmes de collecte de données économiques. Nous nous sommes donc demandé pourquoi il en était ainsi et ce que nous devrions faire pour arriver à une adoption plus générale de la méthodologie dans l'avenir.

7.2 Principaux enseignements tirés de l'expérience jusqu'à présent

Le premier enseignement qui se dégage est peut-être d'arriver à ce que les secteurs opérationnels reconnaissent que la mise en œuvre de la vérification sélective représente un processus de changement important, qui comprend l'élaboration et l'adoption de nouvelles méthodes et de nouveaux procédés et systèmes, ainsi que de nouveaux rôles et compétences pour les employés. Trop souvent, le changement est perçu comme l'apport de modifications relativement mineures aux opérations de collecte, et la complexité du processus est sous-estimée. Les secteurs qui prévoient des ressources appropriées pour effectuer le changement aboutissent invariablement à de meilleurs résultats que ceux qui ne le font pas, car les changements complexes auxquels il faut réfléchir sont nombreux. Ils ont aussi tendance à investir dans l'amélioration des procédures et de la documentation, ce qui les aide à transférer les connaissances d'une génération à la suivante.

Les cadres supérieurs doivent adhérer au processus de changement et offrir leur appui continu. Les secteurs dont les cadres supérieurs cessent d'appuyer le processus de changement en cours de route ou délèguent leurs responsabilités à un cadre de niveau inférieur lorsque d'autres priorités requièrent leur attention adoptent les nouvelles méthodes avec nettement moins de succès. Les cadres supérieurs devraient aussi être tenus responsables de la mise en œuvre du changement.

L'une des meilleures recettes de succès consiste à établir une cible élevée d'économie ou, éventuellement, une réduction importante du délai de diffusion (ou les deux). Ces mesures facilitent vraiment la « concentration ». Donner aux cadres supérieurs la responsabilité d'atteindre la cible d'économie établie les incite également à participer au processus.

Les secteurs opérationnels doivent reconnaître que la modification des processus de vérification représente un engagement à long terme et devraient la traiter comme telle. Il convient d'obtenir l'appui des principaux intervenants non seulement au départ, mais de manière continue. Par exemple, un soutien méthodologique et un soutien des systèmes devraient être disponibles durant la phase de développement, la phase de mise en œuvre et en permanence par après. S'il se produit un changement rapide de personnel, les compétences dans un domaine particulier peuvent être sérieusement appauvries, ce qui limitera fortement les connaissances nécessaires pour appuyer les nouveaux processus. Un bon soutien de la méthodologie et des systèmes aidera ces secteurs opérationnels à se remettre sur les rails. Une bonne documentation et du bon matériel de formation sont également utiles et, idéalement, ces documents devraient être disponibles en ligne.

Dans la même veine, il faudrait s'efforcer de surveiller, d'évaluer et d'améliorer les processus de vérification sélective au cours du temps, afin d'être certain qu'ils demeurent optimaux pour les programmes de collecte de données concernés.

Aux secteurs qui promeuvent ou soutiennent le changement au sein d'une organisation, nous suggérons de se concentrer sur les secteurs opérationnels dont les cadres supérieurs et les employés veulent un changement, du moins de s'adresser à ces secteurs pour commencer. Le succès sera bien plus important en travaillant avec ces secteurs qu'avec ceux ne manifestant que peu d'intérêt pour le changement ou ceux dont les employés sont distraits par d'autres priorités, même si les économies qui peuvent être réalisées semblent importantes. Il est également utile de faire connaître les succès à mesure qu'ils se concrétisent. Vous susciterez ainsi un intérêt et un soutien pour le changement, car il n'est pas facile d'écarter la perspective de vraies économies ou d'autres améliorations.

Enfin, le SigEE a été construit pour traiter des situations de vérification très diverses. Étant donné l'évolution de notre environnement de travail au cours des dix dernières années, certaines personnes estiment aujourd'hui que cela prend trop de temps d'installer l'outil et d'apprendre à s'en servir, et qu'il est trop difficile à utiliser en pratique. Durant le développement de nouveaux outils, nous devrons donc veiller à offrir les fonctionnalités qui permettent de réaliser les gains les plus importants et peut-être omettre certaines fonctionnalités « non essentielles ». L'investissement en vue de rendre les outils conviviaux est également un bon moyen d'obtenir l'adhésion des secteurs opérationnels (par exemple bonnes interfaces, bien intégrées dans l'environnement de traitement, *etc*.).

Bibliographie

Farwell, K. (2004), « The General Application of Significance Editing to Economic Collections », Methodology Advisory Committee Papers, n° 1352.0.55.066 au catalogue de l'Australian Bureau of Statistics, Canberra.

Farwell, K. (2005), « Significance Editing for a Variety of Survey Situations », document présenté à la 55e séance de l'International Statistical Institute, Sydney, 5 au 12 avril.

Farwell, K. (2009), « The Use of Scores to Detect and Prioritise Anomalous Estimates ». Methodology Advisory Committee Papers, n° 1352.0.55.104 au catalogue de l'Australian Bureau of Statistics, Canberra.

Farwell, K., Poole, R. et S. Carlton (2002), « A Technical Framework for Input Significance Editing », Article présenté au DataClean2002, Jyväskylä, Finlande.

Farwell, K. et M. Raine (2000), « Some Current Approaches to Editing in the ABS », *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, États-Unis.

Lawrence, D. et C. McDavitt (1994), « Significance editing in the Australian Survey of Average Weekly Earnings », *Journal of Official Statistics*, vol. 10, n° 4, p. 437 à 447.

SÉANCE 4B

CONFIDENTIALITÉ

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

G-Confid : le logiciel de confidentialité de Statistique Canada

Caroline Rondeau et Jean-Marc Fillion[1]

Résumé

Statistique Canada doit assurer la protection de données confidentielles provenant des répondants en vertu de la *Loi sur la statistique*. La suppression de cellules est une technique utilisée pour assurer la protection des données tabulaires. Le logiciel de confidentialité automatisé G-CONFID (antérieurement CONFID2) développé à Statistique Canada est utilisé pour effectuer cette technique. Ce logiciel est convivial, utilise la même structure que les autres systèmes généralisés de Statistique Canada et permet d'intégrer de nouvelles approches. Il permet aussi de traiter des tableaux à plusieurs dimensions et dont la taille peut être volumineuse. De plus, G-CONFID fait partie du projet d'unification des méthodes et des systèmes de Statistique Canada. L'objectif principal de G-CONFID est de fournir le niveau de protection adéquat aux cellules confidentielles tout en minimisant la perte d'information. La programmation linéaire pour faire la suppression résiduelle de façon optimale est utilisée pour atteindre cet objectif. Cette présentation couvre la fonctionnalité ainsi que les caractéristiques de G-CONFID. L'accent sera mis sur l'utilisation des variables de coût spécifiées par l'utilisateur.

Mots clés : sensibilité ; confidentialité ; suppression résiduelle ; variable de coût.

1. Introduction

1.1 Introduction à G-Confid

Les données de type économique sont fréquemment présentées sous forme de tableaux à différents niveaux de détails. Cette représentation peut engendrer un problème de divulgation dans une des situations suivantes : i) il y a très peu de répondants dans une cellule ou ii) il y a seulement un ou deux répondants qui contribuent le plus dans une cellule, ce qu'on appelle la situation de dominance. La suppression de cellules est une technique utilisée pour assurer la protection des données tabulaires. Le logiciel de confidentialité automatisé G-Confid est utilisé pour effectuer cette technique.

G-Confid est un système généralisé qui effectue la suppression de cellules pour des données de type économique. En plus d'être un outil normalisé et d'être utilisé comme modèle pour d'autres systèmes généralisés à Statistique Canada, ce système est très flexible et convivial. Nous avons la possibilité de modifier l'information puisque les fichiers d'entrée/sortie sont en SAS en plus d'avoir la possibilité de définir les paramètres et options propres à nos besoins. Ce système est utilisé par plusieurs types d'enquêtes à Statistique Canada.

1.2 Structure de l'article

Dans la section 2, nous expliquons les différentes composantes de G-Confid. Dans la section 3, nous approfondissons l'utilisation des variables de coût spécifiées par l'utilisateur à l'aide d'un exemple. En guise de conclusion, les avantages de G-Confid sont présentés à la section 4.

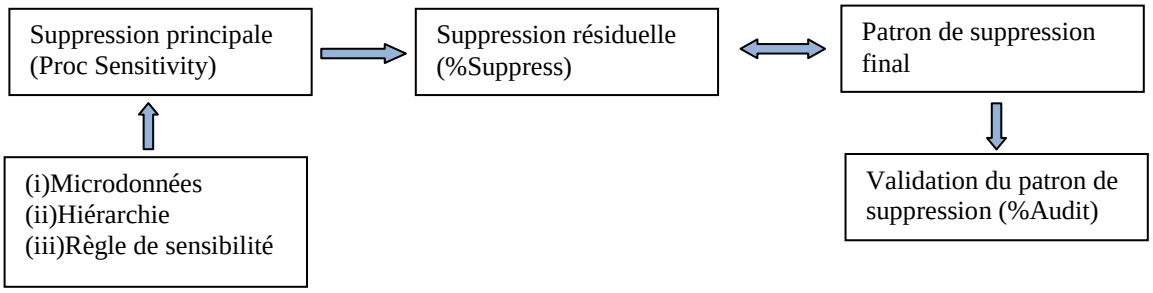
[1]Caroline Rondeau, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6, caroline.rondeau@statcan.gc.ca ; Jean-Marc Fillion, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6, jean-marc.fillion@statcan.gc.ca.

2. Composantes de G-Confid

2.1 Description générale de G-Confid

G-Confid est une suite de trois composantes SAS. La première composante (Proc Sensitivity) effectue la suppression principale. La deuxième composante (macro %Suppress) effectue la suppression résiduelle, c'est-à-dire qu'elle identifie les cellules supplémentaires à supprimer de façon optimale en utilisant la programmation linéaire. La troisième composante (macro %Audit) vérifie que la suppression résiduelle protège adéquatement les cellules confidentielles. La figure 2.1-1 représente le fonctionnement de G-Confid dont chaque composante est expliquée plus en détail dans les sections 2.1.1 à 2.1.3.

Figure 2.1-1
Fonctionnement de G-Confid



2.1.1 Suppression principale

Lors de la suppression principale (la composante Proc Sensitivity), la valeur totale de chaque cellule ainsi que sa sensibilité à partir des microdonnées sont calculées. La sensibilité est une valeur qui détermine si la cellule est considérée comme confidentielle ou non. Trois éléments sont essentiels pour effectuer la suppression principale : i) le fichier de microdonnées, ii) la hiérarchie et iii) la règle de sensibilité. Le fichier de microdonnées doit contenir, entre autres, les informations suivantes : un numéro d'identité (habituellement, le numéro d'entreprise), les variables de dimensions du tableau (par exemple : l'industrie et la région) ainsi que la variable d'intérêt (par exemple : le revenu). La hiérarchie permet de spécifier à G-Confid la structure des tableaux de données à protéger et elle est reliée aux variables de dimensions. La hiérarchie peut être définie à plusieurs niveaux, par exemple, on peut vouloir traiter différentes régions du Canada telles que l'Est, l'Ouest et les Maritimes; chaque région est divisée entre leurs provinces respectives. La règle de sensibilité permet d'identifier les cellules confidentielles. Plusieurs règles existent et elles sont des formes particulières de la règle de sensibilité linéaire générale décrite à la formule (1).

$$S = \sum_{i=1}^{r} \alpha_i x_i \text{ , où } \alpha_1 \geq \alpha_2 \geq \ldots \geq -1 \qquad (1)$$

où
S représente la sensibilité de la cellule,
α_i sont des coefficients fixes,
r représente le nombre de contributeurs à la cellule,
x_i représente les valeurs des différents contributeurs à la cellule en ordre décroissant ($x_i \geq 0$).

Si S est positif, alors la cellule est confidentielle. Il y a plusieurs règles de sensibilité définies dans G-Confid : la règle nk, la règle pq ainsi que d'autres règles spécifiques à Statistique Canada.

Voici un exemple d'appel à la composante Proc Sensitivity.

```
PROC SENSITIVITY
        DATA=data        OUTCONSTRAINT=outconstraint        OUTCELL=outcell
        HIERARCHY= "Total_naics N912 N913 N92 N931 N932 N933 N934 N941 N942;
                    Canada Est Québec Ontario Ouest:
                     Est Terre-Neuve-et-Labrador Maritimes: Ouest Prairies Colombie-Britannique; "
        SRULE="pq 0.15"; ID EntrepriseID;
        VAR revenu; DIMENSION Industrie Region;
```

Comme nous pouvons le constater, la formulation générale est semblable à toutes les procédures SAS. Le fichier SAS de microdonnées est défini dans le paramètre DATA, les fichiers SAS de sorties dans les paramètres OUTCONSTRAINT ET OUTCELL. Le fichier OUTCONSTRAINT identifie les relations entre les cellules à l'aide de coefficients représentant des équations linéaires. Le fichier OUTCELL contient, entre autres, les variables de dimensions, la valeur totale de chaque cellule, sa sensibilité ainsi que son statut. La hiérarchie, la règle de sensibilité (pq 0.15), la variable d'identité (EnterpriseID), la variable d'intérêt (revenu), ainsi que les dimensions (Industrie et Region) doivent être spécifiées dans la procédure.

2.1.2 Suppression résiduelle

L'objectif principal de la suppression résiduelle (la composante %Suppress) est de fournir le niveau de protection adéquat aux cellules confidentielles tout en minimisant la perte d'information. Pour chaque cellule confidentielle, des cellules complémentaires sont identifiées en résolvant le problème de programmation linéaire en minimisant les coûts de suppression sous la contrainte que chaque cellule est bien protégée. Les cellules confidentielles combinées avec les cellules complémentaires forment le patron de suppression final.

Il y a deux étapes (appelées phases) pour la suppression résiduelle. La suppression résiduelle est faite séquentiellement. La deuxième phase permet de faire un petit ménage, c'est-à-dire qu'elle permet de libérer des cellules sur la suppression effectuée à la première phase. Chaque phase utilise une fonction de coût. Les coûts disponibles sont : « constant », « size », « digit » ou « information ».

Voici un exemple d'appel à la composante %Suppress.

%SUPPRESS(INCELL=outcell, CONSTRAINT=outconstraint, CFUNCTION1=size,
CFUNCTION2=information, CVAR1 = , CVAR2 = , OUTCELL=outcell_sprs) ;

Les fichiers de sortie (OUTCELL et OUTCONSTRAINT) de la composante Proc Sensitivity sont utilisés comme fichier d'entrée pour cette macro (INCELL et CONSTRAINT). Les fonctions de coûts pour chaque phase sont définies dans CFUNCTION1 et CFUNCTION2 et le fichier de sortie est dans OUTCELL. Par défaut, la variable d'intérêt (c'est-à-dire le total de la cellule) est utilisée comme variable de coût pour chaque phase (CVAR1 et CVAR2). Cependant, il est possible d'en définir de nouvelles à l'aide des paramètres CVAR1 et/ou CVAR2. Nous y reviendrons à la section 3.

Le tableau 2.1.2-1 présente un exemple sous forme tabulaire de la sortie de la composante %Suppress où les cellules confidentielles sont représentées en rouge et les cellules résiduelles en bleu. Veuillez prendre note que ce ne sont pas les données de sortie de G-Confid, mais une représentation sous forme de tableaux.

Tableau 2.1.2-1
Total du revenu par industrie et région

	Canada	Est	Terre-Neuve-et-Labrador	Maritimes	Québec	Ontario	Ouest	Prairies	Colombie-Britannique
N912	2016	74	0	74	677	342	923	838	85
N913	22115	3	0	3	20197	382	1533	692	841
N92	3875	355	3	352	245	549	2726	2071	655
N931	3014	88	0	88	1164	637	1125	791	334
N932	3435	35	0	35	2750	548	102	98	4
N933	3947	209	0	209	1393	1266	1079	787	292
N934	231	59	0	59	86	32	54	50	4
N941	7019	113	0	113	784	1221	4901	4695	206
N942	66	0	0	0	0	13	53	52	1
Total	45718	936	3	933	27296	4990	12496	10074	2422

Si seulement la suppression principale était effectuée (cellules en rouge), on n'aurait pas un patron de suppression final, car on pourrait toujours dériver certaines des valeurs supprimées. Par exemple, dans la colonne représentant le Canada, seulement une industrie est confidentielle (N932), donc il est assez facile de retrouver sa valeur. La suppression résiduelle permet donc d'obtenir un patron de suppression final.

2.1.3 Validation de la suppression

L'objectif principal de la validation de la suppression (la composante %Audit) est de vérifier la qualité du patron de suppression, qui a été i) créé dans G-Confid et modifié par l'utilisateur ou ii) créé à l'extérieur de G-Confid et fourni par l'utilisateur.

3. Variable de coût

3.1 Avantage

Comme mentionné précédemment, la composante %Suppress permet à l'utilisateur de définir ses propres coûts de suppression. L'utilisation de variables de coût permet, entre autres, i) de préserver la cohérence entre les patrons de suppression pour la même enquête (par exemple : ancienne enquête comparativement à la nouvelle enquête), ii) de supprimer en premier les cellules ayant un coefficient de variation (c.v.) élevé et iii) de favoriser la publication de certains domaines importants pour les utilisateurs.

Pour spécifier à G-Confid comment identifier les cellules à supprimer en priorité, il suffit de le préciser dans les variables de coût (CVAR1 et/ou CVAR2) dans l'appel de la composante %Suppress. La variable spécifiée sera utilisée au lieu de la variable d'intérêt (total de la cellule) dans le calcul du coût de suppression. Plus le coût est faible pour une cellule plus la chance de supprimer cette cellule est élevée. Il y a aussi l'option de changer le statut de la cellule, c'est-à-dire d'imposer un patron de suppression. Mais nous allons nous concentrer sur l'utilisation des variables de coût illustrée à l'aide d'un exemple.

3.2 Exemple

Les données de l'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) seront utilisées pour démontrer l'utilisation des variables de coût. Cette enquête rencontre le problème suivant : les cellules des Territoires (petite population au nord du Canada) sont sélectionnées comme cellules complémentaires parce que le nombre total d'emplois est petit (variable de coût : total d'emploi), mais ces cellules ont une certaine importance dans leurs territoires. Donc, l'EERH aimerait publier plus de cellules dans les Territoires en tenant compte de leur apport en termes d'emplois (proportion d'emplois dans le territoire) au lieu du total d'emplois.

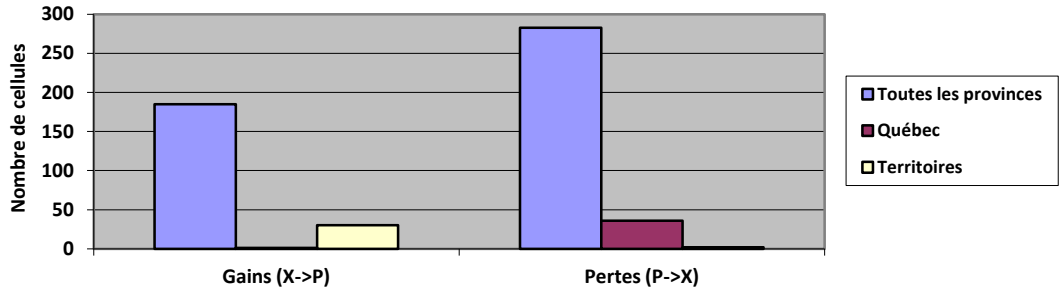
La variable de coût pour la première phase représente la proportion d'emplois pour une industrie donnée dans sa province et est définie selon la formule (2)

$$prop_{ij} = total_{ij} \Big/ \sum_{i \in j} total_i , \qquad \text{où i = industrie et j=province} \tag{2}$$

Cette variable de coût combinée avec la fonction de coût 'size' permet de supprimer les cellules possédant les plus petites proportions d'emplois. Pour la deuxième phase, la variable de coût est le total d'emplois combiné avec la fonction de coût 'size', ce qui aura pour effet de libérer les cellules avec le plus d'emplois.

Une simulation a été faite en utilisant 12 mois de données. Si une cellule est supprimée pour un mois donné, alors la cellule est considérée comme étant supprimée pour le patron annuel. En moyenne, nous avons 5 300 cellules dont 2 000 sont supprimées. Le but de la simulation est de comparer les gains et les pertes entre la méthode de base (la variable de coût est le total d'emplois pour les deux phases) et la nouvelle approche définie précédemment, c'est-à-dire en utilisant la proportion d'emplois. Les gains représentent le nombre de cellules ayant passé du statut de supprimée (X) avec la méthode de base à publier (P) avec la nouvelle approche. La figure 3.2-1 représente les gains (X-P) et les pertes (P-X) suite à la nouvelle approche pour les Territoires, le Québec ainsi que pour toutes les provinces combinées.

Figure 3.2-1
Nombre de cellules ayant subi des pertes/gains en ce qui a trait à la diffusion



Comme on peut le constater, il y a effectivement moins de suppressions avec la nouvelle approche pour les Territoires. Cependant, ce gain au niveau des Territoires entraîne une augmentation des pertes pour les autres provinces, car nous devons aller chercher ailleurs les cellules résiduelles. Ce qui entraîne un plus haut taux de suppression pour des grandes provinces comme le Québec. Veuillez noter que pour cette simulation, il y a environ 425 et 330 cellules pour le Québec et les Territoires respectivement.

Malgré le gain dans les Territoires, cette nouvelle approche n'a pas été mise en production à cause de l'augmentation de suppressions dans les grandes provinces (par exemple le Québec). Une meilleure fonction tenant compte de l'emploi et de la proportion de l'emploi doit être trouvée. Une solution proposée serait d'utiliser l'idée de la proportion, mais pas seulement au niveau de la province, mais aussi au niveau du Canada. On pourrait utiliser un certain pourcentage (p) de la proportion provinciale et 100-p pour la proportion au niveau du Canada, tel que défini avec la formule (3).

$$prop_{ij} = \frac{p}{100}\frac{total_{ij}}{\sum_{i\in j} total_i} + \frac{(100-p)}{100}\frac{total_{ij}}{\sum_{ij} total_{ij}}, \qquad \text{où i = industrie et j = province} \qquad (3)$$

Une simulation a été faite avec p = 25, 50, 75 et 100. Avec p = 100, nous nous ramenons à la formule (2). Les figures 3.2-2 et 3.2-3 représentent les pertes et gains respectivement entre l'approche de base et l'approche proposée selon différentes valeurs de p.

Figure 3.2-2
Nombre de cellules ayant subi des pertes

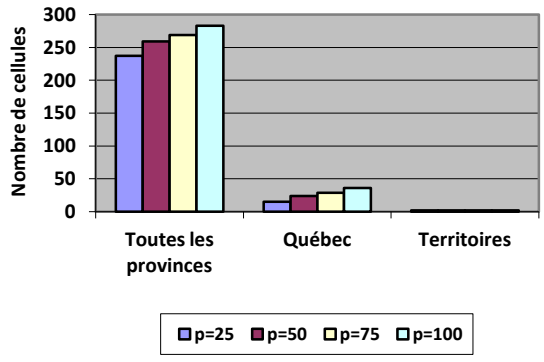
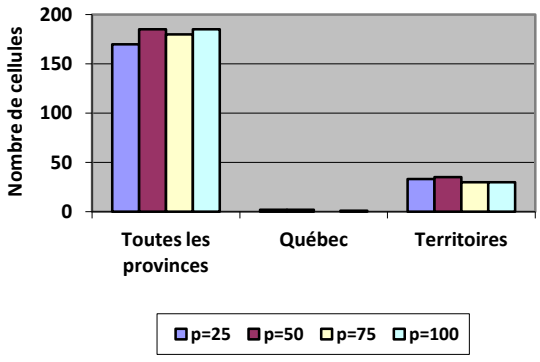


Figure 3.2 -3
Nombre de cellules ayant subi des gains



Au niveau des Territoires, il n'y a pas plus de pertes, peu importe la valeur de p. Pour le Québec, p = 25 ou 50 représentent le moins de pertes sans nuire aux gains. Au niveau du Canada, p = 25 représente le moins de pertes, mais un peu moins de gains que p=50, par exemple. La solution proposée semble être un bon compromis. Il y a moins de pertes pour les grandes provinces et les gains sont toujours satisfaisants pour les Territoires. Veuillez noter que les cellules supprimées pourraient, à la suite d'autres analyses, être négligeables. Ces analyses devraient inclure, entre autres, les pertes et gains en termes de pourcentage d'emplois.

D'autres simulations devront être faites. Comme nous pouvons le constater, G-Confid est très flexible dans l'utilisation des variables de coût.

4. Conclusion

En plus d'être un outil normalisé et très convivial, G-Confid a plusieurs avantages. Il peut être intégré avec d'autre programme SAS. Il est capable de résoudre des tableaux volumineux, il a une bonne performance pour la suppression résiduelle (solveur SAS OPTMODEL) et il permet de traiter les décompositions multiples (par exemple, nous pouvons diviser le Canada en provinces, en régions économiques et en régions métropolitaines de recensement). De plus, puisque G-Confid est développé à l'interne, nous avons la possibilité d'intégrer dans les prochaines années de nouvelles approches. Finalement, G-Confid peut être utilisé avec SAS Enterprise Guide, c'est-à-dire que les fonctions de G-Confid y apparaissent sous forme de tâches personnalisées. Ces tâches permettent de générer le code SAS nécessaire à G-Confid à partir d'une interface graphique.

Remerciements

Nous souhaitons remercier les responsables de l'EERH, en particulier, Yves Morin et Shou Xiang Chen, qui continuent de travailler à trouver une meilleure fonction tenant compte de l'emploi et de la proportion de l'emploi (un compromis entre les deux).

Bibliographie

Frolova, O., Fillion, J.-M. et J.-L. Tambay (2009), « Confid2 : nouveau logiciel de confidentialité des données tabulaires de Statistique Canada », Congrès annuel de la Société statistique du Canada, Recueil de la Section des méthodes d'enquête.

Statistique Canada (2011), « Guide de l'usager G-Confid », document interne.

Tambay, J.L. et J.M. Fillion (2011), « New business survey confidentiality software G-Confid », article présenté à la Réunion de travail CEE/Eurostat sur la confidentialité des données statistiques, à Tarragone (Espagne) du 26 au 28 octobre 2011.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Évaluation du risque de divulgation dans le cas de microdonnées perturbées

Natalie Shlomo[1]

Résumé

Nous rapprochons deux méthodes d'évaluation du risque de divulgation posé par les microdonnées d'enquête qui se fondent sur la probabilité de leur appariement correct à un fichier externe de données sur la population. Ces méthodes sont le cadre de couplage d'enregistrements probabiliste de Fellegi et Sunter (1969) et le cadre de modélisation probabiliste fondé sur le modèle log-linéaire de Poisson. Skinner (2008) a montré que ces deux méthodes sont essentiellement équivalentes. Nous fournissons une preuve empirique de ce résultat et montrons comment peut être évalué le risque de divulgation pour un ensemble de données d'entreprises fortement perturbé tiré de l'enquête de 1982 sur les fermes sucrières au Queensland, en Australie. Nous proposons d'estimer la probabilité d'un appariement correct par une méthode de couplage d'enregistrements probabiliste fondée sur une mesure de distance entre les valeurs originales et perturbées.

Mots clés : Couplage d'enregistrements probabiliste ; modèle log-linéaire de Poisson ; bruit additif ; probabilité d'appariement.

1. Introduction

Il y a risque de divulgation lorsque la probabilité est grande qu'un intrus arrive à identifier une personne au moyen des microdonnées diffusées et que des renseignements confidentiels soient révélés. Dans le cas de microdonnées d'échantillon provenant d'enquêtes sociales, le scénario de risque de divulgation repose habituellement sur l'hypothèse qu'un intrus pourrait les apparier à des données provenant de sources publiques disponibles grâce à un ensemble de variables d'identification clés communes aux deux jeux de données. L'identification d'une personne permet ensuite d'obtenir des renseignements de nature délicate et peut mener à la divulgation d'attributs. Pour limiter le risque d'identification, l'organisme statistique applique aux variables d'identification des méthodes de contrôle de la divulgation comprenant la réduction du niveau de détail (ou agrégation), le sous-échantillonnage ou l'utilisation de méthodes de perturbation qui altèrent les données grâce à l'introduction d'erreurs de classification.

L'évaluation du risque de divulgation repose habituellement sur la notion d'unicité dans la population (voir Bethlehem, Keller et Pannekoek, 1990; Skinner et Holmes, 1998; Elamir et Skinner, 2006 et Skinner et Shlomo, 2008). Le cadre de modélisation probabiliste s'appuie sur des hypothèses distributionnelles formulées pour estimer les paramètres de population en vue d'obtenir des mesures du risque de divulgation fondées sur un modèle. La mesure du risque individuel de divulgation est définie comme étant la probabilité qu'un enregistrement unique dans l'échantillon soit apparié aux données de population au moyen d'un ensemble commun de variables clés. La mesure globale du risque de divulgation se calcule par sommation des probabilités d'appariement des enregistrements uniques pour obtenir le nombre prévu d'appariements corrects. Shlomo et Skinner (2010) ont élargi le cadre initial de modélisation probabiliste d'Elamir et Skinner (2006) et de Skinner et Shlomo (2008) afin de tenir compte des erreurs de mesure concernant les variables clés, que ces erreurs se produisent naturellement au fil des étapes de traitement des données ou qu'elles soient introduites à dessein comme méthode de contrôle de la divulgation par perturbation des données.

Dans le cas des microdonnées sur les entreprises provenant d'enquêtes auprès des établissements, l'évaluation du risque de divulgation doit suivre une autre approche. Ces enquêtes sont habituellement caractérisées par de fortes fractions d'échantillonnage; les grandes entreprises étant échantillonnées avec certitude et les distributions étant asymétriques. Par conséquent, les microdonnées sur les entreprises sont généralement traitées comme un recensement de la population complète. Le risque de divulgation repose sur un scénario de divulgation par déduction dans lequel on présume que l'intrus est l'une des entreprises concurrentes se trouvant dans une cellule

[1]Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, Royaume-Uni, courriel : n.shlomo@soton.ac.uk.

d'un tableau qui a eu préalablement connaissance de la distribution des valeurs dans la cellule et qui peut déduire les valeurs de ses concurrentes. Dans ce cas, les variables sensibles sont aussi les variables clés et elles doivent entrer en ligne de compte dans l'évaluation du risque. Ces dernières années, nombre de recherches ont été consacrées à la diffusion de microdonnées issues d'enquêtes-entreprises en appliquant des techniques de contrôle de la divulgation hautement perturbatrices. Pour ces types d'ensembles de données, la méthode de contrôle de la divulgation repose sur la création d'ensembles de données partiellement synthétiques dont les valeurs sont tirées de modèles fondés sur des variables perturbées ainsi que non perturbées (voir Raghunathan, Reiter et Rubin, 2003 et Reiter, 2005).

L'évaluation du risque de divulgation dans le cas d'ensembles de données très perturbés a généralement été effectuée dans le cadre de couplage d'enregistrements probabiliste de Fellegi et Sunter (1969). L'un des premiers exemples est décrit dans Spruill (1982), qui a couplé des microdonnées d'échantillon perturbées aux données d'échantillon originales par une méthode de couplage d'enregistrements fondée sur la distance. Les exemples plus récents font appel à la théorie du couplage d'enregistrements de Fellegi et Sunter (F-S) (Yancy, Winkler et Creecy, 2002; Hawala, Stinson et Abowd, 2005 et Torra, Abowd et Domingo-Ferrer, 2006). Dans le cadre du couplage d'enregistrements de F-S, un poids d'appariement est attribué à chaque paire potentielle. Les poids d'appariement sont triés et des seuils appropriés sont établis en fonction de bornes d'erreur de type I et de type II prédéterminées. Les paires dont le poids d'appariement est élevé sont considérées comme des appariements corrects et celles dont le poids d'appariement est faible, comme des non-appariements. Les poids d'appariement sont utilisés pour calculer la probabilité qu'un appariement soit correct étant donné une concordance.

Dans le présent article, nous rapprochons les deux méthodes d'évaluation du risque de divulgation telles qu'elles sont décrites dans Skinner (2008). À la section 2, nous présentons la notation et la théorie des deux cadres d'évaluation du risque de divulgation — le cadre de couplage probabiliste d'enregistrements de F-S et le cadre de modélisation probabiliste — et nous fournissons une preuve empirique de la relation entre les deux méthodes lorsque l'on tient compte des erreurs découlant de la perturbation. À la section 3, nous illustrons l'évaluation du risque dans le cas d'un ensemble de données très perturbé sur les fermes sucrières provenant d'une enquête de 1982 sur l'industrie de la canne à sucre au Queensland, en Australie (Chambers et Dunstan, 1986). Nous terminons par une conclusion à la section 4.

2. Notation et théorie

Dans cette section, nous décrivons le cadre de couplage d'enregistrements probabiliste de F-S et le cadre de modélisation probabiliste fondé sur la notion d'unicité dans la population tout en tenant compte de la perturbation. Nous rapprochons ces deux cadres.

2.1 Couplage d'enregistrements probabiliste de Fellegi et Sunter

En utilisant la notation employée dans Skinner (2008), soit $\tilde{X}_a$ la valeur du vecteur de variables d'identification clés croisées pour l'unité *a* dans les microdonnées $(a \in s_1)$ et X_b la valeur correspondante pour l'unité b dans la base de données externe $(b \in s_2)$. Notons que s_2 peut être la population *P* ou tout sous-ensemble $s_2 \subset P$. La notation différente de *X* permet aux deux vecteurs d'avoir des valeurs distinctes en raison de l'erreur de classification naturelle dans les données ou de l'application d'une méthode perturbatrice de contrôle de la divulgation au fichier de microdonnées d'échantillon. Désignons cette matrice d'erreurs de classification par :

$$P(=\tilde{X}_a k \mid X_a = j) = \theta_{kj}. \qquad (1)$$

Selon la théorie du couplage d'enregistrements de F-S, un vecteur de comparaison $\gamma(\tilde{X}_a, X_b)$ est calculé pour les paires d'unités $(a,b) \in s_1 \times s_2$ où la fonction $\gamma(.,.)$ prend des valeurs dans un espace de comparaison fini Γ. Pour le scénario de risque de divulgation, nous supposons que l'intrus utilise le vecteur de comparaison pour repérer des paires d'unités qui contiennent la même unité $(a,a) \in s_1 \times s_2$. En général, l'intrus combinera l'appariement exact et l'appariement probabiliste en ne prennat en compte que les paires qui sont regroupées au moyen d'un appariement

exact sur un sous-ensemble $\tilde{s} \subset s_1 \times s_2$. L'intrus cherche à partitionner l'ensemble de paires dans $\tilde{s}$ en un ensemble d'appariements $M = \{(a,b) \in \tilde{s} \mid a \in s_1, b \in s_2, a = b\}$ et de non-appariements $U = \{(a,b) \in \tilde{s} \mid a \in s_1, b \in s_2, a \neq b\}$. L'approche suivie par F-S consiste à définir le rapport de vraisemblance $m(\gamma)/u(\gamma)$ comme étant le poids d'appariement, où $m(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in M)$ et $u(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in U)$. Nous désignons $m(\gamma)$ comme étant la probabilité *m* et $u(\gamma)$, la probabilité *u*. Les valeurs élevées du rapport de vraisemblance sont plus susceptibles d'appartenir à *M* et les valeurs faibles rapport de vraisemblance sont plus susceptibles d'appartenir à *U*. En outre, sous l'hypothèse d'indépendance, la probabilité *m* et la probabilité *u* peuvent être subdivisées en composantes individuelles pour chaque variable clé. Soit $p = P((a,b) \in M)$ la probabilité que la paire soit dans *M*. La probabilité d'un appariement correct $p_{M|\gamma} = P((a,b) \in M \mid \gamma(\tilde{X}_a, X_b))$ peut être calculée en se servant du théorème de Bayes, ce qui donne :

$$p_{M|\gamma} = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)]. \qquad (2)$$

Un intrus pourrait estimer les paramètres d'appariement $m(\gamma), u(\gamma)$ et *p* en couplant les microdonnées diffusées à un fichier de données externe contenant l'ensemble ou des sous-ensembles de la population. Les paramètres peuvent être estimés en utilisant l'algorithme EM, qui est une méthode d'estimation itérative du maximum de vraisemblance dans le cas de données incomplètes. En fonction des paramètres estimés, nous calculons selon (2) la probabilité $p_{M|\gamma}$ d'un appariement correct étant donné une concordance. Ces probabilités peuvent servir de mesures individuelles du risque de divulgation au niveau de l'enregistrement et être agrégées pour obtenir des mesures globales du risque de divulgation.

2.2 Modélisation probabiliste pour mesurer le risque d'identification

Le cadre de modélisation probabiliste employé pour estimer le risque d'identification repose sur une théorie faisant appel à des modèles pour les variables clés catégoriques. Soit $\mathbf{f} = \{f_j\}$ un tableau de fréquences à *q* dimensions, qui est un échantillon tiré d'une table de population $\mathbf{F} = \{F_j\}$, où $j = (j_1,...,j_q)$ indique une cellule, et f_j et F_j désignent la fréquence d'échantillon et de population, respectivement, dans la cellule *j*. Désignons par *n* et *N* respectivement les tailles de l'échantillon et de population et par *J* le nombre de cellules. Nous supposons que les *q* attributs du tableau sont des variables d'identification clés catégoriques. Le risque de divulgation provient des petites cellules, en particulier lorsque $f_j = F_j = 1$ (enregistrements uniques dans l'échantillon et dans la population). Nous nous concentrons sur une mesure globale du risque de divulgation fondée sur les enregistrements uniques dans l'échantillon : $\tau = \sum_j I(f_j = 1)1/F_j$. Cette mesure est le nombre prévu d'appariements corrects si chaque enregistrement unique dans l'échantillon est apparié à une personne choisie au hasard dans la même cellule de population. Nous examinons le cas où f est connue, F est un paramètre inconnu et la quantité τ doit être estimée. Une estimation de τ est donnée par :

$$\hat{\tau} = \sum_j I(f_j = 1)\hat{E}[1/F_j \mid f_j = 1] \qquad (3)$$

où $\hat{E}$ désigne une estimation de l'espérance. La formule (3) est naïve en ce sens qu'elle ne tient pas compte de la possibilité d'une erreur de classification. Une hypothèse courante dans la littérature sur les tableaux de fréquence est $F_j \sim Poisson(\lambda_j)$, indépendamment, où $\sum_j F_j = N$ est un paramètre aléatoire. L'échantillonnage binomial (ou de Poisson) à partir de F_j signifie que $f_j \mid F_j \sim Bin(F_j, \pi_j)$ indépendamment, où π_j est la fraction d'échantillonnage dans la cellule *j*. Par des calculs standards, nous obtenons alors :

$$f_j \sim Poisson(\lambda_j \pi_j) \text{ et } F_j \mid f_j \sim f_j + Poisson(\lambda_j(1-\pi_j)), \qquad (4)$$

où $F_j \mid f_j$ sont conditionnellement indépendantes.

Nous suivons l'approche élaborée dans Skinner et Holmes (1998), Elamir et Skinner (2006) et Skinner et Shlomo (2008), et utilisons des modèles log-linéaires pour estimer les paramètres de population et ceux du risque

d'identification. Les chiffres d'échantillon $\{f_j\}$ sont utilisés pour ajuster le modèle log-linéaire $\log \mu_j = x_j'\beta$, où $\mu_j = \lambda_j \pi_j$, afin d'obtenir les estimations des paramètres $\hat{\lambda}_j = \hat{\mu}_j / \pi_j$. Sous échantillonnage aléatoire simple et $\pi_j = \pi$ pour toute valeur de *j*, l'estimateur du maximum de vraisemblance (EMV) $\hat{\beta}$ peut être obtenu en résolvant les équations de score : $\sum_j [f_j - \exp(x_j'\beta)]x_j = 0$. Sous un plan de sondage complexe et des poids différentiels, $\hat{\beta}$ peut être estimé par une méthode de pseudo-vraisemblance. En utilisant la seconde partie de l'expression (4), les mesures individuelles prévues du risque de divulgation pour la cellule *j* sont définies par :

$$E_{\lambda_j}(1/F_j \mid f_j = 1) = [1 - e^{-\lambda_j(1-\pi)}]/[\lambda_j(1-\pi)]. \qquad (5)$$

En introduisant $\hat{\lambda}_j$ pour λ_j dans l'expression (5), on obtient les estimations souhaitées $\hat{E}_{\hat{\lambda}_j}[1/F_j \mid f_j = 1]$, puis l'estimation $\hat{\tau}$ de l'expression (3).

L'approche de modélisation probabiliste ne tient pas compte du cas des erreurs de classification qui se produisent naturellement dans les enquêtes ou qui sont introduites délibérément dans les données comme méthode de contrôle de la divulgation. Shlomo et Skinner (2010) ont défini des mesures du risque de divulgation qui tiennent compte des erreurs de classification. Dans le cas qui nous occupe, la mesure du risque de divulgation est :

$$[\theta_{jj}/(1-\pi\theta_{jj})]/[\sum_k F_k \theta_{jk}/(1-\pi\theta_{jk})] \qquad (6)$$

et il s'ensuit que (6) est inférieur à $1/F_j$, l'égalité étant vérifiée en l'absence d'erreur de classification.

Si la fraction d'échantillonnage est petite, comme dans de nombreuses enquêtes sociales, nous pouvons obtenir une approximation de (6) par :

$$\theta_{jj} / \tilde{F}_j \qquad (7)$$

où $\tilde{F}_j$ désigne la population provenant de l'échantillon perturbé. Notons que les approximations en (7) ne dépendent pas de θ_{jk} lorsque $j \neq k$, de sorte qu'il n'est pas nécessaire de connaître ces probabilités pour estimer le risque si des estimations « acceptables » de θ_{jj} et $\tilde{F}_j$ sont disponibles. Dans l'expression (7), la définition du risque s'applique à un enregistrement spécifique. La mesure agrégée sur l'ensemble des enregistrements uniques dans l'échantillon, définie d'après l'expression (7), est

$$\tau_\theta = \sum_{j \in SU} \theta_{jj} / \tilde{F}_j \qquad (8)$$

où *SU* est l'ensemble des valeurs des variables clés qui sont des enregistrements uniques dans l'échantillon perturbé. Ainsi qu'en l'absence d'erreur de classification, cette mesure peut être interprétée comme étant le nombre prévu d'appariements corrects parmi les enregistrements uniques dans l'échantillon.

Puisque les valeurs de F_j ou $\tilde{F}_j$ figurant dans les expressions (7) et (8) sont inconnues, il faut les estimer. L'expression (8) fournit une façon simple d'étendre l'approche de modélisation log-linéaire pour autant que θ_{jj} soit connu (ce qui est le cas si l'organisme statistique perturbe délibérément les données en vue de contrôler la divulgation). À partir des chiffres de l'échantillon perturbé $\tilde{f}_j$, $j = 1,...,J$, nous estimons $1/\tilde{F}_j$ en ajustant un modèle log-linéaire à ces fréquences $\tilde{f}_j$, $j = 1,...,J$. Cela nous donne une estimation $\hat{E}(1/\tilde{F}_j \mid \tilde{f}_j = 1)$ fondée sur les hypothèses que les chiffres de population et d'échantillon suivent une loi de Poisson. Ces estimations doivent être multipliées par les valeurs de θ_{jj} puis additionnées si des mesures agrégées de la forme de (8) sont requises.

2.3 Rapprochement des cadres

Skinner (2008) lie le cadre de couplage d'enregistrements de F-S au cadre de modélisation probabiliste en donnant les exemples qui suivent.

Exemple 1 : Supposons qu'aucune erreur de classification ne s'est produite, c'est-à-dire que $\tilde{X}_a = X_a$ tant dans la population (*P*) que dans l'échantillon (*s*) et que la véritable situation d'appariement est connue de l'organisme. Supposons que l'échantillon (*s*) a été tiré par échantillonnage aléatoire simple de la population *P*. Nous calculons la table de décision du tableau 1 pour chaque $X_a = j$ dans l'échantillon réalisé, où les lignes représentent l'état binaire de concordance/discordance des valeurs sur le vecteur de comparaison $\gamma(X_a, X_b)$ pour les paires $(a,b) \in s \times P$ et où les colonnes indiquent l'état d'appariement. À partir du tableau 1, nous calculons directement $p_{M|\gamma} = 1/F_j$. Nous obtenons aussi $m(\gamma) = f_j / n$, $u(\gamma) = f_j(F_j - 1)/n(N-1)$ et la probabilité d'un appariement correct $p = 1/N$. L'utilisation de la formule de Bayes donne :

$$p_{M|\gamma} = \frac{1/N \times f_j / n}{1/N \times f_j / n + (1-1/N) f_j (F_j - 1)/n(N-1)} = \frac{1}{F_j}. \qquad (9)$$

Une faible valeur de F_j se traduit donc par une probabilité élevée d'appariement correct s'il y a concordance des valeurs dans le vecteur de comparaison.

Tableau 1 : Tableau de contingence de l'état binaire de concordance des variables et de la situation d'appariement pour $X_a = j$ sans erreur de classification

	Non-appariement	Appariement	Total
Discordance	$n(N-1) - f_k(F_k - 1)$	$n - f_k$	$Nn - f_k F_k$
Concordance	$f_k(F_k - 1)$	f_k	$f_k F_k$
Total	$n(N-1)$	n	Nn

Exemple 2 : Dans le prolongement de l'exemple 1, supposons maintenant que les microdonnées présentent des erreurs de classification (qui se sont produites naturellement ou qui ont été pertubées délibérément pour contrôler le risque de divulgation). Soit $\tilde{f}_j$ les chiffres observées dans l'échantillon classé incorrectement, avec $\tilde{X}_a = j$ calculée par $\tilde{f}_j = \theta_{jj} f_j + \sum_{k \neq j} \theta_{jk}$. Nous calculons le tableau de contingence du tableau 2 pour l'échantillon classé incorrectement réalisé pour chaque $\tilde{X}_a = j$, où les lignes représentent l'état binaire de concordance/discordance sur le vecteur de comparaison $\gamma(\tilde{X}_a, X_b)$ pour les paires $(a,b) \in s \times P$ et où les colonnes indiquent la situation d'appariement.

Tableau 2 : Tableau de contingence de l'état binaire de concordance des variables et de la situation d'appariement pour $\tilde{X}_a = j$ avec erreurs de classification

	Non-appariement	Appariement	Total
Discordance	$Nn - n - \tilde{f}_k F_k + M_{kk} f_k$	$n - M_{kk} f_k$	$Nn - \tilde{f}_k F_k$
Concordance	$\tilde{f}_k F_k - M_{kk} f_k$	$M_{kk} f_k$	$\tilde{f}_k F_k$
Total	$Nn - n$	n	Nn

À partir du tableau 2, nous pouvons calculer directement $p_{M|\gamma} = \theta_{jj} f_j / \tilde{f}_j F_j \approx \theta_{jj} / \pi \tilde{f}_j \approx \theta_{jj} / \tilde{F}_j$ où $\tilde{F}_j$ est le nombre d'unités dans la population (*P*) pour lesquelles $\tilde{X}_a = j$ (en imaginant que l'erreur de classification se produit avant l'échantillonnage). Nous obtenons aussi $m(\gamma) = \theta_{jj} f_j / n$, $u(\gamma) = (\tilde{f}_j F_j - \theta_{jj} f_j)/n(N-1)$ et la probabilité d'un appariement correct $p = 1/N$. L'utilisation de la formule de Bayes nous donne :

$$p_{M|\gamma} = \frac{1/N \times \theta_{jj} f_j / n}{1/N \times \theta_{jj} f_j / n + (1-1/N)(\tilde{f}_j F_j - \theta_{jj} f_j)/n(N-1)} \approx \frac{\theta_{jj}}{\pi \tilde{f}_j} \approx \frac{\theta_{jj}}{\tilde{F}_j}. \quad (10)$$

L'expression (10) est semblable à l'expression (8) dérivée du cadre de modélisation probabiliste.

3. Évaluation du risque de divulgation de microdonnées d'entreprises perturbées

Nous montrons comment nous pouvons évaluer le risque de divulgation de microdonnées d'entreprises fortement perturbées. Puisque les microdonnées d'entreprises sont traitées comme des données de recensement, la probabilité d'appariement ne dépend, dans les deux cadres, que de la probabilité θ_{jj} de ne pas être dans l'erreur. Dans ces conditions, nous émettons l'hypothèse d'indépendance conditionnelle de F-S et analysons la probabilité de non-perturbation séparément pour chacune des variables, tant les variables d'identification que les variables sensibles.

3.1 Perturbation de l'ensemble de données

Nous utilisons l'ensemble de données sur les fermes sucrières correspondant à un échantillon de 338 fermes sucrières du Queensland. L'ensemble de données comporte une variable catégorique nominale (région) et cinq variables continues (superficie totale, quantité de récoltes, recettes, coûts et profits, les profits représentant la différence entre les recettes et les coûts). Les techniques classiques de contrôle de la divulgation statistique ont été appliquées pour perturber les données sur les fermes cultivant la canne à sucre dans le but de réduire les risques de divulgation d'identité et de divulgation par déduction. Tout d'abord, cinq enregistrements ont été supprimés parce qu'ils contenaient des valeurs de recettes aberrantes (plus de 300 000 $). La superficie, qui est la variable d'identification clé, a été rendue plus grossière (agrégée) par sa catégorisation en neuf groupes. La réduction du niveau de détail d'une variable d'identification permet de réduire le risque de divulgation en rendant cette variable plus difficile à utiliser en tant que variable d'appariement et procure généralement plus de protection qu'une méthode de perturbation dans la mesure où elle réduit la quantité de données diffusées. Les variables cibles de l'enquête, à savoir les récoltes, les recettes, les coûts et les profits, ont été perturbées par l'ajout d'un bruit suivant une loi normale multivariée dans des petits groupes. La loi normale multivariée est choisie afin de préserver la moyenne et la structure de covariance des variables cibles perturbées, et d'assurer le respect de la contrainte de vérification voulant que les profits soient égaux aux recettes moins les coûts (Shlomo et DeWaal, 2008).

L'ajout d'un bruit s'est effectué comme suit : considérons les quatre variables cibles récoltes (*t*), recettes (*x*), coûts (*y*) et profits (*z*), où $x - y = z$ et supposons qu'elles ont une moyenne conjointe $\boldsymbol{\mu}^{\mathbf{T}} = (\boldsymbol{\mu}_{\mathbf{t}}, \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\mu}_{\mathbf{z}})$ et une matrice de covariance $\boldsymbol{\Sigma}$. Nous générons un bruit aléatoire multivarié $(\varepsilon_t, \varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma})$, où l'exposant T désigne la transposée. Afin de préserver les sous-totaux et de limiter la quantité de bruit, nous avons généré le bruit aléatoire dans des petits groupes d'enregistrements définis par les quintiles des recettes (notons que nous laissons tomber l'indice désignant les quintiles). Le vecteur $\boldsymbol{\mu}'$ contient les moyennes corrigées de chacune des quatre variables *t*, x, y et z dans le quintile sur la base d'un paramètre de bruit δ et en calculant $d_1 = \sqrt{(1-\delta^2)}$ et $d_2 = \sqrt{\delta^2}$:

$$\boldsymbol{\mu}'^{\mathrm{T}} = (\boldsymbol{\mu}'_t, \boldsymbol{\mu}'_x, \boldsymbol{\mu}'_y, \boldsymbol{\mu}'_z) = (\frac{1-d_1}{d_2}\boldsymbol{\mu}_t, \frac{1-d_1}{d_2}\boldsymbol{\mu}_x, \frac{1-d_1}{d_2}\boldsymbol{\mu}_y, \frac{1-d_1}{d_2}\boldsymbol{\mu}_z).$$

La matrice de covariance $\boldsymbol{\Sigma}$ est la matrice de covariance originale des quatre variables dans chaque quintile. Pour chaque variable, nous calculons la combinaison linéaire de la variable originale et du bruit aléatoire généré plus haut, par exemple, pour l'enregistrement *i* : $z'_i = d_1 \times z_i + d_2 \times \varepsilon_{zi}$. Le vecteur de moyenne et la matrice de covariance demeurent les mêmes que pour les données originales et, après la perturbation, la contrainte d'additivité est exactement préservée.

3.2 Méthode d'évaluation du risqué de divulgation proposée

Les microdonnées perturbées sur les fermes sucrières sont traitées comme des données de recensement, si bien que nous avons réduit le niveau de détail de la variable d'identification et perturbé les variables sensibles, et toutes sont utilisées dans le processus d'évaluation du risque. Nous supposons aussi que tous les enregistrements sont des enregistrements uniques dans la population puisque la classification croisée des variables d'identification et des variables sensibles produit des enregistrements uniques. L'évaluation du risque sera effectuée par réappariement de l'ensemble de données perturbé à l'ensemble de données original et le calcul de la probabilité d'appariement selon (8) ou (10), qui ne dépend dans ce cas que de θ_{jj}, soit la probabilité que les enregistrements dans la cellule *j* des variables croisées ne soient pas perturbés. Dans le cas qui nous occupe, toutes les cellules contiennent des enregistrements uniques.

Nous définissons cette probabilité en fonction d'une mesure prenant une valeur comprise entre 0 et 1 qui correspond à la distance (normalisée) entre les valeurs originale et perturbée de chaque variable dans chaque enregistrement. Pour chaque enregistrement, la probabilité finale d'appariement s'obtient par sommation pondérée des mesures sur toutes les variables. Dans le cas de variables continues, la différence entre les valeurs originale et perturbée d'une variable correspond au bruit généré par la loi normale multivariée. Pour la variable dont on a réduit le niveau de détail, nous prenons la distance à partir du point médian de l'intervalle. Pour la valeur *i* d'une variable *p*, définissons la différence entre les valeurs originale et perturbée comme étant ε_i^p.

Nous examinons deux options pour calculer la mesure de distance prenant les valeurs comprises entre 0 et 1 :

1. Option 1 : Normaliser la différence de valeur *i* pour la variable *p* : $Z_i^p = [\varepsilon_i^p - \bar{\varepsilon}_i^p]/s_\varepsilon^2$ et calculer : $Dist_i^p = 1 - |1 - 2\Phi(Z_i^p)|$.
2. Option 2 : Utiliser la fonction exponentielle comme il suit : $Dist_i^p = \exp[-|\varepsilon_i^p|/med(\varepsilon_i^p)]$ où $med(\varepsilon_i^p)$ représente la valeur médiane de ε_i^p.

Nous appliquons la méthode de couplage d'enregistrements de F-S pour obtenir la probabilité d'appariement de chaque paire et agrégeons ces probabilités sur les vrais appariements afin d'obtenir le nombre prévu d'appariements corrects comme dans les expressions (8) et (10). En outre, la méthode de couplage d'enregistrements de F-S permet le calcul d'autres types de mesures du risque qui reflètent l'incertitude quant à la capacité de l'intrus d'inférer un appariement correct, à savoir la proportion d'appariements corrects parmi les paires déclarées et la cote représentant le risque d'un appariement correct étant donné une concordance (nombre de paires déclarées qui sont de vrais appariements divisés par le nombre de paires déclarées qui sont de faux appariements). L'ensemble de données contenant 333 enregistrements a fait l'objet d'un groupage en se fondant sur la région (sans perturbation), ce qui a donné 31 367 paires possibles. Nous procédons ensuite au couplage des enregistrements sur toutes les autres variables : superficie avec niveau de détail réduit et récoltes, recettes et coûts avec ajout d'un bruit. Pour chaque enregistrement, nous proposons de calculer une somme pondérée des mesures de distance entre les variables en posant l'hypothèse d'indépendance conditionnelle du cadre de F-S. Les probabilités *m* du couplage d'enregistrements de F-S sont représentées par les mesures de distance puisqu'elles tiennent compte des erreurs induites par les méthodes de contrôle de la divulgation. Les poids devraient donc refléter les probabilités *u*, c'est-à-dire la cote (odds) exprimant le risque que, étant donné une concordance sur une valeur, la paire soit un vrai appariement. Les poids sont calculés par une régression logistique dans laquelle la variable réponse est égale à 1 pour un vrai appariement et à 0 autrement, et où les variables indépendantes sont les mesures de distance. Les cotes sont ensuite normalisées afin que leur somme soit égale à 1. Nous comparons la moyenne pondérée des mesures de distance à la situation où l'on calcule la moyenne simple des mesures de distance.

Au tableau 3, nous examinons la méthode proposée pour évaluer le risque de divulgation que pose l'ensemble de données sur les fermes sucrières sous deux niveaux de perturbation et deux options de pondération des mesures de distance. Le seuil pour le couplage d'enregistrements de F-S est déterminé par une erreur de type I prédéterminée de 1,4 %.

Tableau 3 : Résultats de la procédure de couplage d'enregistrements pour l'évaluation du risque de divulgation dans l'ensemble de données d'entreprises perturbé (erreur de type I de 1,4 %)

		$\delta = 0.4$		$\delta = 0.7$	
		Option 1	Option 2	Option 1	Option 2
Poids égaux	Vrais appariements/paires déclarées	0,297	0,290	0,160	0,151
	Vrais appariements/faux appariements	0,423	0,409	0,191	0,178
	Somme des probabilités d'appariement	307,5	290,0	289,8	263,9
Poids fondés sur des cotes normalisées	Vrais appariements/paires déclarées	0,307	0,313	0,168	0,175
	Vrais appariements/faux appariements	0,443	0,455	0,201	0,213
	Somme des probabilités d'appariement	309,0	295,6	299,9	292,7

Le tableau 3 montre qu'un risque de divulgation plus élevé est associé au taux de perturbation plus faible $\delta = 0,4.$ L'emploi de la moyenne pondérée des mesures de distance donne plus de puissance au cadre de couplage d'enregistrements que celui de la moyenne simple. Les deux options retenues pour les mesures de distance donnent des résultats incohérents pour les deux mesures du risque de F-S fondées sur le rapport du nombre de vrais appariements au nombre de paires déclarées et au rapport du nombre de vrais appariements au nombre de faux appariements. Alors que l'option 1 produit des mesures du risque de divulgation plus élevées que l'option 2 sous une moyenne simple avec poids égaux, le résultat inverse est observé lorsqu'on utilise la moyenne pondérée par les cotes normalisées, c'est-à-dire que l'option 1 donne des mesures plus faibles du risque de divulgation de F-S que l'option 2. Comme dans le cas de toute méthode de perturbation, le nombre prévu d'appariements corrects obtenu par sommation des probabilités d'appariement sur les vrais appariements est élevé. La perturbation hausse toutefois le niveau d'incertitude relatif aux appariements corrects, comme le montre le taux de vrais appariements par rapport aux faux appariements. Selon les résultats de la présente étude, la méthode de couplage d'enregistrements la plus puissante consisterait à utiliser la mesure de distance de l'option 2 et la moyenne pondérée des mesures de distance fondée sur les cotes normalisées.

4. Conclusion

Dans le présent article, nous avons fourni la preuve du rapprochement entre le cadre de couplage d'enregistrements de F-S et le cadre de modélisation probabiliste pour l'estimation du risque de divulgation au moyen d'une probabilité d'appariement qui peut être calculée à l'aide des deux cadres. Nous montrons comment cette probabilité d'appariement peut être calculée à l'aide du cadre de couplage d'enregistrements de F-S dans le cas d'un ensemble de données d'entreprises fortement perturbé, qui est habituellement traité comme un recensement et dont les variables sensibles ainsi que les variables d'identification sont incluses dans le scénario de risque de divulgation. En outre, le cadre de couplage d'enregistrements de F-S permet d'autres types de mesures du risque qui reflètent l'incertitude quant à la capacité de l'intrus d'obtenir un appariement correct.

Remerciements

Ces travaux sont financés par l'entente de subvention n° 244767, sous le thème 8 du septième programme-cadre de recherche de l'Union européenne, sciences socioéconomiques et humaines : BLUE-ETS.

Bibliographie

Bethlehem, J., Keller, W. et J. Pannekoek (1990), « Disclosure control of microdata », *Journal of the American Statistical Association*, 85, p. 38 à 45.

Chambers, R.L. et R. Dunstan (1986), « Estimating distribution functions from survey data », *Biometrika*, 73, p. 597 à 604.

Elamir, E. et C.J. Skinner (2006), « Record-level measures of disclosure risk for survey micro-data », *Journal of Official Statistics*, 22, p. 525 à 539.

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

Hawala, S., Stinson, M. et J. Abowd (2005), « Disclosure risk assessment through record linkage », dans : Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva.

Raghunathan, T.E., Reiter, J. et D. Rubin (2003), « Multiple imputation for statistical disclosure limitation », *Journal of Official Statistics*, 19, no. 1, p. 1 à 16.

Reiter, J.P. (2005), « Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study », *Journal of the Royal Statistical Society*, A, vol.168, no.1, p. 185 à 205.

Shlomo, N. et T. De Waal (2008), « Protection of micro-data subject to edit constraints against statistical disclosure », *Journal of Official Statistics*, 24, no. 2, p. 1 à 26.

Shlomo, N. et C.J. Skinner (2010), « Assessing the protection provided by misclassification-dased disclosure limitation methods for survey microdata », *Annals of Applied Statistics*, vol. 4, no. 3, p. 1291 à 1310.

Skinner, C.J. (2008), « Assessing disclosure risk for record linkage », dans J. Domingo-Ferrer et Y. Saygin (éds.) *Privacy in Statistical Databases*, Lecture Notes in Computer Science 5262, Berlin: Springer, p. 166 à 176.

Skinner, C. et D. Holmes (1998), « Estimating the re-identification risk per record in microdata », *Journal of Official Statistics*, 14, p. 361 à 372.

Skinner, C.J. et N. Shlomo (2008), « Assessing identification risk in survey microdata using log-linear models », *Journal of the American Statistical Association*, vol. 103, no. 483, p. 989 à 1001.

Spruill, N.L. (1982), « Measures of confidentiality », *Proceedings of the Survey Research Methods Section of the American Statistical Association*, p. 260 à 265.

Torra, V., Abowd, J.M. et J. Domingo-Ferrer (2006), « Using Mahalanobis distance-based record linkage for disclosure risk assessment » dans J. Domingo-Ferrer et L. Franconi (éds.) *Privacy in Statistical Databases*, Lecture Notes in Compute Science, 4302, Berlin: Springer, p. 233 à 242.

Yancey, W.E.,Winkler, W.E. et R.H. Creecy (2002), « Disclosure risk assessment in perturbation micro-data protection », dans J. Domingo-Ferrer (éd.) *Inference Control in Statistical Databases*, New York: Springer, p. 135 à 151.

Le couplage d'enregistrements probabiliste préservant la confidentialité de A à Z : un exemple d'utilisation du Système généralisé de couplage d'enregistrements à SwissLinkage*

Adrian Spoerri, Kurt Schmidlin, Rainer Schnell et Kerri Clough-Gorr[1]

Résumé

Le couplage d'enregistrements probabiliste figure en tête des priorités de recherche actuelles en matière de santé. La contraction des ressources de recherche limite la production de nouvelles données. Par conséquent, les données existantes – souvent des données recueillies de manière courante – devraient être analysées dans la mesure du possible. Les questions de protection des renseignements empêchent d'utiliser les meilleures données existantes dans de nombreux projets de couplage d'enregistrements. L'élaboration récente de méthodes de couplage d'enregistrements préservant la confidentialité, comme l'emploi de filtres de Bloom, permet le chiffrement des noms, des adresses et d'autres données personnelles protégées. Jusqu'à récemment, il était impossible d'appliquer des méthodes probabilistes de couplage d'enregistrements pour chiffrer des données. La présente communication a pour but de faire la démonstration du processus complet d'un couplage d'enregistrements probabiliste préservant la confidentialité qui consiste à combiner les données du Registre suisse du cancer de l'enfant (RSCE) avec celles de l'Institut national pour l'épidémiologie et l'enregistrement du cancer (NICER) en utilisant des filtres de Bloom et le logiciel de couplage d'enregistrements SGCE. Comme l'équipe chargée du couplage d'enregistrements n'a pas le droit d'examiner des données personnelles protégées (pour assurer le respect de la confidentialité), le prétraitement des variables de couplage doit être automatisé dans les locaux où se trouvent les registres. Nous ferons la démonstration d'un outil de prétraitement, ainsi que de chiffrage et d'exportation des données qui sera utilisé par la personne responsable au RSCE et au NICER. Le processus de couplage d'enregistrements sera ensuite exécuté par SwissLinkage (tierce partie indépendante). L'équipe du couplage d'enregistrements n'aura jamais accès aux données protégées (par exemple, les noms) et exécutera le couplage en se servant du Système généralisé de couplage d'enregistrements (SGCE). Le SGCE de Statistique Canada n'a pas été conçu pour effectuer des couplages d'enregistrements portant sur des variables chiffrées. Nous montrons l'application des fonctions de hachage des filtres de Bloom et le calcul d'une mesure de similitude (coefficient de Dice) en utilisant le SGCE. La capacité d'apparier des sources de données en se servant de renseignements discriminants, comme le nom d'un patient, sans violer l'obligation de secret offerte par la méthode de couplage d'enregistrements probabiliste préservant la confidentialité pourrait transformer la recherche en épidémiologie de manière éthique en permettant d'avoir accès, en préservant l'anonymat, à des données sur le cancer qui ne pouvaient être utilisées avant cela.

[1]Adrian Spoerri et Kurt Schmidlin, Institut de médecine sociale et préventive, Université de Berne, Suisse ; Rainer Schnell, Département des sciences sociales, University of Duisburg-Essen, Allemagne ; Kerri Clough-Gorr, Institut de médecine sociale et préventive, Université de Berne, Suisse et Institut national pour l'épidémiologie et l'enregistrement du cancer, Suisse.

* SwissLinkage est un centre d'excellence en couplage d'enregistrements à l'Université de Berne.

Produits normalisés et classifications de niveau régional pour les populations minoritaires dans le cadre du Recensement de 2011 de l'Angleterre et du pays de Galles

Joe Traynor et Emma White[1]

Résumé

L'avantage qu'offre un recensement comparativement aux enquêtes par sondage est la capacité de fournir un large éventail de renseignements sur la population d'une petite région géographique.

Des contraintes de contrôle de la divulgation sont conçues pour l'ensemble de la population en vue de protéger les renseignements personnels, y compris ceux des groupes minoritaires, au niveau géographique le plus détaillé. Cependant, il existe des régions où les populations minoritaires sont suffisamment grandes pour permettre de produire des données détaillées à leur sujet.

Pour aborder cette question, l'Office for National Statistics prévoit produire un algorithme qui pourrait être appliqué à toute population minoritaire, par exemple un groupe ethnique ou religieux, de façon que, dès que la population en question atteint une taille seuil spécifiée à un niveau de détail géographique déterminé, des données de sortie prédéfinies soient produites pour répondre aux besoins de cette collectivité particulière.

La communication décrira les étapes suivies pour concevoir et tester l'algorithme, y compris le processus connexe de consultation.

La conception fructueuse d'un tel algorithme offrira plusieurs avantages, à savoir :

- l'ajout de valeur aux données de recensement déjà recueillies ;
- la détermination des régions où les groupes minoritaires sont concentrés ;
- la production d'une plus grande quantité de données ayant trait directement aux groupes minoritaires, qui permettra une planification et des décisions stratégiques plus éclairées et représentatives ;
- l'apport d'un soutien pour le dossier de l'égalité.

[1]Joe Traynor et Emma White, Office for National Statistics, Angleterre.

SÉANCE 5A

DISCOURS DU GAGNANT DU PRIX WAKSBERG

Modélisation des données d'enquêtes complexes : Pourquoi les modéliser? Pourquoi est-ce un problème? Comment pouvons-nous le résoudre?

Danny Pfeffermann[1]

Résumé

Les données d'enquête sont souvent utilisées pour procéder à des inférences analytiques sur des modèles statistiques que l'on suppose être vérifiés pour la population de laquelle est tiré l'échantillon. Les données d'enquête diffèrent habituellement des autres ensembles de données en ce qui a trait à cinq aspects importants.

1- Les échantillons sont tirés au hasard en appliquant des probabilités de sélection connues, ce qui permet d'utiliser la distribution aléatoire sur toutes les sélections possibles d'échantillons comme base de l'inférence plutôt que la distribution qui sous-tend le modèle de population. L'utilisation d'une combinaison des deux distributions est fréquente.

2- Les probabilités de sélection de l'échantillon, au moins à certains degrés de l'échantillonnage, sont généralement inégales; quand ces probabilités sont reliées à la variable de résultat du modèle, l'échantillonnage devient informatif et le modèle vérifié pour l'échantillon diffère alors du modèle de la population cible.

3- Les données d'enquête sont presque inévitablement sujettes à diverses formes de non réponse, souvent d'une grandeur considérable, qui de nouveau peuvent fausser le modèle de population si la propension à répondre est corrélée à la variable de résultat d'intérêt.

4- Les données sont souvent groupées à cause de l'utilisation d'échantillons en grappes. Les grappes sont des « unités naturelles » et les observations à l'intérieur d'une même grappe sont par conséquent corrélées.

5- Les données dont dispose le modéliseur sont parfois masquées afin de préserver l'anonymat des répondants.

De nombreuses approches ont été proposées dans la littérature pour estimer les modèles de population d'après des données d'enquête complexes possédant les caractéristiques susmentionnées. Les approches se distinguent par les conditions qui sous tendent leur utilisation, les données requises pour leur application, les tests d'adéquation de l'ajustement du modèle, les objectifs d'inférence qu'elles permettent de satisfaire, l'efficacité statistique, les demandes de ressources informatiques et les compétences que doivent posséder les analystes qui ajustent les modèles. Cette hétérogénéité signifie qu'aucune approche ne peut être considérée comme étant la meilleure dans toutes les situations. Cela étant, la question fondamentale est de savoir quelle(s) approche(s) il convient d'utiliser pour une application particulière.

Dans la présente communication, je passe en revue les diverses approches proposées dans la littérature pour traiter ces caractéristiques, en discutant leurs mérites et leurs limites à la lumière des propriétés susmentionnées. Je présente aussi les résultats de simulations conçues pour comparer les approches sur le plan du biais, de la variance et du taux de couverture dans le cas de l'estimation de modèles de régression stratifiés.

L'article est publié dans *Techniques d'enquête*, décembre 2011.

[1]Danny Pfeffermann, Hebrew University of Jerusalem, Israël, et University of Southampton, Royaume-Uni.

SÉANCE 6A

NORMES ET LIGNES DIRECTRICES POUR LA CONCEPTION ET LA MISE À L'ESSAI DE QUESTIONNAIRES INTERNET

Lignes directrices pour l'élaboration de questionnaires électroniques :problèmes et défis dans un environnement en évolution

Anne-Marie Côté, David Lawrence et Paul Kelly[1]

Résumé

L'utilisation de modes de collecte multiples, et plus particulièrement l'emploi de questionnaires à remplir soi même en ligne, présente des défis pour les enquêtes de Statistique Canada. Afin de veiller à ce que les données recueillies répondent à nos exigences en matière de qualité, nous devons élaborer avec beaucoup de soin le cadre conceptuel pour les questionnaires sur Internet. Les procédures de conception et de mise à l'essai, ainsi que le contenu de l'enquête, doivent être pris en compte. Une solution intégrée de création de questionnaires en ligne, qui comprend l'élaboration de processus normalisés pour réaliser les questionnaires en ligne, est présentement mise en œuvre. Le présent article porte sur certains des enjeux et défis liés à l'élaboration de lignes directrices concernant la conception et l'élaboration de questionnaires sur Internet à Statistique Canada.

Mots clés : Questionnaire électronique ; lignes directrices ; normes ; mode de collecte.

1. Introduction

Statistique Canada offre diverses solutions de collecte électronique des données (CED) aux répondants, depuis un certain temps. En 2010, une nouvelle initiative à l'échelle de l'organisme a été approuvée, en vue de l'utilisation de questionnaires électroniques (QE) comme principal mode de collecte. L'initiative est le résultat de tentatives répétées de préparation d'une solution de CED conforme aux exigences strictes en matière de confidentialité et de sécurité de Statistique Canada, techniquement durable et respectant les attentes des répondants. L'option doit aussi être conforme aux lignes directrices sur la Normalisation des sites Internet (NSI) du gouvernement du Canada et répondre aux exigences en matière d'accessibilité, afin d'assurer un accès équitable à l'ensemble du contenu des sites Web du gouvernement du Canada.

Le présent article fournit un bref aperçu de l'initiative intégrée, surtout en ce qui a trait à la collecte de données au moyen de QE. Il porte aussi sur l'établissement et la mise en œuvre de lignes directrices et de normes pour les QE à Statistique Canada, c'est-à-dire la façon dont les lignes directrices actuelles ont été établies, et comment elles continueront d'évoluer.

2. Nouveau projet de QE

2.1 Objectifs

L'accent qui est mis sur les QE comme principal mode de collecte est motivé par l'évolution des attentes de la population canadienne, ainsi que par la situation économique actuelle. La vision de Statistique Canada est de concevoir des enquêtes dans lesquelles les QE seront le principal mode de collecte et le premier mode dans un environnement séquentiel à plusieurs modes. Cette démarche est motivée par la nécessité de trouver des façons plus efficaces de procéder à la collecte des données, ainsi que de capitaliser sur les fonctions génériques, en vue de réduire la nécessité d'utiliser des outils de collecte personnalisés.

[1]Anne-Marie Côté, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (anne-marie.cote@statcan.gc.ca); David Lawrence, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) Canada, K1A 0T6 (dave.lawrence@statcan.gc.ca); Paul Kelly, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) Canada, K1A 0T6 (paul.kelly@statcan.gc.ca).

En offrant une option de QE, Statistique Canada atteindra les objectifs suivants :

- Fournir un mode de collecte qui est pratique pour les répondants et disponible 24 heures par jour, 7 jours par semaine.
- Réduire les coûts liés aux enquêtes administrées par des intervieweurs et envoyées par la poste, sous réserve de taux appropriés d'adoption de la réponse par Internet.
- Améliorer l'actualité des réponses par Internet par rapport aux réponses sur questionnaires papier.
- Réduire le volet d'élaboration puisque Statistique Canada prévoit une période d'élaboration plus courte pour les applications de QE que celle pour les applications d'Interview Assistée par Ordinateur (IAO).
- Maintenir et améliorer la qualité des données, grâce à l'utilisation de contrôles en ligne, d'enchaînements automatisés des questions et de fonctions d'aide supplémentaires, pour faire en sorte que les questions et les concepts d'enquête soient compris par les répondants, ce qui les amènera à fournir des réponses de plus grande qualité.
- Répondre aux attentes des répondants, ceux-ci étant de plus en plus nombreux à indiquer qu'ils préféreraient remplir un QE, plutôt que de passer du temps avec un intervieweur ou de remplir un questionnaire papier.

2.2 Élaboration et mise en œuvre

Le processus d'élaboration des QE comporte six étapes principales : évaluation initiale, spécifications d'entrée, production des questionnaires électroniques, mise à l'essai, collecte des données et évaluation rétrospective (bilan).

Étant donné que chaque enquête est unique, la première étape est celle de l'évaluation initiale. Chaque enquête est passée en revue par des experts de la collecte, de concert avec le secteur spécialisé, afin de déterminer comment la stratégie de questionnaire électronique correspondra aux exigences existantes en matière de collecte et de traitement.

Ensuite, les spécifications d'entrée des QE sont élaborées, à partir de fonctions génériques et en respectant les lignes directrices et les normes en matière de QE existantes. Au besoin, des fonctions supplémentaires peuvent être élaborées pour répondre aux besoins d'une enquête particulière. Toutefois, ces fonctions doivent être génériques pour que Statistique Canada puisse les réutiliser dans d'autres enquêtes.

Les QE sont générés automatiquement à partir d'un modèle générique de spécifications d'entrée dynamiques. Autrement dit, une fois les spécifications des questionnaires (par exemple, questions, enchaînements, messages de validation, *etc*.) sont préparées, le fichier des spécifications est entré dans le Système de génération de questionnaires électroniques (SGQE), qui traite automatiquement les entrées et produit un QE complet. Le SGQE a été élaboré à l'interne par Statistique Canada. On s'attend à ce qu'il réduise le temps requis pour élaborer et mettre à l'essai chaque application de QE. Par ailleurs, les applications qui sont générées par le SGQE sont aussi conformes aux exigences en matière de confidentialité et de sécurité de Statistique Canada et aux lignes directrices en matière de normalisation des sites Internet (NSI) du gouvernement du Canada, et elles respectent les exigences en matière d'accessibilité. Le SGQE représente la pierre angulaire du nouveau projet de QE.

Par suite de la production de l'application de QE, une mise à l'essai rigoureuse est menée pour s'assurer qu'elle est conforme aux normes de qualité de Statistique Canada. L'étape de mise à l'essai comprend un examen d'assurance de la qualité, un essai d'acceptation par les utilisateurs, un essai de convivialité pour les utilisateurs finaux et une vérification de la conformité aux exigences en matière d'accessibilité et de NSI.

On passe alors à la collecte des données. Un courriel d'invitation est envoyé à chaque répondant. Il comprend une brève introduction à l'enquête et fournit un lien et un code d'accès sécuritaire pour entrer dans le QE. Une fois le lien activé, le répondant peut accéder au portail électronique sécuritaire et est en mesure de remplir son questionnaire en ligne. Lorsque le répondant soumet le questionnaire, les données recueillies sont intégrées à celles obtenues au moyen d'autres modes de collecte dans le système de traitement des données existant.

Enfin, on procède à une évaluation rétrospective, les observations et les leçons apprises servant à améliorer le processus pour les enquêtes ou les cycles suivants.

2.3 Partenaires de collecte

Un processus d'élaboration et de mise en œuvre uniformisé a été créé pour répondre aux besoins de tous les partenaires de collecte. On a pris soin de tenir compte de tous les aspects, comme le contenu du questionnaire, les outils de collecte, les systèmes de traitement, etc. Le calendrier du processus est généralement court et fait intervenir différents secteurs de service, ce qui nécessite une coordination continue des activités entre tous les partenaires concernés et une utilisation rigoureuse des outils de gestion de projet.

À Statistique Canada, les divisions et centres de ressources différents offrent des services spécialisés aux secteurs de programme d'enquête. Pour ce qui est de l'élaboration et de la mise en œuvre des QE, les services spécialisés sont offerts pour la Division des opérations et de l'intégration (DOI), la Division des systèmes et de l'infrastructure de collecte (DSIC), la Division de la planification et de la gestion de la collecte (DPGC), ainsi que le Centre de ressources en conception de questionnaires (CRCQ).

Le projet d'élaboration des QE est coordonné et supervisé par la DPGC. On s'y assure aussi que le nouveau mode de collecte s'intègre bien au processus de collecte existant.

Dans une enquête typique, les secteurs spécialisés fournissent des spécifications à la DOI, qui est responsable de la conception des QE et de la préparation d'un modèle en Excel, qui est utilisé dans le SGQE. La DOI et la DPGC sont aussi responsables de veiller à ce que les spécifications respectent l'ensemble des lignes directrices et normes en matière de QE. Une fois le modèle de QE en Excel prêt, la DOI produit l'application de QE et lance le processus de mise à l'essai.

La DSIC est responsable de l'élaboration et de la mise à jour de l'ensemble de la plateforme de QE, y compris le SGQE. La DSIC s'assure aussi que les fonctions requises pour une enquête donnée sont prêtes et disponibles dans le SGQE.

Les experts de la conception de questionnaires du CRCQ contribuent à l'élaboration du modèle et des spécifications des QE. Le CRCQ assure aussi la coordination et la mise à l'essai des QE auprès des utilisateurs finaux.

3. Lignes directrices et normes pour la conception des QE

3.1 Comité des normes en matière de QE

À Statistique Canada, le Comité des normes en matière de QE a été créé en 2010, en vue d'élaborer et de compiler un ensemble de lignes directrices pour la conception des QE, afin d'assurer une interface commune pour les répondants, l'uniformité de l'affichage et de la présentation à l'écran, l'application des exigences du gouvernement du Canada en matière de normalisation des sites Internet, ainsi que des fonctions et des approches uniformisées.

Le Comité est une équipe multidisciplinaire. Ce groupe comprend des concepteurs de systèmes, des experts de la convivialité, des spécialistes de la conception de questionnaires, des experts de l'accessibilité, des coordonnateurs de collecte des enquêtes auprès des entreprises et des enquêtes sociales et des gestionnaires des opérations.

3.2 Lignes directrices

Les lignes directrices et normes de QE ont pour objectif de fournir une interface commune aux utilisateurs, en réduisant les délais d'élaboration, de mise à l'essai et de formation, ainsi que les coûts.

La première version des lignes directrices et normes de conception de Statistique Canada est disponible à l'interne depuis mai 2011. Les normes ont été formulées à partir de documents et de la littérature externes, de pratiques existantes à Statistique Canada et de constatations et d'observations découlant de la mise à l'essai auprès des utilisateurs finaux. Jusqu'à maintenant, un ensemble unique de lignes directrices et de normes a été établi, tant pour les enquêtes auprès des entreprises que pour les enquêtes sociales.

Les normes décrivent comment les QE devraient être conçus et présentés. Pour le moment, elles englobent les caractéristiques de conception des enquêtes sur le Web, comme la navigation, la convivialité, l'accessibilité et les affichages à l'écran. Les composantes traditionnelles de la conception des questionnaires, comme le libellé ou l'ordonnancement du contenu, n'ont pas encore été abordées. Les normes continueront d'évoluer, par suite de la mise à l'essai des QE, des résultats de la recherche et des résultats de la collecte de données en cours. Elles seront aussi élargies, afin d'inclure les nouvelles exigences d'enquête. Le Comité des normes se penche actuellement sur des aspects plus complexes, comme l'harmonisation du contenu commun, qui doivent être incluses dans les lignes directrices.

3.3 Éléments inclus

Le Comité a examiné et établi des normes pour les éléments suivants :

- éléments de la NSI du gouvernement du Canada et conformité ;
- interface réservée à l'enquête ;
- uniformité des barres d'outils ;
- fonction et positionnement des touches et boutons de navigation ;
- catégories de questions et positionnement ;
- instructions ;
- catégories de réponses, positionnement et sélection ;
- enchaînement des questions ;
- polices de caractère et styles ;
- utilisation uniforme des couleurs et du noir et blanc ;
- restriction du nombre de caractères ;
- fonctions pour les valeurs négatives, les décimales et les milliers ;
- contenu générique pour la page d'introduction et les données de contact ;
- affichage de messages (contrôles, mises en garde) ;
- fonction de verrouillage automatique après un délai d'inactivité ;
- navigation et impression.

4. État d'avancement du projet

Jusqu'à maintenant, la collecte au moyen des QE a été effectuée ou est actuellement en cours pour 16 enquêtes auprès des entreprises, deux enquêtes agricoles et deux enquêtes sociales auprès des établissements. Environ 10 autres enquêtes devraient être en production d'ici le 31 mars 2012.

On procède aussi à une mise à l'essai continue pour les enquêtes sociales. Certains modules choisis de l'Enquête sur la population active (EPA) ont été mis à l'essai auprès d'anciens répondants à l'EPA. Une version électronique intégrée du questionnaire de l'Enquête sociale générale a aussi fait l'objet d'un essai qualitatif auprès des répondants potentiels.

L'objectif de Statistique Canada est de mener jusqu'à 200 enquêtes auprès des entreprises et des ménages au moyen des QE comme mode de collecte, d'ici le 31 mars 2015.

5. Défis

5.1 Mise en œuvre d'une nouvelle méthodologie

La mise en œuvre d'une nouvelle méthodologie, comme la collecte des données au moyen de QE, a suscité de nombreux défis, et tous les partenaires de collecte ont dû s'adapter. Par exemple, de nombreuses enquêtes auprès des entreprises étaient depuis toujours administrées au moyen de questionnaires papier. Les secteurs spécialisés ont dû comprendre qu'un QE ne peut pas toujours être une réplique électronique d'un instrument sur papier. Dans ce cas,

des mesures ont été prises pour veiller à ce que les données recueillies au moyen de QE puissent être intégrées dans les systèmes existants de collecte et de traitement et soient comparables au niveau analytique avec celles recueillies au moyen d'autres modes. Des efforts spéciaux ont été déployés pour tenter d'atténuer les effets de mode possibles.

5.2 Communication

Le maintien d'une ligne de communication constante et continue représentait un défi, du fait qu'il y a de nombreux collaborateurs et que les lignes directrices en matière de conception sont nouvelles et en constante évolution. Le processus opérationnel lié à l'élaboration et à la mise en œuvre était nouveau, et de nombreux secteurs spécialisés participaient à un tel processus pour la première fois.

5.3 Contenu du questionnaire

Parmi les autres défis liés à la nouvelle méthode de collecte figure la nécessité d'uniformiser certains éléments pour le contenu du questionnaire. Même si une politique de révision et de mise en essai des questionnaires est en place à Statistique Canada depuis presque 20 ans, il n'existe jusqu'à maintenant aucune norme de conception documentée pour les instruments sur papier des enquêtes auprès des entreprises. Au fil du temps, chaque questionnaire papier a été « personnalisé ». Au cours du processus de spécification du QE, il faut tenir compte d'un nombre plus grand d'éléments que dans le cas du questionnaire papier. Une gamme variée d'éléments peuvent affecter l'aspect visuel du QE, comme le libellé et le positionnement des questions et des instructions, la numérotation des questions, les en-têtes de section et les messages de contrôle. Ces caractéristiques doivent être bien définies. Pour améliorer le processus d'élaboration des QE, l'examen du contenu doit être rigoureux et les experts de la conception de questionnaires doivent participer tôt au processus, idéalement à l'étape des spécifications d'entrée.

5.4 Planification et mise en œuvre des essais qualitatifs

La planification et la mise en œuvre des essais qualitatifs des QE présentaient des défis. À Statistique Canada, les essais qualitatifs des questionnaires papier faisaient intervenir uniquement le secteur spécialisé et les experts de la conception de questionnaires (CRCQ). Dans le cas des QE, la mise en œuvre des essais qualitatifs prend plus de temps, fait intervenir un plus grand nombre de partenaires de collecte et exige une coordination soigneuse des activités.

Jusqu'à maintenant, des contraintes de temps ont limité la capacité de mettre à l'essai pleinement le QE dans un environnement réel à partir de l'ordinateur du répondant. Les essais qualitatifs ont principalement pris la forme d'une mise en œuvre d'une version de mise à l'essai du QE sur les ordinateurs portables de Statistique Canada. Jusqu'à maintenant, cela a limité la capacité d'examiner tous les aspects de l'expérience des QE, comme la réception d'une invitation par courriel et l'entrée dans le portail du QE, au moyen de réglages propres à l'utilisateur établis par les répondants sur leur ordinateur.

5.5 Changements techniques et technologiques

Du fait de l'augmentation du nombre d'enquêtes pour lesquelles on prévoit utiliser le QE comme mode de collecte et comme chaque enquête porte sur un contenu spécialisé différent, des nouvelles fonctions doivent être élaborées pour répondre aux différents besoins des enquêtes. Par ailleurs, compte tenu du nombre important d'applications et d'enquêtes que l'on retrouve sur Internet, les répondants savent de plus en plus ce qui est disponible du point de vue des fonctions et de la présentation visuelle. Afin de maintenir les taux de réponse et de réduire le fardeau de réponse, Statistique Canada doit se tenir au fait de ces innovations lorsqu'il offre une option de réponse en ligne.

La rapidité avec laquelle les nouvelles technologies voient le jour et leur évolution constante représentent aussi un défi pour Statistique Canada. Par exemple, la diffusion d'une nouvelle version d'un navigateur commun mène à un plus grand nombre d'essais et de rajustements possibles de la plateforme de QE, l'organisme devant s'assurer que la nouvelle version respecte les exigences en matière d'accessibilité des sites Web du gouvernement du Canada.

6. À venir

Le processus de conception et d'élaboration des QE continue d'évoluer, et tous les partenaires de collecte continuent d'apprendre et de s'adapter. Les résultats des essais qualitatifs et de la collecte des données doivent être intégrés dans les lignes directrices et les normes en matière de QE et, un jour, les aspects traditionnels de la conception des questions seront aussi intégrés.

Afin de réaliser des économies, les experts de la conception de questionnaires participeront aux premières étapes des processus liés au contenu, aux spécifications et à la production des QE. Les experts de la conception de questionnaires cherchent aussi des façons d'améliorer les stratégies et les compétences en matière d'essais préliminaires. Différentes approches reposant sur des interviews cognitives pour la collecte d'information, comme les paradonnées ou les nouveaux outils technologiques (par exemple, les systèmes de poursuite oculaire), sont envisagées.

Les membres de l'équipe d'élaboration des QE de Statistique Canada se tiennent au fait des technologies nouvelles et émergentes. Il faut notamment s'assurer que le SGQE est compatible avec les différents navigateurs et systèmes d'exploitation. L'équipe évalue aussi la faisabilité d'une application de collecte pour appareils mobiles à Statistique Canada et les gains pouvant en découler.

L'utilisation de la collecte par QE pour les enquêtes sociales est toujours en voie d'élaboration. Les experts des questionnaires, les partenaires de collecte et diverses équipes spécialisées explorent différentes méthodes de transition des questionnaires administrés par des intervieweurs à des instruments à remplir soi-même. L'avènement du mode de collecte par QE pour les enquêtes sociales a aussi fait ressortir la nécessité d'étudier la faisabilité d'utiliser des données déjà intégrées recueillies dans des interviews précédentes. Les problèmes liés à la confidentialité et aux réactions des répondants seront évalués.

Enfin, tous les partenaires de collecte de Statistique Canada sont tenus au courant de la recherche et des progrès en cours par leurs collègues internationaux.

Bibliographie

Karaganis, M. (2011), « Development and Implementation of E-questionnaire as a Primary Collection Mode at Statistics Canada », document interne non publié, Ottawa, Canada, Statistique Canada.

Lawrence, D. (2011), « Developing Electronic Questionnaires at Statistics Canada: Experiences and Challenges in a Changing Environment », document présenté dans le cadre du Internet Survey Methodology Workshop, La Haye, Pays-Bas.

Statistique Canada (2011) « Normes et lignes directrices en matière de conception de questionnaires électroniques », document interne non publié, Ottawa, Canada.

GINO++, un système généralisé pour les enquêtes en ligne

Renato Torelli[1]

Résumé

Le système généralisé, appelé GINO++, permet au chercheur de : concevoir et de mettre en œuvre des questionnaires en ligne de façon indépendante et rapide, grâce à une interface graphique; d'accompagner les questions de trucs et de conseils concernant l'outil; d'inclure des vérifications dans les données entrées; de contrôler les progrès de l'enquête en temps réel.

L'utilisateur final (répondant à l'enquête) peut : remplir le questionnaire en plusieurs séances, c'est-à-dire en le sauvegardant chaque fois et en l'envoyant uniquement à la fin de la compilation; imprimer localement (html ou excel) le questionnaire; insérer des notes pour chaque variable individuelle (à condition que cela soit prévu par le chercheur au moment de la conception).

Chaque fois que le répondant sauvegarde les données entrées, celles-ci sont directement insérées dans des tableaux d'une base de données en Oracle produite par l'application, à partir d'un nom défini par l'utilisateur.

GINO++ comporte deux domaines principaux : gestion et conception de questionnaires et de métadonnées; surveillance des enquêtes.

Mots clés : Enquêtes en ligne ; logiciel généralisé ; conception de questionnaires ; surveillance des enquêtes.

1. Introduction

1.1 Description

GINO++ [*bien plus que recueillir de l'information en ligne*] permet au statisticien d'exécuter lui-même (ce qui signifie sans développeurs de logiciel ou informaticiens) trois étapes d'une enquête : conception, saisie et surveillance.

À l'« étape de la conception », le statisticien peut concevoir un questionnaire, mais peut aussi le modifier, juste avant l'enquête (et en améliorer la disposition par la suite), ainsi qu'insérer des règles pour vérifier les valeurs entrées.

À l'« étape de la saisie », il peut obtenir des données en ligne et les intégrer directement dans une base de données. Cela permet par la suite de visionner chaque questionnaire, même partiellement rempli, exactement comme il figure dans le compilateur (après une commande « Sauvegarder » ou « Envoyer »). Il peut aussi exporter des données qui viennent d'être entrées en format excel, par exemple, et voir les séries au moyen de diagrammes (à condition que des séries chronologiques soient prévues).

À l'« étape de la surveillance », le chercheur peut constamment surveiller, étape par étape, les activités des répondants et des superviseurs, par exemple, le nombre d'accès et l'état d'avancement des questionnaires. Par ailleurs, il peut intervenir auprès des retardataires immédiatement, à toutes les étapes de l'enquête : enregistrement initial de l'utilisateur, transmission provisoire, transmission finale. Enfin, il peut procéder à une analyse de la qualité à partir de rapports détaillés au sujet des erreurs.

[1]Renato Torelli, Istat, via Cesare Balbo 16, Rome, Italie, 00184, torelli@istat.it.

1.2 Étapes d'une enquête

Dans le cycle de vie d'une enquête, GINO++ se situe après la conception de l'enquête et avant les étapes de la vérification et de l'imputation, ainsi que de la diffusion des données. Il porte par conséquent sur les étapes intermédiaires d'une enquête : conception du questionnaire et saisie des données.

Il est important de noter que l'utilisation de ce système permet d'obtenir deux résultats : décharger les développeurs du fardeau de mettre en œuvre des questionnaires ponctuels et donner de l'autonomie aux chercheurs.

1.3 Évolution

À Istat, nous avons entrepris le développement de la première version du système en 2008, et celui-ci a été utilisé pour la première fois dans le cadre de l'enquête sur l'« *environnement urbain en Italie* ». Année après année, nous avons pu améliorer le système, de nouvelles fonctions ayant été mises en œuvre, ainsi que de nouvelles enquêtes, sous forme de cas types.

Jusqu'à maintenant, nous avons traité trois principaux domaines d'intérêt : statistiques sur l'environnement, statistiques sociales et statistiques économiques.

2. Architecture

2.1 Composantes technologiques

Le système comprend deux principales composantes technologiques : PHP représente le côté serveur du langage de programmation sur le Web, et Oracle, le système de gestion de bases de données relationnelles.

La base de données est constituée de deux ensembles de tableaux : les tableaux de données et les tableaux de métadonnées. Comme cela est habituellement le cas, les tableaux de données sont généralement peu nombreux et gros, tandis que les tableaux de métadonnées sont nombreux et petits.

Par ailleurs, les tableaux de données se divisent de la façon suivante :

- *Données de source* : il s'agit d'un tableau statique qui peut comprendre des données initiales pour la personnalisation du questionnaire

et

- *Données entrantes* : il s'agit d'un tableau dynamique qui comprend les données fournies par les répondants.

2.2 Configurations

La structure de gestion la plus simple d'une enquête appuyée par GINO++ comprend deux niveaux : le premier est celui du propriétaire, c'est-à-dire le responsable de l'enquête, et le deuxième, celui du répondant; les répondants pouvant être des personnes, des ménages, des organisations, des institutions, des compagnies, *etc*.

Une structure plus complexe d'enquête appuyée par GINO++ comprend aussi deux niveaux (propriétaire et répondant), mais chaque répondant est interviewé au sujet de nombreux thèmes (c'est-à-dire de nombreux questionnaires pour la même enquête). Dans cette configuration, deux rôles coexistent au niveau du répondant : un seul *responsable thématique* et plusieurs *répondants thématiques*.

Une troisième structure possible comprend trois niveaux et diffère des précédentes en raison de la présence d'un « organisme intermédiaire ».

La dernière configuration peut comporter jusqu'à quatre niveaux de surveillance.

2.3 Gestion des utilisateurs

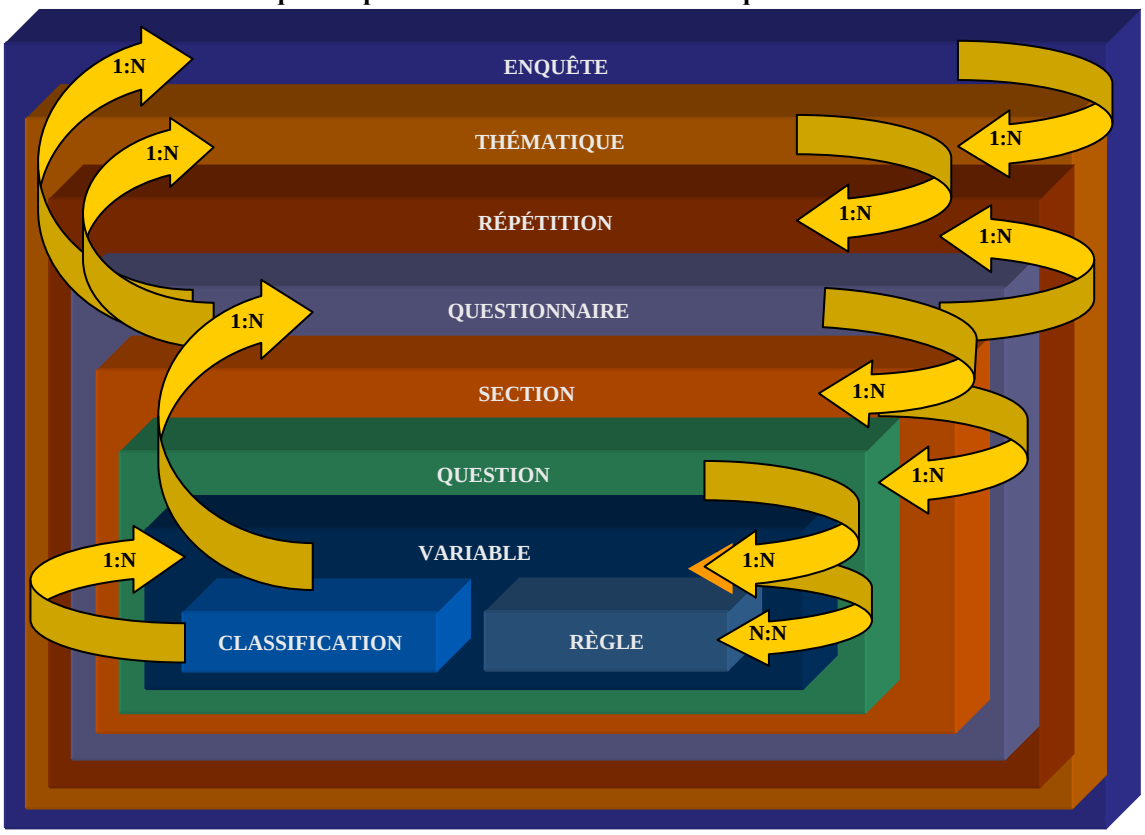
Les permissions de lecture et d'écriture sont établies de façon dynamique par le responsable de l'enquête pour chaque rôle, selon chaque *état d'avancement* du questionnaire (*non traité, en traitement, soumis, validé* et *achevé*).

La concurrence entre les utilisateurs est gérée. Ainsi, lorsque le premier utilisateur ayant une permission d'écriture entre dans le questionnaire, tous les autres utilisateurs (de même qu'une autre session ouverte par le même utilisateur) perdent leur permission d'écriture.

2.4 Composantes de base

Les éléments de base de l'enquête sont les suivants : *thématique*, *répétition*, *questionnaire*, *section*, *question*, *variable*, *règle* et *classification*.
Comme le montre la figure 2.4-1, une enquête peut porter sur plusieurs *thématiques*, les résultats étant le fait de plusieurs questionnaires thématiques.
Une thématique peut faire l'objet de plusieurs *répétitions* (c'est-à-dire dans différentes versions de la même enquête).
Un *questionnaire* peut être utilisé pour plusieurs enquêtes, thématiques et répétitions. Il est en outre composé de plusieurs sections.
Une *section* est constituée de plusieurs questions, à moins qu'il ne s'agisse seulement d'un en-tête.
Une *question* est composée de plusieurs variables.
La même *variable* peut apparaître dans plusieurs questionnaires. Par ailleurs, une variable peut figurer dans plusieurs *règles* et une règle peut être utilisée pour plusieurs variables.
Enfin, une *classification* peut être utilisée pour plusieurs variables.

Figure 2.4-1
Éléments de base de l'enquête représentés comme des cases imbriquées



2.5 Fonctions

Le statisticien utilise deux principaux domaines dans le système : un pour la gestion des métadonnées et un pour la surveillance de l'enquête.

Le premier domaine porte sur la définition de l'enquête et sur la conception du questionnaire.

Le deuxième a trait à la surveillance de l'enregistrement des utilisateurs, de l'état d'avancement des questionnaires et du processus d'enquête en général.

3. Gestion des métadonnées

3.1 Description

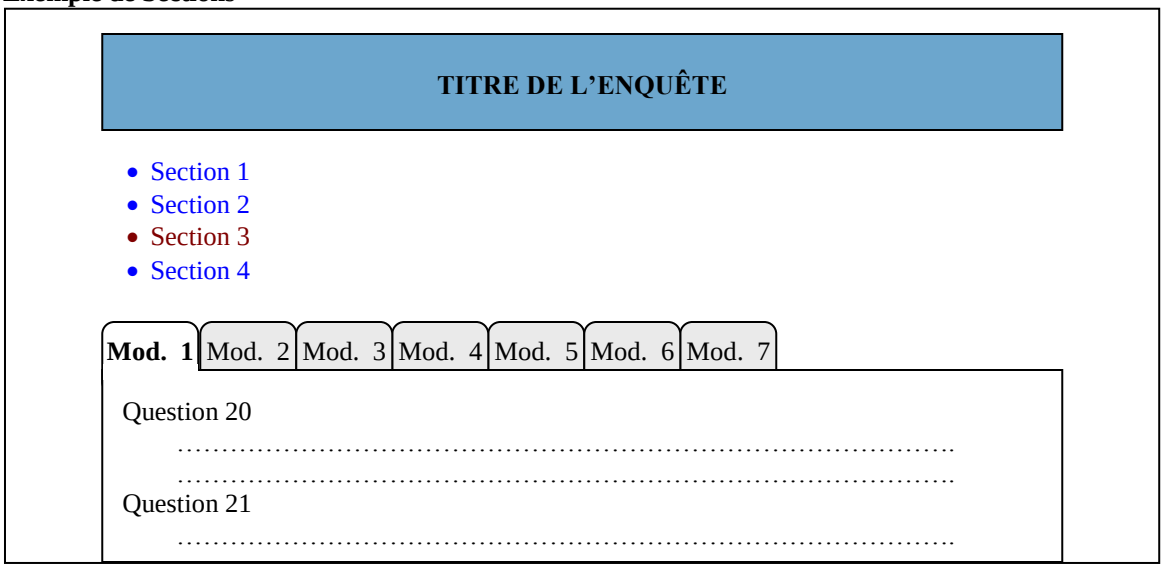
La gestion des métadonnées permet de créer, de mettre à jour et de supprimer :
- les fonctions d'une enquête, y compris les thématiques, répétitions, états d'avancement et utilisateurs ;
- le domaine complet des variables et des classifications ;
- le questionnaire organisé en sections, en questions et en règles.

3.2 Sections

Voici les types possibles de sections :
- en-tête/pas d'en-tête ;
- paginée/non paginée ;
- modulaire/non modulaire.

Dans la Figure 3.2-1
- Le titre de l'enquête avec l'arrière-plan bleu ciel a été obtenu au moyen d'une section de type en-tête ;
- Les puces, en caractères bleus, permettent d'accéder à une section par page ;
- En bas de la section 3 se trouvent sept modules. Un module est un bloc de questions qui se répète, comme pour les membres d'un ménage ou les sections d'une entreprise.

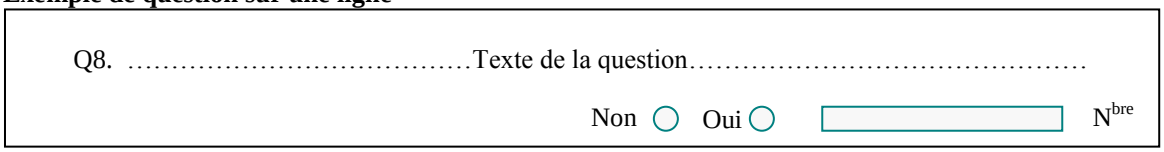
Figure 3.2-1
Exemple de Sections



3.3 Questions

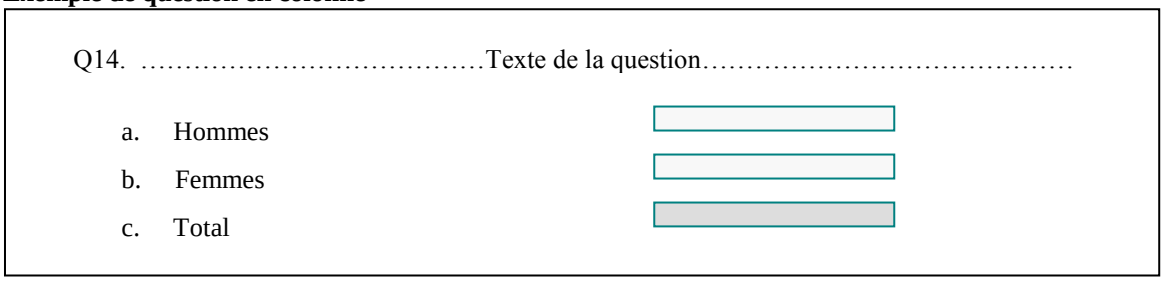
Divers formats sont disponibles pour les questions :
Le premier type est la question sur une « *ligne* » (figure 3.3-1), c'est-à-dire que toutes les variables liées à la question sont placées horizontalement.

Figure 3.3-1
Exemple de question sur une ligne



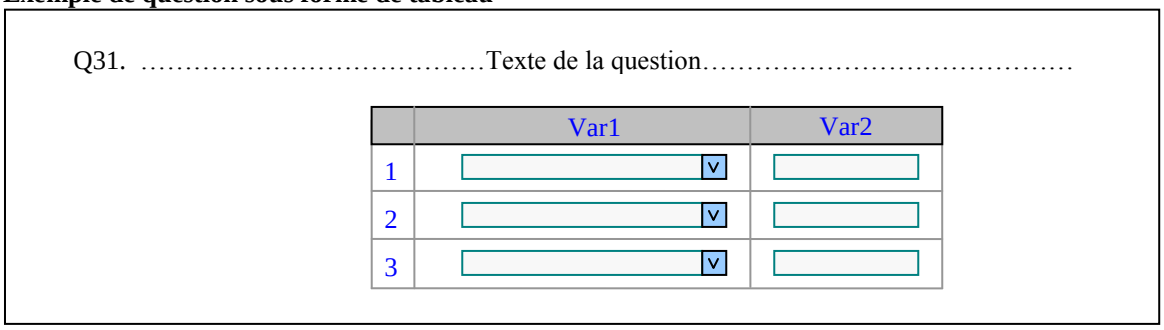
Un deuxième type est la question en « *colonne* » (figure 3.3-2), c'est-à-dire que toutes les variables associées à la question sont placées verticalement.

Figure 3.3-2
Exemple de question en colonne



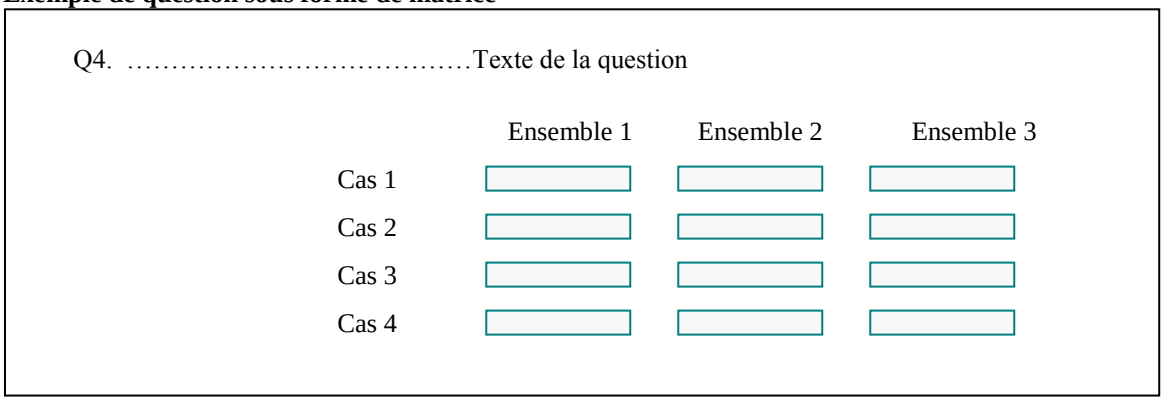
La question en « *tableau* » (figure 3.3-3) représente une ligne de variables reprise un certain nombre de fois, qui n'est parfois même pas fixé au départ.

Figure 3.3-3
Exemple de question sous forme de tableau



Le quatrième type de question est la question sous forme de « *matrice* » (figure 3.3-4), qui représente une façon d'organiser m prédéterminé multiplié par n variables.

Figure 3.3-4
Exemple de question sous forme de matrice



Certaines options sont disponibles pour les questions :
- trucs : texte additionnel, en vue d'explorer/de compléter certains concepts ;
- conseil concernant l'outil : lorsque le curseur est sur une question, un message apparaît ;
- sources : permet au compilateur de décrire la source des données pour cette question ;
- notes en encadré : carré avec arrière-plan coloré servant à informer le compilateur de certaines notes ;
- possibilité de numéroter automatiquement les questions au moyen de chiffres ou de lettres.

3.4 Variables

Les variables peuvent avoir les formats suivants :
- case à cocher pour les réponses à choix multiples ;
- bouton radio et menu déroulant pour les réponses à choix unique ;
- champ de texte, champ numérique et champ de date.

Par ailleurs, les variables peuvent comporter les options suivantes :
- des notes de champ, qui permettent à l'utilisateur d'expliquer sa réponse au moyen d'un texte. Si quelque chose est écrit dans le champ, l'icône change et un crayon apparaît sur une image de feuille de papier ;
- possibilité de consulter des séries chronologiques de variables ;
- mode lecture seulement ;
- numérotation automatique ou manuelle.

3.5 Règles

Les règles associées à chaque variable peuvent être des types suivants :
- activation/désactivation d'un champ ;
- saut de sections/questions/variables ;
- vérification de la cohérence (par exemple, A+B doit être inférieur à C) ;
- variables calculées (par exemple, A doit être égal à B+C) ;
- variable filtre sur la valeur d'une autre variable (par exemple, « liste des provinces » contrôlant la « liste des villes »).

Les règles peuvent aussi comporter certaines options :
- une règle peut être activée ou désactivée pendant l'enquête ;
- une règle peut entraîner le groupage ou le non-groupage pour l'envoi du questionnaire.

4. Surveillance

4.1 Description

Pour que l'activité de surveillance soit efficace, plusieurs formulaires de rapport sont disponibles. Cinq d'entre eux sont énumérés ci-après.

4.2 Premier formulaire

Dans ce formulaire (figure 4.2-1), pour chaque répondant, on montre :
- certains détails, afin de mieux l'identifier et le sélectionner/le filtrer ;
- l'état d'avancement du questionnaire (par exemple, en traitement, envoyé, *etc*.) ;
- une icône de roue dentée pour modifier l'état d'avancement et son report en avant ou en arrière ;
- une icône de cadenas pour indiquer que le questionnaire est actuellement en compilation (cadenas fermé). Cette icône est aussi utilisée pour déverrouiller les sessions qui ont été bloquées en raison d'une fermeture soudaine du navigateur sans sortie proprement dite de l'application ;
- un sémaphore pour indiquer le degré de violation des règles.

Figure 4.2-1
Premier formulaire de surveillance

Répondant	Données territoriales	État d'avance-ment	Dernière modification	Dernier utilisateur	Modification de l'état d'avancement	Utilisateurs concernés	Questionnaire	En compilation	Degré de violation
Répondant 1	Rome	Envoyé	20/07/11	Utilisateur 3					
Répondant 2	Milan	En traitement	14/10/11	Utilisateur 8					
Répondant 3	Florence	En traitement	03/09/11	Utilisateur 53					

4.3 Deuxième formulaire

Un deuxième formulaire montre, pour chaque type d'utilisateur, des données d'identification et la façon de communiquer avec lui. De façon plus particulière, on montre :
- le répondant ;
- le type d'utilisateur (répondant thématique, responsable thématique, organisme intermédiaire) ;
- le nom de l'utilisateur ;
- la date d'enregistrement ;
- le numéro de téléphone ;
- l'adresse de courriel ;
- toute affiliation.

4.4 Troisième formulaire

Un troisième formulaire montre l'état d'avancement de l'enquête et fournit la valeur absolue et le pourcentage d'un questionnaire pour chaque étape de traitement (non traité, en traitement, soumis, validé et achevé) et à différents niveaux territoriaux.

4.5 Quatrième formulaire

Un quatrième formulaire donne une indication de la qualité des questionnaires en fournissant une valeur absolue et un pourcentage de champs remplis à toutes les étapes du traitement. Évidemment, ces renseignements ne sont significatifs que dans le cas des questionnaires non personnalisés (c'est-à-dire lorsque certaines des questions ne sont pas déjà compilées au préalable).

4.6 Cinquième formulaire

Le cinquième et dernier formulaire fournit deux renseignements sommaires à différents niveaux territoriaux et pour chaque thématique :

- un rapport sur le nombre de répondants thématiques enregistrés et le nombre de ceux qui ne le sont pas ;
- un rapport sur le nombre de questionnaires *non traités, en traitement, soumis, validés* et *achevés*.

5. Autres fonctions

5.1 Chargement/Téléchargement

Une fonction utile pour les répondants (ou organismes intermédiaires) qui disposent déjà des données requises en format électronique, est la possibilité de charger les données à différents niveaux :

- l'ensemble du questionnaire ;
- des sections individuelles ;
- des questions individuelles.

Il convient de souligner que les mêmes contrôles sont possibles à partir de la compilation en ligne.
Le téléchargement est aussi autorisé.

5.2 Analyse de la qualité

Une autre caractéristique est l'analyse de la qualité du questionnaire, tant du « *côté répondant* » que du « *côté conception* » :

- la qualité du « côté répondant » représente la qualité des réponses, c'est-à-dire, le nombre d'erreurs au moment de la dernière sauvegarde multiplié par le type d'erreur (grave, intermédiaire, légère – partitionnement de l'ensemble de règles défini par l'utilisateur) ;
- la qualité du « côté conception » signifie l'analyse de la qualité des questions, des enchaînements, *etc*. et est le résultat du dénombrement des erreurs pour toutes les sauvegardes.

5.3 Divers

Enfin, la dernière version de l'application permet :

- la gestion de langues multiples ;
- ce que l'on appelle les questionnaires multiples, c'est-à-dire, la création dynamique d'un plus grand nombre de questionnaires pour chaque utilisateur, chaque fois qu'un certain événement a lieu ;
- les rappels aux gestionnaires, grâce à la production de listes d'adresses ou par l'envoi direct de courriels.

5.4 Points forts

Les points forts du système sont les suivants :

- un grand niveau de stabilité (quatre versions et plusieurs enquêtes l'utilisent) ;
- un vaste ensemble de fonctions ;
- l'uniformité de la présentation pour les utilisateurs externes et la même interface d'application pour les utilisateurs internes ;
- la réduction des coûts.

Remerciements

J’ai eu la chance de rencontrer un grand nombre de personnes intelligentes qui ont collaboré à ce projet, et je ne peux malheureusement pas toutes les nommer. J’aimerais simplement remercier Corrado Carmelo Abbate, qui m’a permis d’entreprendre ces travaux. Ceux-ci n’auraient pas pu aller aussi loin sans l’appui de Linda Laura Sabbadini et Saverio Gazzelloni, qui croyaient fermement en leur utilité. Heureusement, Silvia Montagna a eu le courage de faire l’expérience d’un logiciel encore jeune, de suggérer des améliorations et de le proposer à d’autres. Des remerciements particuliers vont à Teresa Di Sarro, pour ses efforts constants et sa grande capacité de penser en termes de concepts généraux, ainsi que pour l’exactitude de l’étape d’essai. Enfin, de nombreux remerciements vont à Angela Ciocci, qui a contribué de façon significative au projet, grâce à son expertise technologique.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Expérience intégrée sur des méthodes de suivi des cas de non-réponse visant la collecte de données au moyen d'un questionnaire électronique

Milana Karaganis, Karla Fox, Jeannine Claveau, Joanne Leung et Wei Lin[1]

Résumé

À l'heure actuelle, Statistique Canada entreprend une refonte générale de ses programmes de production de statistiques sur les entreprises. L'un des objectifs est de faire de la collecte électronique des données le mode principal de collecte pour les enquêtes-entreprises. Jusqu'à maintenant, les méthodes de suivi appliquées aux questionnaires électroniques étaient fondées sur les stratégies servant aux méthodes de collecte sur support papier (rappels par télécopieur et/ou par téléphone). Afin d'établir une stratégie de suivi uniforme de la collecte pour les enquêtes annuelles auprès des entreprises au moyen des questionnaires électroniques comme principal outil de collecte, Statistique Canada a conçu un modèle expérimental permettant de comparer différentes méthodes de suivi de la non réponse. Les auteurs résument les premiers résultats de cette expérience.

Mots clés : Collecte ; non-réponse ; suivi ; paradonnées ; modèle expérimental.

1. Introduction

À Statistique Canada, la collecte des données des enquêtes-entreprises comporte de nombreuses étapes et utilise plus d'un mode de collecte. Dans de nombreuses enquêtes auprès des entreprises, on continue d'utiliser des questionnaires papier pour la collecte des données. Les progrès récents des technologies Internet ont eu des répercussions considérables sur la collecte des données d'enquête, l'utilisation de questionnaires électroniques (QE) pour la collecte des données ayant connu un essor considérable au cours des dix dernières années. Les enquêtes au moyen de QE peuvent prendre une gamme variée de formes, de la simple enquête par courriel à des systèmes d'enquête perfectionnés sur le Web. À l'heure actuelle, Statistique Canada entreprend une refonte générale de ses programmes de production de statistiques sur les entreprises. L'un des objectifs est de faire de la collecte électronique des données le mode principal de collecte pour les enquêtes-entreprises.

Jusqu'à maintenant, les méthodes de suivi utilisées pour les enquêtes au moyen de QE à Statistique Canada étaient fondées sur les méthodes servant à la collecte sur support papier, soit une combinaison de tentatives de prise de contact par télécopieur et par téléphone. Des expériences menées dans d'autres pays et à Statistique Canada ont montré que les profils de suivi des répondants qui fournissent des données électroniques diffèrent des profils observés autrement. Afin d'établir une stratégie uniforme de suivi de la collecte pour les enquêtes annuelles auprès des entreprises au moyen de questionnaires électroniques comme principal outil de collecte, Statistique Canada a élaboré un modèle expérimental pour comparer les différentes méthodes de suivi des cas de non réponse (SCNR), combinant des rappels par téléphone et par courriel à différents moments tout au long de la période de collecte. Sept enquêtes relevant du Programme unifié des statistiques sur les entreprises (PUSE), pour le cycle de collecte de 2011, ont servi à cette expérience. Un plan factoriel équilibré intégré a été utilisé pour réaliser cette expérience. La tenue d'une expérience intégrée visait à trouver une stratégie de suivi produisant les meilleurs taux de réponse et étant la plus efficace du point de vue des coûts. Nous voulions aussi savoir jusqu'où nous pourrions aller en envoyant des rappels par courriel seulement, et déterminer s'il était important d'effectuer le premier suivi par téléphone plutôt que par courriel.

[1]Milana Karaganis, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (milana.karaganis@statcan.gc.ca) ; Karla Fox, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (karla.fox@statcan.gc.ca) ; Jeannine Claveau, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (jeannine.claveau@statcan.gc.ca) ; Joanne Leung, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (joanne.leung@statcan.gc.ca) ; Wei Lin, Université de Toronto, Ontario, Canada (wei.lin@utoronto.ca).

Le présent document résume les résultats initiaux de cette expérience. La section 2 donne un aperçu de la collecte des enquêtes du PUSE. La section 3 explique la méthodologie du modèle expérimental. La section 4 présente les résultats de l'expérience, y compris les résultats de l'analyse des tests de variance. La section 5 résume les conclusions et la section 6 fournit des recommandations pour les enquêtes futures au moyen de QE.

2. Aperçu de la collecte des données des enquêtes du PUSE

Le PUSE comprend près de 60 enquêtes annuelles auprès des entreprises, qui sont intégrées du point de vue du contenu, de la collecte et du traitement des données. En 2011, des questionnaires électroniques (QE) sur le Web ont été élaborés et utilisés pour la collecte des données des sept enquêtes du PUSE. Avant 2011, seulement une enquête de ce programme utilisait le questionnaire électronique comme mode de collecte.

Le processus de collecte du PUSE se déroule en deux étapes. La première étape du cycle de collecte annuelle des enquêtes du PUSE a pris la forme d'un contact préliminaire par téléphone auprès des nouvelles entreprises sélectionnées dans l'échantillon, afin de confirmer leurs coordonnées, ainsi que leurs codes d'activité, selon le Système de classification des industries de l'Amérique du Nord (SCIAN). En 2011, pour sept enquêtes au moyen de QE, un contact préliminaire par téléphone a été effectué, non seulement pour confirmer les données existantes, mais aussi pour obtenir les adresses de courriel des répondants. Les répondants ont été informés que la collecte serait effectuée par questionnaire électronique et ont dû fournir leur adresse de courriel. Les seuls répondants pour lesquels on a utilisé des questionnaires papier étaient ceux qui avaient refusé catégoriquement d'utiliser le QE, ou ceux que l'on n'avait pas pu joindre lors de la prise de contact préliminaire. Dans le cas des unités qui avaient déjà fait l'expérience de la collecte au moyen de QE, un courriel de rappel leur a plutôt été envoyé pour les informer que leurs réponses seraient encore une fois recueillies au moyen d'un QE. Dans l'ensemble, l'échantillon de 2011 des sept enquêtes du PUSE utilisant le mode de collecte du QE s'est établi à 9 324 unités, et 6 457 unités (environ 70 %) ont été affectées au mode de collecte au moyen de QE après l'étape de la prise de contact préliminaire.

L'étape suivante de la collecte a consisté à envoyer des questionnaires papier par la poste ou des invitations par courriel aux unités échantillonnées. L'étape finale de la collecte des données pour les enquêtes du PUSE a consisté à recevoir les questionnaires remplis, à mettre en image les questionnaires papier, à télécharger les questionnaires mis en image et électroniques dans le système central de collecte (Blaise) pour les vérifications de contrôle, à assurer le suivi auprès des non répondants (SCNR) et des répondants dont les questionnaires avaient été rejetés au contrôle dans Blaise (SQRC pour suivi des questionnaires rejetés au contrôle), à mettre la dernière main aux cas dans Blaise et à envoyer le produit à la division spécialisée pour poursuivre le traitement.

Pour les sept enquêtes au moyen de QE en 2011, le suivi des unités utilisant le QE qui n'avaient pas soumis de questionnaires remplis a fait l'objet d'une expérience. Les unités répondant au moyen d'un questionnaire papier ont fait l'objet de la stratégie habituelle de SCNR (suivi par téléphone et/ou rappels par télécopieur). Pour l'ensemble du SCNR par téléphone, peu importe le mode de collecte, un maximum de cinq tentatives a été déterminé avant qu'une unité soit définitivement considérée comme une unité non répondante.

3. Méthodologie du modèle expérimental

3.1 Modèle intégré

Les chercheurs conviennent qu'il est important pour les méthodologistes d'enquête d'étudier comment les différentes méthodologies d'enquête et stratégies de mise en œuvre affectent la non réponse, la qualité et l'efficience (Van den Brackel et Rensenn (2005), Jackle et coll. (2010), Groves (2010)). Des essais contrôlés de façon aléatoire constituent la façon la plus courante de tenter de déterminer s'il existe un rapport de cause à effet entre une intervention particulière et un résultat. D'autres modèles d'étude peuvent permettre de déterminer les associations entre une intervention et un résultat. Toutefois, elles ne peuvent éliminer la possibilité que l'association soit causée par un troisième facteur, lié à la fois à l'intervention et au résultat. Même si les essais aléatoires sont des outils puissants, leur utilisation est souvent limitée par des préoccupations éthiques et pratiques. Dans le cas du PUSE, il n'est pas faisable, en raison du coût, de tenir une expérience distincte de l'enquête pour étudier les différentes stratégies de

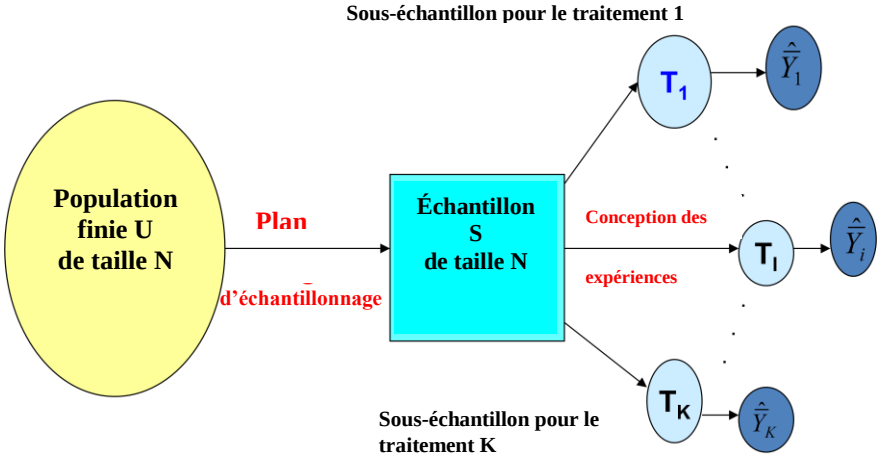
suivi des cas de non réponse. Toutefois, des ouvrages publiés récemment montrent que les expériences intégrées dans des enquêtes sur échantillon permanentes sont particulièrement appropriées pour évaluer les effets des diverses méthodes d'enquête sur le comportement des répondants ou les estimations de la population finie cible (Van den Brackel et Renseen, 2005, Van den Brackel et Berkel, 2002). Même si on s'apercevait que les approches de rechange ont des effets importants, nous pourrions toujours utiliser les données recueillies au moyen de l'approche d'enquête courante pour la publication (c'est-à-dire l'estimation) et utiliser l'ensemble des données pour vérifier les différences de traitement (c'est-à-dire l'inférence).

Une expérience à l'intérieur d'une enquête peut être perçue comme une variante d'un plan d'enquête à deux phases, comme l'illustre la figure 3.1 1 ci après. Ainsi, pour vérifier si les traitements diffèrent de façon significative, nous devons utiliser des méthodes qui tiennent compte de l'étape de l'expérience et de l'étape de l'échantillonnage (Van den Brackel et Renssen, 1998).

Pour notre expérience, étant donné qu'il n'existait pas d'estimation précédente de la grandeur de l'effet, nous avons utilisé un modèle factoriel équilibré intégré. Les traitements ont été répartis de façon équilibrée et aléatoire entre les strates d'enquête. Cette répartition aléatoire s'est faite au moment de la conception de l'enquête. On a procédé ainsi pour des raisons opérationnelles, la répartition aléatoire des unités non répondantes n'étant pas faisable dans la structure de système Blaise actuelle. La répartition aléatoire à cette étape permet d'effectuer des comparaisons entre les traitements, mais nous devons étudier l'effet des différences de réponse sur les traitements à partir du moment de la répartition aléatoire (Jackle et coll., 2010).

L'expérience intégrée devait se tenir au moyen d'enquêtes actives, les données recueillies devant servir à produire les estimations habituelles généralement produites à chaque cycle. Cela a imposé une contrainte importante à l'expérience. Nous devions nous assurer que les résultats finaux de la collecte ne soient pas touchés. Autrement dit, nous ne pouvions pas compromettre l'objectif ultime de produire des estimations à partir de cette collecte. Par conséquent, l'expérience a été conçue pour se tenir au cours des quatre premiers mois de la collecte. Une fois l'expérience terminée, une opération éclair finale auprès de toutes les unités non répondantes a été menée pour améliorer les taux de réponse.

Figure 3.1-1 : Illustration d'une expérience intégrée dans une enquête sur échantillon

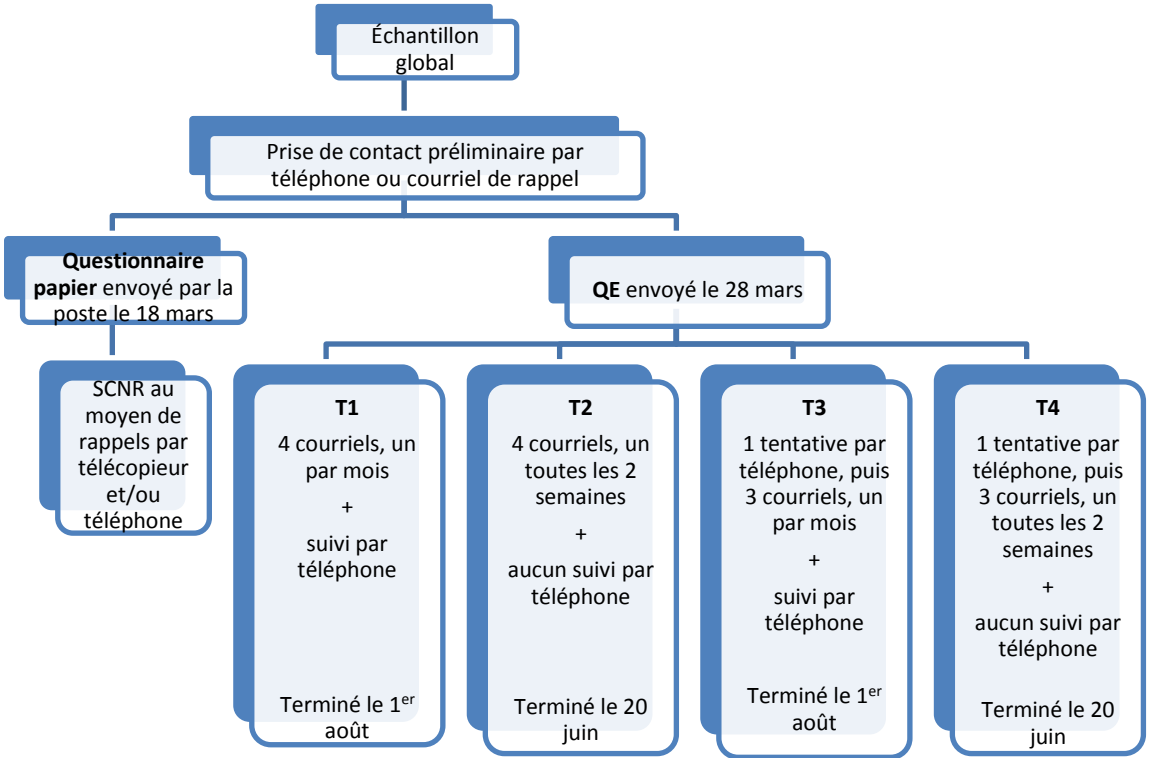


3.2 Répartition des traitements

La préparation de l'expérience est illustrée dans la figure 3.2 1. À noter que la répartition entre les unités recevant un questionnaire papier et celles recevant un questionnaire électronique n'a pas été faite de façon aléatoire. Elle a plutôt été déterminée par les répondants. Par conséquent, nous ne pouvons pas comparer les résultats entre la collecte des données par l'entremise de questionnaires papier et la collecte au moyen de QE. En fait, la collecte sur papier a été

menée auprès d'entreprises qui n'ont pas été contactées lors de la prise de contact préliminaire et de celles qui ont refusé d'utiliser le QE. Ainsi, ces entreprises sont plus sujettes à la non réponse et on croit qu'elles diffèrent de façon qualitative des unités qui ont adopté le mode de collecte au moyen de QE.

Figure 3.2-1 : Préparation de l'expérience



Les unités utilisant le mode de collecte au moyen de QE ont été affectées de façon aléatoire à un des quatre traitements, chaque traitement englobant environ 25 % de l'échantillon de collecte au moyen de QE. La répartition aléatoire a été faite à l'intérieur de chaque combinaison d'enquêtes, de groupes de strates et de types de questionnaires. Les trois groupes de strates sont les suivants : 1) petite strate à tirage partiel, 2) grande strate à tirage partiel et 3) strate à tirage complet et strate à tirage obligatoire. Le type de questionnaire a été déterminé selon la longueur, c'est à dire version longue ou version abrégée. Il convient de souligner que la stratification a été appliquée à six enquêtes, tandis que l'Enquête auprès des sièges sociaux a pris la forme d'un recensement. Par conséquent, nous nous sommes retrouvés avec un plan en blocs aléatoire pour les industries de service (enquêtes et strates utilisées comme blocs) et un plan complètement aléatoire pour l'Enquête auprès des sièges sociaux. Ainsi, les résultats ont dû être analysés séparément pour les enquêtes sur les services et l'Enquête auprès des sièges sociaux, afin de tenir compte des différences de plan de sondage.

Le 28 mars 2011, chaque unité des traitements 1 à 4 (T1 à T4) a reçu une invitation par courriel comprenant un hyperlien et un code d'accès pour répondre à l'enquête en ligne. Le suivi des cas de non réponse a commencé environ un mois après le début de la collecte, soit le 26 avril 2011. À ce moment là, toutes les unités non répondantes ont été envoyées au SCNR, selon une approche déterminée indiquée pour les T1 à T4. Chaque traitement comportait une approche de SCNR différente, utilisant une combinaison de rappels par courriel et de tentatives par téléphone. On interrompait le SCNR pour une unité, dès que le questionnaire était renvoyé à Statistique Canada.

Les quatre traitements ont été conçus de la façon suivante. Le traitement 1 (« standard ») a été conçu pour correspondre aux stratégies habituelles de SCNR pour les questionnaires papier. Autrement dit, nous avons assuré un suivi par téléphone tout au long de la collecte et, une fois par mois (26 avril, 26 mai, 23 juin et 25 juillet), nous avons envoyé un rappel par courriel pour que les non répondants remplissent leurs questionnaires (remplaçant ainsi les rappels par télécopieur par des rappels par courriel). Le traitement 1 devait servir de base pour la comparaison avec d'autres stratégies. Il devait prendre fin le 1er août, afin de permettre l'opération éclair finale par les intervieweurs.

Le traitement 2 a été conçu pour vérifier jusqu'où nous pouvions aller avec les rappels par courriel seulement, c'est à dire les types de taux de réponse que nous pouvions obtenir sans suivi par téléphone. Nous n'avons pas effectué de suivi par téléphone, mais les intervieweurs pouvaient répondre aux appels téléphoniques des répondants et prendre rendez vous pour recueillir les données, si tel était le souhait des répondants. Comme il n'y a pas eu de suivi téléphonique du tout, nous avons décidé d'envoyer des rappels par courriel plus fréquemment, soit une fois toutes les deux semaines (26 avril, 9 mai, 26 mai et 7 juin). Ainsi, ce traitement devait se terminer le 20 juin, afin de permettre l'envoi du même nombre de rappels par courriel que pour les autres traitements.

Le traitement 3 a été conçu pour mesurer les répercussions de la première tentative de suivi par téléphone plutôt que par courriel. Dans ce cas, la première tentative de suivi a été effectuée au téléphone, et le reste des rappels de suivi ont été effectués par courriel, une fois par mois (26 29 avril (tentative par téléphone), 26 mai (courriel), 23 juin (courriel) et 25 juillet (courriel)). Des appels de suivi par téléphone ont été effectués comme pour le traitement 1. Dans le cas du traitement 3, l'expérience a pris fin le 1er août.

Le traitement 4 était conçu pour combiner à la fois la première tentative de suivi par téléphone et le passage à des rappels par courriel seulement (26 29 avril (tentative par téléphone), 9 mai (courriel), 26 mai (courriel) et 7 juin (courriel)). Ainsi, nous n'avons pas effectué de suivi par téléphone, mais les intervieweurs étaient autorisés à répondre aux appels téléphoniques des répondants et à prendre des rendez vous pour recueillir les données. Dans le cas du traitement 4, l'expérience a pris fin le 20 juin encore une fois, comme pour le traitement 2, étant donné que nous avons choisi de comprimer le calendrier de suivi en envoyant des rappels par courriel toutes les deux semaines.

À la fin de l'expérience (20 juin ou 1er août), toutes les unités non répondantes ont été envoyées au SCNR par téléphone. Même si pour les traitements 2 et 4, la fin de l'expérience avait été fixée au 20 juin, le SCNR par téléphone a commencé le 8 juillet seulement pour ces traitements. À noter que dans le cas de ces deux traitements, un autre rappel par courriel a été envoyé le 7 juillet. Après le 7 août, l'expérience était terminée pour tous les traitements et l'opération éclair de SCNR par téléphone pour les quatre traitements a été lancée. La collecte active s'est poursuivie jusqu'au 14 octobre 2011, date à laquelle toutes les mesures de suivi ont pris fin. Cela a marqué la fin du cycle de collecte pour 2011.

4. Résultats

4.1 Statistiques descriptives

Au total, 9 324 unités ont été échantillonnées pour ces sept enquêtes. Parmi elles, 6 457 ont été affectées à la collecte au moyen de QE et ont par la suite été divisées de façon aléatoire entre quatre traitements de la même taille environ. Les quatre traitements comptaient 1 615, 1 613, 1 615 et 1 614 unités respectivement.

Au moment du suivi de la non réponse, les répondants pouvaient demander de changer de mode de collecte. Ainsi, pendant la collecte, 338 unités sont passées de la collecte sur papier à celle au moyen de QE, et 521 unités sont passées du QE à la collecte sur papier. Par ailleurs, 1 098 unités ont été considérées comme hors du champ de l'enquête ou retirées des affaires, comme l'a confirmé la collecte. À la fin de la collecte, le nombre d'unités comprises dans le champ de l'enquête pour les quatre traitements était de 1 375, 1 350, 1 394 et 1 376 respectivement.

4.2 Taux de renvoi

Le taux de renvoi, qui indique le pourcentage de questionnaires remplis et renvoyés, sert souvent de mesure clé des progrès de l'enquête. Il est déterminé lorsque le questionnaire est soumis à Statistique Canada, ou si l'unité est déclarée comme répondante au moyen d'un autre mode de collecte (interview téléphonique assistée par ordinateur, télécopieur, *etc*.). Les enquêtes après des entreprises comportent des populations très asymétriques, ce qui signifie qu'un nombre relativement faible d'unités peut représenter une partie importante de l'activité économique. Par conséquent, les taux de renvoi doivent être calculés à la fois de façon pondérée et non pondérée. Le taux de renvoi non pondéré indique le pourcentage de questionnaires reçus ou remplis, parmi toutes les unités comprises dans le champ de l'enquête, tandis que le taux de renvoi pondéré représente un pourcentage de la contribution au revenu des unités reçues ou complètes, par rapport à la contribution totale au revenu des unités comprises dans le champ de l'enquête.

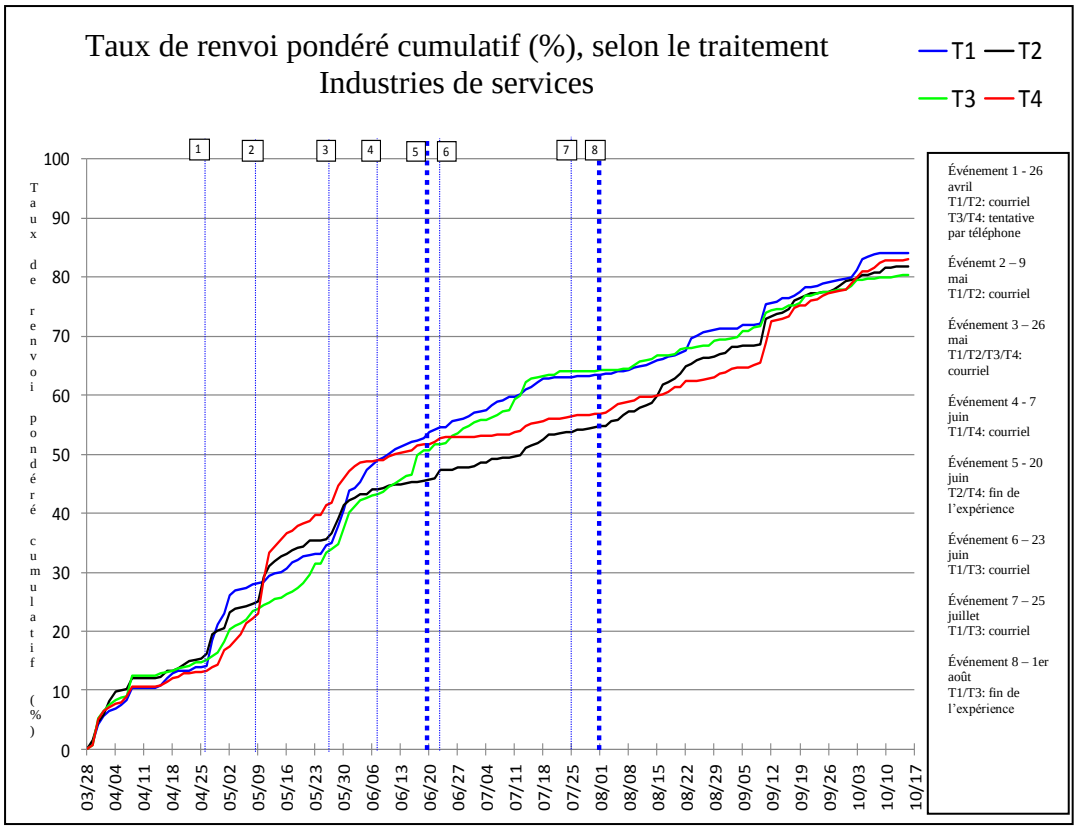
La progression du taux de renvoi pondéré total pour chaque traitement, de l'envoi des QE (28 mars) jusqu'à la fin de la collecte (14 octobre), est décrite dans le graphique 4.2 1. Les taux de renvoi ont été calculés sur la base de toutes les unités comprises dans le champ de l'enquête à la fin de la collecte. Tous les événements qui se sont produits aux dates clés sont indiqués par les lignes pointillées désignées « E1 » à « E8 ».

Au début du suivi des cas de non réponse (26 avril, E1), le taux de renvoi pondéré pour les quatre traitements était similaire (14 % à 16 %). Le 26 avril, les unités des traitements 1 et 2 avaient reçu un rappel par courriel, et celles des traitements 3 et 4 avaient fait l'objet d'une tentative par téléphone cette semaine là. Deux semaines plus tard (9 mai, E2), tous les traitements et la collecte sur papier obtenaient des taux de renvoi pondérés similaires (23 % à 25 %). Le 9 mai, les unités des traitements 2 et 4 ont reçu un rappel par courriel. Il semble que cela ait été utile et, après le 9 mai, le traitement 2 et, le traitement 4 plus particulièrement, venaient en tête du point de vue du taux de renvoi, par rapport aux deux autres traitements. Cette avance s'est maintenue jusqu'à la fin de mai pour le traitement 2 et jusqu'à la mi juin pour le traitement 4. Il semble que l'envoi d'un rappel toutes les deux semaines, plutôt que tous les mois, a fait en sorte que les traitements 2 et 4 ont pris de l'avance au début. Le 26 mai (E3), les unités des quatre traitements ont reçu un rappel par courriel. Le 7 juin (E4), les unités du traitement 2 ont reçu leur quatrième rappel par courriel et celles du traitement 4, leur troisième. Toutefois, ce quatrième rappel par courriel pour le traitement 2 ne semble pas avoir eu des répercussions positives similaires aux trois précédents.

Le 20 juin (E5, fin de l'expérience pour les traitements 2 et 4), les taux de renvoi étaient d'environ 50 % pour les quatre traitements. Nous avons pu récupérer plus de 40 % des questionnaires uniquement en envoyant des rappels par courriel et environ 50 %, uniquement en procédant à une tentative par téléphone complétée par des rappels par courriel. Il est encore plus intéressant de noter que le traitement 4, pour lequel une seule tentative par téléphone a été effectuée entre le 26 avril et le 29 avril, a obtenu presque les mêmes résultats que le traitement 1 (traitement standard), et que le traitement 3, dans lesquels un suivi par téléphone, selon la fonction de score, a été effectué continuellement à partir du 26 avril.

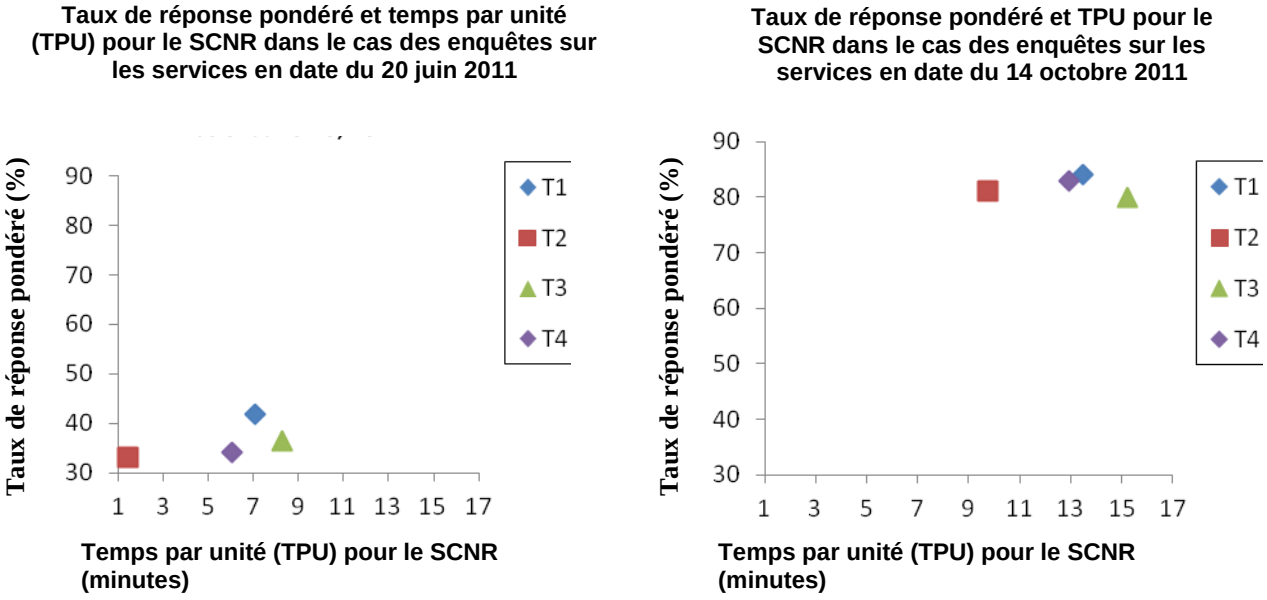
Après juin, les traitements 1 et 3 ont commencé à dépasser les traitements 2 et 4. Le 1er août (E8), ces traitements dépassaient les traitements 2 et 4 de plus de 10 %. Comme le suivi régulier par téléphone pour les traitements 2 et 4 a commencé après la fin de l'expérience (en fait, le 8 juillet, en raison de problèmes opérationnels), les répercussions du suivi par téléphone ont été perçues uniquement à la fin de l'été. À la fin de septembre, les quatre traitements de SCNR affichaient des taux de renvoi très similaires. Cette tendance est demeurée constante jusqu'à la fin de la collecte, l'ensemble des traitements prenant fin avec un taux de renvoi de plus de 80 %.

Graphique 4.2-1 : Taux de renvoi pondéré cumulatif observé chaque jour au cours de la collecte pour les industries de services



Puis, nous avons examiné les coûts de la collecte et nous les avons comparés entre les différents traitements. Les dates de début et de fin de chaque tentative par téléphone sont inscrites dans le système Blaise. Nous pouvons donc calculer la durée de chaque tentative, que l'on appelle aussi temps par unité (TPU). Des diagrammes de dispersion du taux de réponse pondéré par rapport au TPU pour le suivi des cas de non-réponse, en date des 20 juin et 14 octobre, figurent dans le graphique 5. Le traitement 2 (carrés rouges) comportait des taux de réponse similaires aux trois autres traitements, mais un TPU moins élevé consacré au SCNR. Comme il est plus coûteux de consacrer du temps au suivi par téléphone qu'à l'envoi de rappels par courriel (qui est un processus automatisé), un traitement a été considéré comme plus coûteux si le TPU du SCNR est plus élevé. Par conséquent, à partir des diagrammes de dispersion, nous avons observé que le traitement 2 pouvait produire des taux de réponse similaires, avec un coût plus faible de suivi par téléphone, par rapport aux traitements 1, 3 et 4.

Graphique 4.2-2 : Diagrammes de dispersion du taux de réponse pondéré par rapport à la durée moyenne de temps par unité (TPU) pour le SCNR (en date du 20 juin et du 14 octobre 2011)



4.3 Analyse de la variance

On a procédé à une analyse de la variance (ANOVA) pour les taux de renvoi pondérés et non pondérés, le nombre de jours écoulés avant la réception du questionnaire, le nombre de tentatives de SCNR et le temps par unité du SCNR.

Il convient de se rappeler que l'Enquête auprès des sièges sociaux est un recensement dans lequel chaque unité comporte le même poids d'échantillonnage, tandis que pour les enquêtes sur les services, on choisit un échantillon stratifié. Du fait des différences dans les plans d'échantillonnage, des tests d'analyse de la variance ont été exécutés séparément pour l'Enquête auprès des sièges sociaux et pour les enquêtes sur les services. Dans le cas des tests pour les unités des sièges sociaux, on a eu recours au test F de la procédure PROC GLM en SAS. Pour les enquêtes sur les services, on a eu recours au test de Wald rajusté pour tenir compte du plan d'enquête (Van den Brakel et Renssen, 2005).

Dans le tableau 4.3-1 ci-dessous, les résultats des tests ANOVA ont montré que, tant le 20 juin que le 14 octobre (fin de la collecte), les taux de renvoi non pondérés et pondérés n'étaient pas significativement différents dans les quatre traitements pour toutes les enquêtes. Le nombre de tentatives de SCNR et le TPU du SCNR des quatre traitements différaient de façon significative. Le nombre de jours écoulés avant la réception du questionnaire différait de façon significative dans les quatre traitements pour l'Enquête auprès des sièges sociaux, en date du 20 juin, mais pas de façon significative le 14 octobre. Il ne différait pas de façon significative pour les enquêtes sur les services, tant le 20 juin que le 14 octobre.

Tableau 4.3-1 : Essai relatif aux différences de traitement le 20 juin et le 14 octobre 2011

Principaux effets : T1 = T2 = T3 = T4	Sièges sociaux		Services	
	20 juin valeur p	14 octobre valeur p	20 juin valeur p	14 octobre valeur p
Taux de renvoi pondéré	—	—	0,9996	0,9999
Taux de renvoi non pondéré	0,0621	0,1590	0,6512	0,9999
Nombre de jours écoulés avant la réception du questionnaire	<0,0001	0,7637	0,0978	0,5231
Tentatives de SCNR	<0,0001	0,0256	<0,0001	0,0066
TPU du SCNR	<0,0001	0,0363	<0,0001	0,0104

5. Conclusion

Nous avons mis en œuvre avec succès une expérience intégrée pour vérifier quatre stratégies de suivi des cas de non-réponse des questionnaires électroniques pour sept enquêtes du PUSE, du cycle de collecte de 2011. La stratégie consistant à envoyer un rappel par courriel toutes les deux semaines au début de la collecte a permis d'obtenir les premiers 40 % de réponses à un coût plus faible. Les résultats ont montré que même si le suivi par téléphone commence trois mois plus tard que dans la stratégie actuelle de suivi, on peut obtenir des taux de réponse finaux très similaires, avec un moins grand nombre de tentatives par téléphone et moins d'effort.

Les tests initiaux effectués pour comparer les traitements mis en œuvre ont semblé appuyer les conclusions tirées de l'analyse descriptive. Toutefois, une analyse plus détaillée doit être effectuée pour mettre la dernière main aux conclusions globales de cette expérience. De façon plus particulière, nous aimerions examiner les répercussions de l'utilisation de modes différents de collecte et déterminer s'ils ont des répercussions sur les estimations.

6. Remerciements

Cette expérience a été rendue possible grâce aux efforts de nombreux secteurs de Statistique Canada. Les auteurs aimeraient souligner les efforts considérables et la contribution de la Division des industries de service, de la Division de la statistique des entreprises, de la Division des méthodes d'enquêtes auprès des entreprises, de la Division des méthodes d'enquêtes sociales, de la Division des systèmes et de l'infrastructure de collecte, de la Division des opérations et de l'intégration, de la Division de la planification et de la gestion de la collecte et du bureau régional de Sturgeon Falls, qui ont rendu cette étude possible.

Bibliographie

Groves, R.M. et L. Lyberg. (2010), « Total Survey Error: Past, Present, and Future », *Public Opinion Quarterly*, Oxford Journals, vol. 74, n° 5, p. 849 à 879.

Jäckle, A., Roberts, C. et P. Lynn (2010), « Assessing the effect of data collection Mode on Measurement », *International Statistical Review*, Wiley, vol. 78, n° 1, p. 3 à 20.

Van den Brakel, J. et R.H. Renssen (1998), « Design and Analysis of Experiments Embedded in Sample Surveys ». *Journal of Official Statistics*, vol. 14, n° 3, p. 277 à 295.

Van den Brakel, J. et R.H. Renssen (2005), « Analyse d'expériences intégrées dans des plans de sondage complexes », *Techniques d'enquête*, vol. 31, n° 1, p. 4 à 23.

Van den Brakel, J. et C.A.M. Van Berkel (2002), « A Design-based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys. An Application in the Dutch Labor Force Survey », *Journal of Official Statistics*, vol. 18, no 2, p. 217 à 231.

SÉANCE 6B

DONNÉES ABERRANTES ET IMPUTATION

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Intégration d'une procédure simplifiée de détection des valeurs aberrantes dans un système généralisé complexe

Laura T. Bechtel[1]

Résumé

Les distributions des données économiques ont tendance à être très asymétriques. Ceci rend difficile la détection des valeurs aberrantes, car certaines valeurs signalées comme étant des valeurs aberrantes ne le sont pas. Souvent, les procédures de détection des valeurs aberrantes appliquées aux données économiques s'appuient sur la comparaison de ratios. Le contrôle statistique d'Hidiroglou-Berthelot (« contrôle HB ») est utilisé couramment par les programmes économiques du U.S. Census Bureau. Le contrôle HB a été longtemps la seule méthode de détection des valeurs aberrantes disponible dans le Standard Economic Processing System (StEPS) (système normalisé de traitement des données économiques). Toutefois, ce contrôle présente deux contraintes : les données d'entrée doivent être positives et il ne peut pas être utilisé pour déceler les valeurs aberrantes sur un seul item. Ces deux contraintes ont empêché son utilisation par le programme du Quarterly Financial Report (QFR), dont le revenu (qui peut être négatif) est l'un des items clés pour lesquels des données sont recueillies. Par conséquent, le module de détection des valeurs aberrantes du StEPS a été amélioré par ajout de la méthode des limites robustes qui a donné d'excellents résultats pour d'autres applications du U.S. Census Bureau. Le présent article explique comment nous avons intégré les limites robustes dans le logiciel existant de détection des valeurs aberrantes, et décrit les défis que nous avons dû relever pendant ce processus et la façon dont nous l'avons fait.

1. Introduction

Les programmes économiques du U.S. Census Bureau recueillent et publient des données pour plus de 100 enquêtes entreprises. Nombre d'entre elles procèdent à la détection des valeurs aberrantes en utilisant le contrôle statistique par le test du ratio proposé par Hidiroglou et Berthelot (« contrôle HB ») en 1986. L'usage du contrôle HB est répandu dans le contexte des enquêtes entreprises, parce qu'il tient compte de la taille de l'unité pour détecter les ratios aberrants. Par conséquent, pendant plus de dix ans, le Standard Economic Processing System (StEPS) a été doté uniquement d'un module de contrôle HB au lieu d'un module de détection des valeurs aberrantes. Ce module était utile pour la plupart des programmes d'enquête qui utilisent le StEPS, mais certaines enquêtes produisent des données qui ne se prêtent pas à l'utilisation du contrôle HB.

Dans le présent article, nous discutons des procédures suivies pour ajouter au StEPS la méthode des limites robustes pour la détection des valeurs aberrantes. Nous commençons par donner un bref aperçu du StEPS et de la terminologie connexe à la section 2. À la section 3, nous présentons la méthode des limites robustes pour la détection des valeurs aberrantes. Les principales caractéristiques de la mise en œuvre du contrôle HB dans le StEPS sont expliquées à la section 4. La section 5 décrit les étapes que nous avons suivies pour intégrer les limites robustes dans le module existant de détection des valeurs aberrantes du StEP. Nous concluons à la section 6 en réfléchissant aux leçons apprises et en énumérant les avantages du processus d'intégration.

[1]Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau, Washington, DC 20233 (Laura.Bechtel@census.gov). Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche et de favoriser les discussions. Les opinions exprimées en ce qui concerne les questions méthodologiques ou opérationnelles sont celles de l'auteure et ne reflète pas forcément celles du U.S. Census Bureau. Je remercie Anne Russell, James Hunt, Xijian Liu et Katherine Thompson de leur examen minutieux assorti de commentaires constructifs des versions antérieures du présent manuscrit, ainsi que les membres du SMAG de leurs commentaires et suggestions concernant l'exposé.

2. Contexte du StEPS

Le Standard Economic Processing System (StEPS) est un logiciel généralisé utilisé pour mettre en œuvre de nombreuses enquêtes économiques différentes. Les modules qui constituent le StEPS comprennent la collecte des données, la vérification des données, la détection des valeurs aberrantes, l'imputation et l'estimation.

Comme tout logiciel, le StEPS possède une certaine terminologie qui lui est propre. Le terme **item** est utilisé pour décrire une variable contenant des données numériques stockées pour la publication ou pour le traitement des données, habituellement sur l'instrument d'enquête. Par exemple, si une enquête est conçue pour recueillir la valeur des ventes, la variable Ventes sera un **item**. En outre, une unité de déclaration ou de totalisation est désignée par **ID**. Lorsque l'on discute de données d'enquête traitées dans le StEPS, le **niveau ID-item** fait référence à la variable particulière déclarée pour une unité de déclaration ou de totalisation particulière. Par exemple, le module de détection des valeurs aberrantes permet à l'utilisateur de définir les spécifications pour un contrôle des valeurs aberrantes au niveau ID-item. Cela signifie que chaque ID sera soumis au test de contrôle des valeurs aberrantes pour l'item spécifié dans le test en question.

En plus de posséder sa propre terminologie, le StEPS possède son propre processus de contrôle des changements. Afin d'apporter un changement au système, une demande de changement (DR) doit être soumise au comité de contrôle des changements (CCB, pour *Change Control Board*) du StEPS. Tous les utilisateurs du StEPS sont représentés au sein du CCB et les demandes de changement doivent être approuvées par ce dernier pour pouvoir être mises en œuvre. Si la demande est approuvée, trois étapes doivent être accomplies avant que le changement soit introduit dans le StEPS : 1) déterminer et décrire les exigences, 2) harmoniser les exigences avec le logiciel existant et 3) procéder à la mise à l'essai et à la mise en œuvre. Quand la demande de changement concerne des questions méthodologiques, telles que la détection des valeurs aberrantes, le processus est supervisé par le comité consultatif de la méthodologie du StEPS (SMAG, pour *StEPS Methodology Advisory Group*), qui est un comité permanent formé de méthodologistes représentant les secteurs de programme qui utilisent le StEPS.

3. Méthodologie de détection des valeurs aberrantes

3.1 Contrôle HB

Le contrôle HB est une procédure sélective de vérification de ratios qui identifie moins de données comme étant « douteuses » comparativement aux tests de ratio habituels. Les ratios sont de la forme $R_i=x_i/y_i$ où x_i et y_i sont des variables positivement corrélées déclarées pour l'ID i. Habituellement, y est une valeur de x déclarée antérieurement, mais il peut également s'agir d'un item différent déclaré durant la même période statistique.

Le contrôle HB est utilisé pour produire des seuils de tolérance en vue de repérer les ratios qui sont aberrants et ceux qui ne le sont pas. Avant d'établir les seuils de tolérance, les données subissent plusieurs transformations. La première est la transformation de centrage à la suite de laquelle les observations transformées (S_i) sont centrées autour du ratio médian, R_m. Les observations centrées sont ensuite soumises à la transformation d'importance de taille, $E_i=S_i\times\{\max(x_i,y_i)\}^u$ où E_i est la statistique HB et u est le paramètre de taille qui varie dans l'intervalle [0, 1], et dont la valeur par défaut est fixée à 0,5.

Une fois la statistique HB calculée, l'étape suivante consiste à générer les seuils de tolérance (bornes supérieure et inférieure). En général, on utilise les quartiles et leurs différences. Cependant, le contrôle HB comprend un second terme pour éviter les différences nulles entre les quartiles :

- $D_{q1}=\max\{(E_m-E_{q1}),|A\times E_m|\}$ où E_m est la statistique HB médiane, E_{q1} est le premier quartile de la statistique HB et A est un multiplicateur (dont la valeur par défaut est habituellement 0,05) utilisé quand $E_m - E_{q1}$ est proche de zéro.

- $D_{q3}=\max\{(E_{q3}-E_m),|A\times E_m|\}$ où E_{q3} est le troisième quartile de la statistique HB.

Une fois que D_{q1}et D_{q3} sont calculées, les observations dont la statistique HB se situent en dehors de l'intervalle [E_m - $c\times D_{q1}$, $E_m + c\times D_{q3}$] ,où c est une constante prédéterminée, sont signalées comme des valeurs aberrantes.

Pour qu'un programme d'enquête utilise le contrôle HB, ses données doivent satisfaire certaines hypothèses sous-jacentes. Les deux hypothèses clés sont que les éléments de données comparés sont positivement corrélés et que les éléments de données étudiés doivent être **strictement non négatifs**. Ces conditions sont satisfaites par de nombreuses enquêtes qui utilisent le StEPS, mais pas toutes. Par exemple, le programme du Quarterly Financial Report (QFR) ne peut pas appliquer le contrôle HB à sa variable de revenu, parce que celui-ci peut être négatif. Toutefois un examen de la variable de revenu au niveau des macrodonnées est nécessaire avant la publication. Par conséquent, on a proposé d'ajouter la méthode des limites robustes dans le StEPS pour les enquêtes telles que le QFR.

3.2 Méthodes des limites robustes

Comparativement à la méthode du contrôle HB, la méthode des limites robustes est une forme beaucoup plus simple de détection des valeurs aberrantes, parce qu'aucune transformation des données n'est nécessaire. En fait, le contrôle HB est une forme particulière de limites robustes. La méthode des limites robustes comprend des seuils de tolérance qui sont calculés d'après les quartiles de la distribution pour signaler les observations aberrantes. Dans chaque cellule d'analyse spécifiée, on spécifie la distribution des variables à analyser comme étant celle d'un item unique, du ratio d'un item à sa valeur précédente déclarée ou du ratio de la valeur d'un item à celle d'un autre item fortement corrélé observée durant la même période de collecte. Après que l'on ait défini la variable à analyser, la méthode produit les statistiques suivantes : q_1, le premier quartile ; m, la médiane ; q_3, le troisième quartile ; H, l'écart interquartile (q_3 - q_1).

Ces statistiques servent alors à générer les bornes qui seront utilisées pour signaler les valeurs aberrantes dans les microdonnées. Il convient de souligner que ces quartiles peuvent être produits pour des données pondérées ou non pondérées, selon la préférence de l'utilisateur. Les deux méthodes générales de calcul de ces bornes sont les suivantes (Thompson, 1999) :

- **Limites robustes classiques –** signaler la valeur de la variable à analyser si elle est inférieure à $q_1 - k \times H$ ou supérieure à $q_3 + k \times H$;

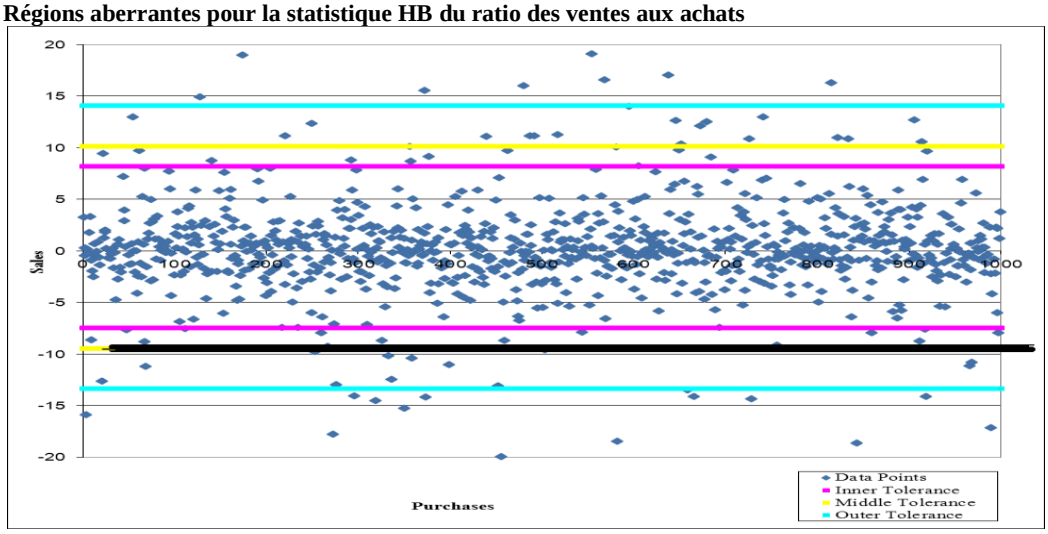
- **Limites robustes asymétriques –** signaler la valeur de la variable à analyser comme une valeur aberrante si elle est inférieure à $q_1 - k \times (m - q_1)$ ou supérieure à $q3 - k \times (q_3 - m)$

où k est une constante pré-spécifiée qui détermine la largeur des « limites ». Le paramètre k est analogue à la valeur c utilisée pour déterminer les seuils de tolérance du contrôle HB.

4. Le contrôle HB dans le StEPS

Le contrôle HB dans le StEPS peut être défini au niveau ID-item; chaque test de contrôle HB est spécifié de manière à signaler un item particulier. Le StEPS offre à l'utilisateur la souplesse de spécifier jusqu'à trois régions aberrantes pour chaque test en lui permettant de fournir jusqu'à trois valeurs de c. La figure 1 qui suit illustre comment ces trois régions (tolérances) différentes fonctionnent dans le StEPS.

Figure 4.1 :
Régions aberrantes pour la statistique HB du ratio des ventes aux achats



Les valeurs des observations comprises entre les limites de tolérance internes (droites roses) de la figure 1 sont considérées comme des valeurs légitimes selon le contrôle HB et ne sont pas signalées comme des valeurs aberrantes. Les valeurs des observations situées entre les limites de tolérance internes (roses) et moyennes (jaunes) sont des valeurs qui ne s'écartent que légèrement de la distribution et qui sont signalées en vue d'un examen par un analyste. Les valeurs comprises entre les limites de tolérance moyennes (jaunes) et extérieures (bleues) sont un peu plus douteuses et sont signalées en vue d'être supprimées de la base de données d'imputation en plus d'être signalées en vue d'un examen. Enfin, les valeurs qui se situent à l'extérieur des limites de tolérance externes (bleues) sont si extrêmes qu'elles sont signalées en vue d'une imputation en plus d'être signalées pour la suppression et l'examen.

Le jeu de paramètres qui suit peut être spécifié pour le contrôle HB au niveau ID-item, à savoir le numérateur, le dénominateur, trois valeurs différentes de *c*, la valeur de *u*, la valeur de *A*, et la variable de classe utilisée pour définir la cellule de détection des valeurs aberrantes.

Plusieurs commentaires doivent être faits au sujet des paramètres du contrôle HB. Ces paramètres sont spécifiés au niveau ID-item, ce qui signifie que l'utilisateur ne peut pas spécifier des paramètres de contrôle HB différents pour chaque cellule de détection des valeurs aberrantes. Seul l'item au numérateur est signalé comme une valeur aberrante. Un item peut être utilisé à plusieurs reprises comme un numérateur; un item peut être associé à plusieurs tests de contrôle HB différents. Cependant, un seul indicateur final de valeur aberrante est attribué au niveau ID-item. Le programme du StEPS est conçu pour définir les cellules de détection des valeurs aberrantes en utilisant pas plus d'une variable de classification, de sorte que, pour définir une cellule de détection des valeurs aberrantes en se servant de plus d'une variable de classe, l'utilisateur doit créer une variable recodée qui combine toutes les variables souhaitées en une seule variable de classe.

5. La procédure d'intégration

5.1 Demande de changement

Comme il est décrit à la section 2, une demande de changement (DC) doit être soumise pour approbation au CCB du StEPS avant que le changement puisse être mis en œuvre dans le StEPS. Au départ, le programme d'enquête intéressé par les limites robustes a soumis une DC pour une version très précise de la méthode, à savoir les limites robustes asymétriques. Cependant le CCB a demandé immédiatement que la DC soit examinée par le SMAG qui a déterminé que la demande devait être généralisée afin que les deux versions de la méthode des limites robustes soient mises en œuvre. En collaboration avec l'émetteur original de la DC, le SMAG a soumis une DC plus générale, qui a été approuvée par le CCB.

5.2 Détermination des exigences

Avant de commencer à déterminer les exigences, certains précédents résultant de la façon dont le contrôle HB a été mis en œuvre dans le StEPS devaient être satisfaits :

- Les paramètres devaient être spécifiés au niveau ID-item.
- Jusqu'à trois jeux de seuils de tolérance pouvaient être spécifiés pour chaque test de détection des valeurs aberrantes.
- Le niveau de la cellule de détection des valeurs aberrantes pour l'application du test pouvait être spécifié.
- Un seul indicateur final de valeur aberrante serait attribué au niveau ID-item.
- Les tests de ratio (positifs) devaient être analogues au contrôle HB.

En plus des précédents établis par le contrôle HB, le SMAG devait veiller à ce que les exigences permettent au programme demandeur (QFR) de mettre en œuvre ses procédures de détection des valeurs aberrantes. Simultanément, les exigences devaient être suffisamment générales pour que d'autres secteurs de programme puissent utiliser la méthode. Clairement, le SMAG devait tenir compte de nombreux points de vue. Il n'est donc pas étonnant qu'il y ait eu de nombreux points de friction qui ont entraîné beaucoup de discussions et très peu de résolution des problèmes.

Finalement, comme nous avions le sentiment de piétiner, le SMAG a examiné les méthodes employées dans le passé pour essayer de résoudre les désaccords. Une solution efficace, mais équitable, consistait à produire une liste des questions en suspens. Le procédé était simple – les exigences sur lesquelles on n'avait pu se mettre d'accord ont été énumérées et chaque représentant de division a proposé une solution. Si nous ne pouvions pas prendre une décision unanime, nous avons adopté la règle de la majorité, en donnant un peu plus de poids à l'opinion de la division demandeuse. Cela a permis d'accélérer le processus de détermination des exigences qui ont été établies en quelques réunions.

Une question n'a pu être résolue en utilisant la liste des questions en suspens : que devrait être l'indicateur final de détection des valeurs aberrantes si un item est soumis à plus d'un test de détection des valeurs aberrantes et que les indicateurs résultants ne concordent pas? La figure 2 qui suit illustre ce phénomène. L'ID et l'item sont énumérés dans la première et la deuxième colonne, respectivement. Les trois autres colonnes présentent les indicateurs résultant de trois tests différents de détection des valeurs aberrantes pour la combinaison ID-item présentée dans les deux premières colonnes. Comme il est mentionné plus haut, il n'existe qu'un seul indicateur de valeur aberrante au niveau ID-item. Donc, la question qui se pose pour les ID 0001 à 0003 est : quel est l'indicateur final de détection des valeurs aberrantes pour le revenu trimestriel (QREV)? Pour l'ID 0001, la réponse est simple – attribuer un indicateur d'imputation. Pour les ID 0002 et 0003, la réponse n'est pas aussi simple et un ensemble de règles est nécessaire pour attribuer l'indicateur final de valeur aberrante.

Tableau 5.2.1
Indicateurs de détection de valeurs aberrantes pour un item

ID	Item	Test des limites robustes 1	Test des limites robustes 2	Test du contrôle HB 1
0001	QREV	Imputer (I)	Imputer (I)	Imputer (I)
0002	QREV		Adresser à un analyste (A)	Supprimer (S)
0003	QREV	Supprimer (S)		Imputer (I)

Quand le contrôle HB a été conçu, aucun choix n'a été offert. Si plus d'un test de contrôle HB était exécuté sur la même combinaison ID-item et que les tests produisaient des indicateurs de valeurs aberrantes différents, celui correspondant à la valeur « la plus » aberrante était choisi comme indicateur final de valeurs aberrantes. Le secteur

de programme qui a demandé la méthode des limites robustes souhaitait exactement l'opposé, c'est-à-dire qu'on sélectionne l'indicateur correspondant à la valeur « la moins » aberrante. En terme plus technique, les utilisateurs voulaient résoudre ce problème en choisissant de minimiser l'erreur de type I ou l'erreur de type II. L'erreur de type I est la probabilité de signaler une valeur légitime comme étant une valeur aberrante, et l'erreur de type II est la probabilité de ne pas signaler une vraie valeur aberrante. Dans la figure 3, les deux dernières colonnes montrent quels indicateurs seraient attribués aux ID 0002 et 0003 en minimisant respectivement l'erreur de type I ou l'erreur de type II.

Tableau 5.2.2
Indicateurs de détection des valeurs aberrantes en minimisant l'erreur de type I ou de type II

ID	Test des limites robustes 1	Test des limites robustes 2	Test du contrôle HB 1	Minimiser l'erreur de type I	Minimiser l'erreur de type II
0001	I	I	I	I	I
0002		A	S		S
0003	S		I		I

L'exigence résultante a été de donner à l'utilisateur le choix de minimiser l'erreur de type I ou l'erreur de type II pour chaque item soumis à la détection des valeurs aberrantes, l'option par défaut dans le StEPS étant de minimiser l'erreur de type II.

5.3 Conception et mise en œuvre

Si la détermination des exigences a semblé être un exercice très difficile pour les méthodologistes, la tâche des programmeurs a été tout aussi difficile, voire plus. Premièrement, ils ont disséqué le code existant pour trouver le module du contrôle HB, puis ont décidé comment ils allaient programmer nos exigences dans ce module. La communication et le compromis ont été des éléments essentiels durant cette phase du processus d'intégration.

Les programmeurs nous ont expliqué les étapes de la conception, une fois que celle-ci a été achevée. La conception concordait en grande partie avec nos exigences, mais quelques-unes n'avaient pas été incorporées. Par exemple, nous avions spécifié que le module des limites robustes devrait permettre à l'utilisateur de spécifier jusqu'à six variables pour définir une cellule de détection des valeurs aberrantes. Après avoir examiné le problème, les programmeurs nous ont fait savoir que cela n'était pas possible en raison des contraintes du code existant du contrôle HB. En particulier, ils auraient été obligés de réécrire le code du contrôle HB pour pouvoir satisfaire à cette demande et cela était trop risqué. Au lieu de cela, comme le contrôle HB, le module de limites robustes n'aurait la capacité de spécifier qu'une seule variable. Bien que nous souhaitions avoir la souplesse de spécifier six variables de classe au lieu d'une, nous n'avons pas insisté sur cet aspect parce d'un côté nous ne voulions pas compromettre la fonctionnalité du module de contrôle HB existant et de l'autre nous pouvions utiliser six variables, à condition de les combiner en une seule. Il existait des exigences plus nécessaires pour lesquelles aucun compromis n'était acceptable.

Parfois, il ne s'agissait pas d'en arriver à un compromis avec les programmeurs, mais entre les membres du SMAG. Durant l'explication de la conception, les membres du SMAG ont eu de la difficulté à se mettre d'accord sur des questions telles que la façon dont les choses devraient être affichées à l'écran et les conventions d'attribution de nom pour les nouvelles variables. Enfin, après de nombreuses négociations, la conception du module a été achevée et prête à être mise à l'essai.

5.4 Essai et mise en œuvre

Il était essentiel que cette nouvelle fonctionnalité du StEPS soit mise à l'essai complètement avant de passer officiellement à l'environnement de production. Cependant, comme il s'agissait d'une nouvelle fonctionnalité, de nombreux utilisateurs ne savaient pas comment la mettre à l'essai. Par conséquent, le SMAG a établi un plan généralisé de mise à l'essai indiquant précisément aux utilisateurs quels paramètres ils devaient spécifier et à quelles données de sortie ils devraient s'attendre. Certains membres du SMAG ont également élaboré et mis en œuvre des séances de formation concernant la mise à l'essai. Après avoir reçu la formation, les responsables de l'essai se sont servi du plan pour effectuer la mise à l'essai. Naturellement, de petits problèmes ont été découverts et corrigés. Finalement, la méthode des limites robustes a été mise en œuvre dans le StEPS.

6. Conclusion

En dernière analyse, le processus d'intégration a duré environ un an. Beaucoup de temps a été consacré à la détermination des exigences et à la conception du logiciel. Cela n'a peut-être pas été le processus d'intégration le plus long qui soit, mais il est certain qu'il aurait pu être amélioré. Nous avons tiré quelques leçons qui pourraient rendre des projets similaires plus efficaces.

Premièrement, lorsque l'on modifie un module de logiciel, il est important que les utilisateurs courants du logiciel soient consultés. Ces utilisateurs devraient bien connaître les avantages et les inconvénients du logiciel existant. Dans notre cas, nos spécialistes des logiciels ont fourni des informations précieuses pour la détermination des exigences. Deuxièmement, il importe de choisir judicieusement ses batailles. Ce conseil, qui paraît évident, peut être oublié rapidement lorsque l'on essaie de développer le logiciel « le meilleur qui soit ». À mesure que progresse le processus de détermination des exigences, les désaccords sont inévitables; il est important de faire la distinction entre les exigences « obligatoires » et celles qui sont « souhaitables ». Troisièmement, il faut mettre sur pied un groupe spécialisé pour déterminer les exigences. Cela aide à garder le cap et à maintenir l'élan, et à empêcher le groupe de se laisser distraire par d'autres questions. Quatrièmement, il faut dresser la liste des questions en suspens afin d'en faire le suivi et de les résoudre. Le fait de consigner par écrit les questions en suspens et les solutions proposées réduit les discussions et accélère le processus de prise de décisions. Enfin, il convient de créer des groupes distincts pour déterminer les exigences fonctionnelles (façon de procéder à la mise en œuvre) et non fonctionnelles (méthodologie). Laisser les méthodologistes décider de la façon dont un programme devrait être mis en œuvre (par exemple, à quoi les écrans devraient ressembler) pourrait entraver le processus de prise de décisions s'ils ne sont pas les utilisateurs finaux.

Fait agréable, participer au processus de modification a des avantages inattendus. La nouvelle fonctionnalité pour la méthode des limites robustes a été mise en œuvre dans le StEPS d'une manière telle qu'elle s'est prêtée à l'amélioration de la fonctionnalité existante du contrôle HB. Durant le processus de détermination des exigences, la correction des défauts connus de la façon dont le contrôle HB avait été mis en œuvre a pu être intégrée dans la mise en œuvre de la méthode des limites robustes. Notre objectif était de justifier encore davantage l'apport de ces améliorations au contrôle HB afin qu'il soit en harmonie avec son homologue de détection des valeurs aberrantes. Enfin, le processus complet d'intégration de la nouvelle méthode a facilité les discussions et a permis de mieux comprendre les différentes approches de la mise en œuvre de méthodes identiques (ou similaires) de détection des valeurs aberrantes. En dernière analyse, nous avons non seulement offert à nos utilisateurs un système de détection des valeurs aberrantes plus souple, mais nous leur avons aussi donné un aperçu plus général des méthodologies et approches existantes.

Bibliographie

Hidiroglou, M.A. et J.-M. Berthelot (1986), « Contrôle statistique et imputation dans les enquêtes-entreprises périodiques », *Techniques d'enquête*, vol. 12, n°1, p. 85 à 97.

Thompson, K.J. (1999), « Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods », Statistical Policy Working Paper 29 (document de travail) Federal Committee on Statistical Methodology, U.S. Census Bureau (http://www.fcsm.gov/99papers/thompson.pdf).

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Outil de détection de valeurs aberrantes à Statistique Canada

Nelson Émond[1]

Résumé

Plusieurs méthodes de détection de valeurs aberrantes peuvent être utilisées et le sujet est bien couvert dans la littérature. Le choix final d'une méthode reposera sur l'expérience du spécialiste du sujet et sur l'expertise du méthodologiste puisqu'ils tiendront compte de la structure des données. Par contre, des contraintes de temps, de ressources et de budget restreignent les responsables d'enquêtes à considérer plusieurs solutions de rechange et ils s'en remettent souvent à une méthode déjà en place et plus facile d'accès au détriment d'une autre qui serait, peut être, mieux adaptée.

Pour combler cette lacune, un outil a été développé en SAS. Il regroupe les méthodes les plus usuelles à Statistique Canada. Cela permet de conserver l'expertise au profit d'autres enquêtes ayant le même profil. L'originalité de cet outil est de donner une vue d'ensemble des données en permettant de les visualiser à l'aide de graphiques interactifs afin de comparer les méthodes entre elles. Ainsi, il n'y a pas de coût de développement pour les enquêtes et il est possible de choisir la méthode la plus appropriée tout en optimisant les paramètres. Il ne faut pas négliger les conséquences d'un choix inadéquat de méthodes et de paramètres, car ils peuvent augmenter le fardeau de travail à l'étape de l'imputation et influencer la qualité des estimations.

Mots clés : Détection ; aberrant ; influent ; atypique ; robuste.

1. Introduction

1.1 Définition

Jusqu'à présent, il n'y a pas de consensus pour définir ce qu'est une valeur aberrante. Il y a eu une tentative intéressante par Hawkins (1983) qui propose que la majorité des observations suivent un modèle à priori et les observations suffisamment éloignées de ce modèle seraient déclarées aberrantes. Cette notion de modèle sera omniprésente dans cet article. Plusieurs méthodes de détection seront présentées et le choix dépendra du modèle sous-jacent.

1.2 Description

La structure des observations ainsi que l'objectif de la détection influencent le choix d'une méthode de détection de valeurs aberrantes par rapport à une autre. C'est dans cette optique qu'un outil a été conçu pour offrir plusieurs méthodes de détection et permettre de les comparer. L'outil permet de faire une représentation graphique des données et de visualiser l'impact du choix des paramètres. C'est essentiellement un outil de développement et d'analyse des données. Contrairement au système généralisé BANFF qui fait, entre autres, de la détection de valeurs aberrantes, cet outil n'est pas orienté vers la production. Par contre, le système BANFF ne possède pas d'interface graphique permettant de visualiser les données. Les méthodes Hidiroglou-Berthelot et interquartile sont déjà instaurées dans BANFF. Une troisième méthode, l'écart-sigma, est en cours d'intégration. L'outil, présenté ici, inclut ces trois méthodes et en contient cinq autres.

[1]Nelson Émond, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (nelson.emond@statcan.gc.ca).

2. Méthodes

2.1 Contexte

Lors d'un sondage interne à Statistique Canada demandant aux responsables d'enquêtes qu'elle était la méthode qu'ils employaient, les méthodes Hidiroglou-Berthelot et de l'écart-sigma se sont avérées les plus populaires. Les méthodes interquartiles et des contributeurs influents sont souvent envisagées, car elles sont très simples. Les autres méthodes sont souvent plus complexes et il y aura seulement une description sommaire de ces méthodes. L'outil offre le choix de huit méthodes : Higiroglou-Berthelot, écart-sigma, interquartile, processus bayésien séquentiel, estimation M, distance de Mahalanobis, méthodes classiques et contributeurs influents. L'outil a été développé avec le logiciel SAS et ne traite que les données continues; il est facile et rapide d'utilisation et comprend plusieurs fonctions pour aider l'utilisateur qui seront décrites plus en détail dans une section subséquente.

2.2 Méthodes

2.2.1 Hidiroglou-Berthelot

C'est la méthode la plus utilisée à Statistique Canada (voir Hidiroglou et Berthelot, 1986). C'est une méthode bivariée qui a été développée, à l'origine, pour des données historiques, mais peut aussi être utilisée pour des variables corrélées, par exemple revenu (x) contre profit (z). Son principal intérêt vient du fait qu'elle tient compte de la taille comme facteur de déclaration de valeurs aberrantes. Il y a trois restrictions dont il faut tenir compte pour appliquer cette méthode :

- les observations doivent être strictement positives et continues ;
- il faut qu'il y ait une relation linéaire fondée sur le ratio entre les deux variables ;
- la droite passe par l'origine.

Les étapes de la méthode pour une unité *i* dans une classe spécifique *h* sont :

1. Calcul du ratio : $r_{hi} = \dfrac{x_{hi}}{z_{hi}}$

2. Transformation du ratio: $s_{hi} = \begin{cases} 1 - \dfrac{r_{hM}}{r_{hi}} & si \quad 0 < r_{hi} < r_{hM} \\ \dfrac{r_{hi}}{r_{hM}} - 1 & si \quad r_{hi} \geq r_{hM} \end{cases}$ où r_{hM} est la médiane des r_{hi}.

3. Calcul des effets : $e_{hi} = s_{hi}\left[\text{Max}\left(x_{hi}, z_{hi}\right)\right]^U$ où $0 \leq U \leq 1$, U étant le facteur de courbure.
4. Calcul du premier quartile (e_{hq1}), de la médiane (e_{hM}) et du troisième quartile (e_{hq3}) des effets (e_{hi}).
5. Distance : $\begin{aligned} d_{hq1} &= \text{Max}\left(e_{hM} - e_{hq1}, \left|A \cdot e_{hM}\right|\right) \\ d_{hq3} &= \text{Max}\left(e_{hq3} - e_{hM}, \left|A \cdot e_{hM}\right|\right) \end{aligned}$

   où *A (=0,05* par défaut) est déterminé par l'utilisateur et assure une valeur minimale de la distance.
6. Une valeur sera considérée aberrante si : $\begin{cases} e_{hi} < e_{hM} - C_{crit} \cdot d_{hq1} \\ e_{hi} > e_{hM} + C_{crit} \cdot d_{hq3} \end{cases}$

   où les variables *A*, C_{crit} et *U* sont les paramètres déterminés par l'utilisateur.

Dans la figure 2.2.1-1, on voit le principal intérêt de la méthode qui tient compte de la variable de taille en abscisse. Le ratio d'acceptation est d'autant plus faible que la taille est grande. La figure 2.2.1-2 est une représentation différente du même phénomène.

La reproduction graphique est établie selon un standard qui inclut les courbes qui délimitent la zone d'acceptation au-delà de laquelle les valeurs seront aberrantes. Ces courbes sont disponibles lorsque cela est possible de le faire. Des renseignements supplémentaires sont fournis dans le titre. Il est également possible de mettre sur le graphique

(voir Figure 2.2.1-2) ou en légende la liste des identificateurs des observations aberrantes. Toutes les autres méthodes utilisent le même standard graphique.

Figure 2.2.1-1
Représentation du taux de variation c. la taille

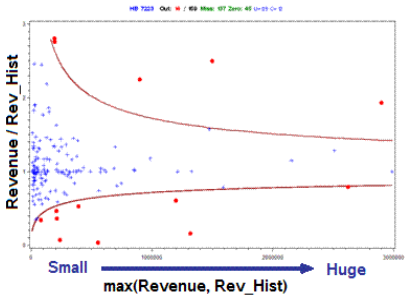
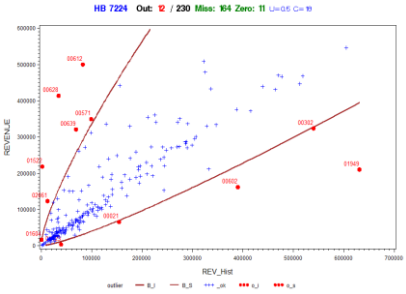


Figure 2.2.1-2
Représentation du revenu c. revenu historique



2.2.2 Écart-sigma

C'est une méthode univariée ou bivariée. Dans le cas où elle est bivariée, elle peut se comparer à la méthode Hidiroglou-Berthelot, car elle utilise le ratio. Les données n'ont pas à être exclusivement positives. C'est une méthode qui se fonde sur la distance entre deux observations consécutives selon la valeur de l'écart-type. Les étapes de la méthode dans une classe *h* sont :

1. Posons $r_{hi} = x_{hi}$ dans le cas univarié et $r_{hi} = \frac{x_{hi}}{z_{hi}}$ dans le cas bivarié.

2. Calcul de l'écart-type (σ_h) selon une des quatre méthodes disponibles dans l'outil.

3. Mettre en ordre croissant les r_{hi}.
4. L'utilisateur fournira le percentile (*K*) à partir duquel le critère de distance sera appliqué. Donc, toutes les valeurs en deçà du *K* ne sont pas admissibles à être aberrantes. Pour les autres, lorsque la distance entre deux observations consécutives satisfait la condition suivante :

   $$r_{hi} - r_{hj} > C * \sigma_h \text{ où } j = i - 1 \text{ et } C > 0$$

   toutes les observations plus grandes que r_{hj} seront déclarées aberrantes.

L'utilisateur aura à déterminer les paramètres *K*, *C* et la méthode de calcul de l'écart-type. Par défaut, le système utilise la méthode *Mad* pour le calcul de σ_h.

2.2.3 Interquartile

C'est une méthode univariée ou bivariée fondée sur le ratio. Elle ne tient pas compte de la taille de l'entreprise. Elle est un cas particulier de la méthode Hidiroglou-Berthelot dans le cas bivarié avec U=0. Elle est souvent utilisée, car elle offre la possibilité à l'utilisateur d'exercer un bon contrôle sur le nombre de données aberrantes. Voici les étapes à suivre :

1. Posons $r_{hi} = x_{hi}$ dans le cas univarié et $r_{hi} = \frac{x_{hi}}{z_{hi}}$ dans le cas bivarié.

2. Mettre en ordre croissant les r_{hi}.

3. Calculer les valeurs du 25^{e} (Q_1), la médiane (Q_2) et le 75^{e} (Q_3) percentile.

4. Une donnée est aberrante si elle satisfait un des critères suivants :

   $$r_{hi} \begin{cases} < Q_2 - K_{inf} * [Q_2 - Q_1] \\ > Q_2 + K_{sup} * [Q_3 - Q_2] \end{cases}$$

L'utilisateur aura à fournir les paramètres K_{inf} et K_{sup} qui sont un multiple de la distance interquartile.

2.2.4 Processus bayésien séquentiel

Cette méthode suppose à priori qu'il existe une relation linéaire entre les variables explicatives et la variable d'intérêt. C'est une méthode complexe (voir Philips et Gutman (2006) pour des explications détaillées). Sommairement, c'est un processus itératif qui enlève une observation à chaque itération. Au cours d'une itération, chaque observation reçoit une probabilité d'être aberrante selon l'influence qu'elle a sur la droite de régression. L'observation qui sera enlevée est celle dont la probabilité d'être aberrante est la plus élevée. La probabilité dépend du ratio RStudent/Student.

La recherche de valeurs aberrantes s'arrête lorsqu'on a atteint une corrélation acceptable ou lorsqu'il n'existe plus d'observation avec une probabilité d'être aberrante suffisamment élevée ou que le pourcentage maximal d'observations aberrantes, déterminé par l'utilisateur, est atteint. L'utilisateur aura à fournir le pourcentage maximum qui est acceptable.

2.2.5 Estimation M

Un des buts de la détection de valeurs aberrantes est de trouver un estimateur robuste. Plusieurs techniques utilisent des méthodes non paramétriques qui ne présument pas d'une distribution à priori. Mais de nouvelles méthodes paramétriques ont vu le jour au début des années 1960 comme l'estimation M dont le pionnier est Huber (1964). Il y a eu beaucoup de développement depuis cette époque. En somme, la technique de l'estimation M peut se résumer comme suit :

$$Minimize_{\theta} \sum_{i=1}^{n} \rho(r_i)$$

où ρ est une fonction symétrique, c'est-à-dire $\rho(-t) = \rho(t), \boldsymbol{\theta}$ est le vecteur des paramètres de la droite de régression et $r_i = y_i - \boldsymbol{X}_i^T \boldsymbol{\theta}$.
En dérivant par rapport aux coefficients θ_j on obtient la fonction suivante :

$$\sum_{i=1}^{n} \psi(r_i) x_i = 0$$

où x_i est le vecteur des variables explicatives. Donc, cela donne p équations (p=nombre de variables explicatives + 1) dont la solution est souvent difficile. En pratique, un processus itératif est utilisé pour trouver une solution. Il y a plusieurs choix de la fonction ψ, mais celle offerte par le système est celle de Huber. L'originalité de la méthode proposée est d'appliquer la méthode de détection des valeurs aberrantes des interquartiles sur les résidus robustes $\hat{r}_i = y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\theta}}$, où $\hat{\boldsymbol{\theta}}$ est l'estimateur robuste de θ, ce qui permet de déterminer les valeurs aberrantes.

2.2.6 Distance de Mahalanobis

Cette méthode présume que la majorité des données sont autour d'un point central. Elle est bien résumée dans l'article de Franklin, Thomas et Brodeur (2000). Les principales étapes sont :

1. Centrer les données en utilisant l'estimateur L_1 (voir Rousseeuw et Leroy, 1984).
2. Initialiser les poids de départ à $\delta_i^n = 1$ pour $i = 1, \ldots n$.
3. Déterminer un nombre d'itérations souhaitées (10 itérations sont habituellement suffisantes) :
   a) générer aléatoirement un vecteur « Y_1 » normalisé ;
   b) calculer les autres vecteurs $Y_2, \ldots, Y_p$ où p est le nombre de variables pour qu'ils forment un ensemble de vecteurs orthogonaux-orthonormés ;
   c) calculer les poids δ_i pour chaque enregistrement ;
   d) si $\delta_i^k < \delta_i$ alors $\delta_i = \delta_i^k$;

4. Calculer les nouveaux vecteurs pondérés $(\hat{\boldsymbol{u}})$ et de la matrice de variance-covariance $(\hat{\boldsymbol{V}})$;
5. Calculer la distance de Mahalanobis robuste : $D_i = (\boldsymbol{x}_i - \hat{\boldsymbol{u}})^T \hat{\boldsymbol{V}}^{(-1)} (\boldsymbol{x}_i - \hat{\boldsymbol{u}})$;
6. Effectuer le test de détection : si $\frac{(n-p)n}{(n^2-1)p} D_i > F_{\alpha;p,n-p}$, alors l'unité *i* est déclarée aberrante.

C'est une méthode robuste multivariée itérative développée initialement par Patak (1990) qui ne présume pas que les observations suivent un modèle de régression.

2.2.7 Méthodes classiques

On retrouve sous cette rubrique un ensemble de tests statistiques dont l'accent est mis sur les observations ayant une grande influence sur l'estimateur des moindres carrés. Les diagnostics consistent en une combinaison de statistiques numériques et graphiques. Plusieurs diagnostics sont fondés sur les résidus et d'autres sont fondés sur l'incidence de la suppression d'une observation.

Parmi ces tests, nommons les suivants : matrice de projection, distance de Cook, DfBetas, DfFits, CovRatio et RStudent. Une bonne documentation peut être consultée dans Weisberg (1985). Des procédures SAS sont utilisées pour le calcul de ces tests.

Ladiray et Ramsay (2003) ont appliqué ces tests dans le cas bivarié. Ils ont mis l'accent sur l'effet des observations atypiques sur les différents coefficients de corrélation, tels que Kendall, Spearman, Pearson et le R-carré, en plus de calculer la pente et l'ordonnée à l'origine avec et sans la présence de données atypiques. L'utilisateur peut choisir un ou plusieurs tests et une observation est déclarée atypique dès qu'elle échoue à au moins un test. De plus, les valeurs atypiques peuvent provenir de l'une ou l'autre des deux variables.

2.2.8 Contributeurs influents

C'est une méthode très intuitive fondée sur l'hypothèse que les valeurs aberrantes sont celles qui sont les plus susceptibles d'influencer les estimations. Par conséquent, c'est une méthode qui s'intéresse davantage aux grandes valeurs. Elle laisse beaucoup de latitude à l'utilisateur d'où sa popularité. L'utilisateur peut choisir de déterminer :

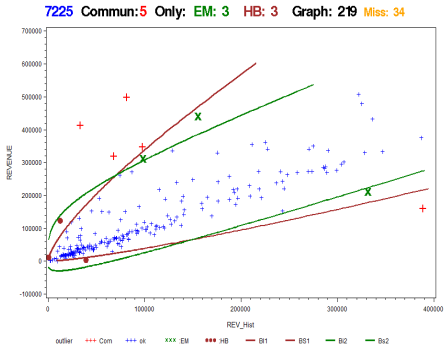
- celles qui ont les plus grandes valeurs par strate ;
- celles qui contribuent à plus qu'un certain pourcentage du total de la strate ;
- celles qui sont au-delà d'une certaine valeur seuil par strate ;
- celles dont la différence entre deux valeurs consécutives est plus grande qu'un certain pourcentage ;
- celles qui sont les plus grandes observations qui contribuent à plus d'un certain pourcentage du total (par exemple : 3,85, est-ce que les trois plus grandes valeurs contribuent à plus de 85 % du total? Si oui, les identifier comme valeurs influentes).

Un attrait de cette méthode est qu'elle identifie, habituellement, peu de valeurs influentes.

2.3 Comparaison de méthodes à l'aide de l'outil

C'est un module qui a été développé pour voir graphiquement l'effet d'une méthode par rapport à une autre. Divers renseignements liés aux méthodes sont indiqués sur le graphique. Dans le titre, on retrouve la strate, le nombre d'observations aberrantes qui sont communes aux deux méthodes, le nombre d'observations aberrantes détectées par chaque méthode, le nombre d'observations sur le graphique et combien sont manquantes. Il est aussi possible de comparer une seule méthode avec deux ensembles de paramètres différents. De plus, les courbes qui délimitent chaque méthode seront représentées lorsque cela est possible de le faire.

Figure 2.3-1
Estimation M (EM) c. Hidiroglou-Berthelot (HB)



2.4 Options de l'outil

L'utilisateur dispose de plusieurs options :

- déclarer la variable de poids dans les calculs pour certaines méthodes ;
- déclarer la variable de la structure de variance du modèle de régression pour certaines méthodes ;
- exclure les valeurs égales à zéro, manquantes et/ou négatives ;
- exporter un graphique dans un fichier html ;
- mettre le numéro de l'identificateur sur le graphique ou dans la légende ;
- agrandir une zone particulière du graphique ;
- valider le nombre d'observations par strate, le type de variables, le nom, *etc.* ;
- exporter la liste des données aberrantes dans un fichier.

3. Développements futurs

L'accent sera mis sur la rédaction d'un Guide de référence et d'un Guide de référence rapide. La mise à l'essai pour vérifier si des méthodes bivariées peuvent être appliquées avec des observations multivariées est en cours. Il existe une possibilité de créer une interface plus conviviale si le temps le permet. La priorité ne sera pas mise sur le développement de nouvelles méthodes dans un proche avenir. Toutefois, il sera possible d'ajouter de nouvelles fonctionnalités aux méthodes déjà en place selon les besoins et les demandes des utilisateurs.

4. Conclusion

La détection de valeurs aberrantes est souvent la partie négligée dans les enquêtes par manque de temps, de ressources ou d'un budget restreint. Il ne faudrait pas oublier qu'une mauvaise détection de valeurs aberrantes a une incidence néfaste sur l'imputation et l'estimation. On s'en remet souvent à une méthode acceptable qui est déjà en place faute de temps pour développer une méthode de rechange. Un outil comme celui présenté dans cet article permet de disposer d'une multitude de méthodes regroupées dans un seul environnement tout en permettant de faire des analyses graphiques. Il est important de mentionner qu'il n'y a pas une méthode qui est meilleure qu'une autre; une méthode sera préférable par rapport à une autre selon la structure des données et si elle répond aux besoins des utilisateurs.

Bibliographie

Hawkins, D.M. (1983), « Outliers », *Encyclopedia of Statistical Science*, éds, S. Kotz et N.L. Johnson., New York : JohnWiley and Sons.

Hidiroglou, M. et J.-M. Berthelot (1986), « Contrôle statistique et imputation dans les enquêtes-entreprises périodiques », *Techniques d'enquêtes*, vol.12, p.79 à 89.

Huber, P.J. (1964), « Robust estimation of a location parameter », *Annals of Mathematical Statistics*, vol. 35, n^{o} 1, p. 73 à 101.

Ladiray, D. et L. Ramsay (2003), « Statistical evaluation of the CoA-based comparison between tax and survey data », Statistique Canada, document de travail, Ottawa, Canada.

Philips, R. et I. Guttman (2006), « Towards the Robust Estimation of Parameters in the Univariate Linear Model », publication interne, Ottawa, Canada: Statistique Canada

Rousseeuw, P.J. et A.M. Leroy (1987), « Robust regression and outlier detection ». *Psychometrika*, New York : John Wiley and Sons, vol. 55, n^{o} 1, p. 182 à 183.

Weisberg, S. (1985), *Applied Linear Regression*, 2^{e} édition, chapitre 5, University of Minnesota St-Paul, Minnesota, New York : John Wiley and Sons.

Franklin, S., Thomas, S. et M. Brodeur (2000), « Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Modified Stahel-Donoho estimators », International Conference on Establishment Survey (ICES) II.

Patak, Z. (1990), « Robust Principal Component Analysis Via Projection Unit », Master's Thesis, Université de la Colombie-Britannique, Canada.

Évaluation de méthodes d'imputation de l'exposition au risque dans le profil de risque des acteurs d'un modèle pour la microsimulation

Deirdre Hennessy, Carol Bennett, Meltem Tuna, Claude Nadeau, William Flanagan et Douglas Manuel[1]

Résumé

L'imputation des données manquantes sur l'exposition au risque représente un aspect courant et important des modèles de microsimulation. Les données pour la microsimulation sont souvent puisées à plusieurs sources et nécessitent l'imputation des réponses aux enquêtes manquantes ainsi que des variables manquantes (c'est-à-dire les variables nécessaires pour la modélisation qui ne sont pas disponibles dans la source principale de données utilisées pour initialiser le modèle). Alors que l'imputation pour corriger la non réponse aux enquêtes peut être effectuée par des méthodes validées, l'imputation des variables manquantes pose plus de difficulté, parce qu'elle comprend l'utilisation de données provenant de « donneurs » externes qui peuvent ou non être comparables à celles de la source initiale de données. Les spécialistes de la microsimulation ont suivi diverses approches pour imputer les variables manquantes; cependant, on ne sait pas clairement laquelle produit les résultats les plus valides. En outre, les avantages des diverses approches d'imputation pour la microsimulation dépendent du rôle de la variable dans l'éventuel modèle de microsimulation. En prévision de l'élaboration d'un modèle de microsimulation de la santé de la population (POHEM) pour la maladie cardiovasculaire et la consommation de sel, nous avons étudié diverses techniques permettant de créer des variables imputées pour la pression artérielle et le taux de cholestérol, qui représentent des risques fondamentaux de maladie cardiovasculaire. Nous nous sommes servis du cycle 2.2 de l'Enquête sur la santé dans les collectivités canadiennes comme source initiale de données pour la plupart des risques de maladie cardiovasculaire, y compris l'apport de sodium. L'information sur la pression artérielle et le cholestérol a été tirée de données de « donneurs » (venant de l'Enquête canadienne sur les mesures de la santé). Nous évaluons deux approches d'imputation. La première est l'imputation par la régression qui comporte l'utilisation d'une ou de plusieurs variables communes aux deux fichiers de données. La seconde englobe les méthodes d'imputation « hot-deck » qui consistent à attribuer à chaque acteur du modèle une valeur réelle de la pression artérielle ou du taux de cholestérol obtenue d'un « donneur ». Nous avons comparé les méthodes en nous fondant sur l'exactitude, la discrimination et la validité des valeurs imputées ; cependant, nous avons également dû examiner minutieusement les propriétés statistiques de chaque méthode et les répercussions de l'utilisation des données résultantes dans la microsimulation.

Mots clés : Imputation ; données d'enquête ; microsimulation ; pression artérielle ; cholestérol.

1. Introduction

1.1 Imputation pour la microsimulation

L'imputation de données manquantes sur l'exposition au risque est un problème fréquent, complexe et important que posent presque tous les modèles de microsimulation pour la santé. Plusieurs sources de données sont souvent nécessaires pour créer les profils de santé des acteurs d'un modèle de microsimulation, y compris les caractéristiques socioéconomiques, l'exposition aux risques pour la santé (exposition au risque) et la situation concernant la maladie. Habituellement, on se sert d'une source principale de données pour initialiser chaque profil d'acteur, puis on regroupe les caractéristiques manquantes et on les impute en se servant de plusieurs sources de « donneurs » de données. Diverses approches sont suivies pour imputer les caractéristiques manquantes, mais on ne sait pas clairement laquelle produit les meilleurs résultats. C'est-à-dire un ensemble complet de données contenant des

[1]Deirdre Hennessy, Institut de recherche de l'Hôpital d'Ottawa et Statistique Canada, 1 053, avenue Carling, Ottawa, Ontario, K1Y 4E9 (deirdre.hennessy@statcan.gc.ca) ; Carol Bennett, Institut de recherche de l'Hôpital d'Ottawa, 1 053, avenue Carling, Ottawa, Ontario, K1Y 4E9 (cbennett@ohri.ca) ; Meltem Tuna, Institut de recherche de l'Hôpital d'Ottawa, 1 053, avenue Carling, Ottawa, Ontario, K1Y 4E9, (mtuna@ohri.ca) ; Claude Nadeau, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Ontario, K1A 0T6 (claude.nadeau@statcan.gc.ca) ; William Flanagan, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Ontario, K1A 0T6, (william.flanagan@statcan.gc.ca) ; Douglas Manuel, Institut de recherche de l'Hôpital d'Ottawa et Statistique Canada, 1 053, avenue Carling, Ottawa, Ontario, K1Y 4E9 (dmanuel@ohri.ca).

variables imputées qui se comportent de la même façon que dans l'ensemble de données « donneur » et qui s'associent de manière prévue (en fonction d'autres résultats publiés) aux variables qui figurent dans la source initiale de données. Dans le contexte de l'élaboration d'un modèle de microsimulation de la santé de la population (POHEM : MCV) en vue d'illustrer et de projeter la relation entre l'apport alimentaire de sodium et la maladie cardiovasculaire (MCV), l'imputation était essentielle, parce qu'aucune source unique de données ne reflète toutes les expositions importantes. Nous avons tiré du cycle 2.2 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) les données pour la plupart des caractéristiques du profil de santé des acteurs, y compris l'apport de sodium et d'autres renseignements nutritionnels; cependant, ces données ne comprenaient pas les mesures de la pression artérielle et du taux de cholestérol, qui sont des facteurs de risque fondamentaux utilisés dans tous les algorithmes de calcul du risque de MCV. Par conséquent, nous avons dû imputer ces mesures en nous fondant sur l'Enquête canadienne sur les mesures de la santé (ECMS), puis valider les résultats de l'imputation avant de procéder à la spécification du modèle de microsimulation.

Nous avons évalué deux approches générales d'imputation. Pour commencer, nous nous sommes penchés sur l'imputation par la régression, qui comprend la spécification d'un modèle en se servant d'une ou de plusieurs variables communes (variables communes aux deux fichiers de données, par exemple l'âge, le sexe et l'indice de masse corporelle (IMC)). L'un des avantages de cette méthode est que la variable qui doit être imputée peut être modélisée en s'appuyant sur un ensemble éventuellement grand d'autres variables qui expliquent la relation causale (Durrant, 2005). En outre, les modèles de régression permettent de décrire la relation entre les variables en se servant des coefficients du modèle qui peuvent être comparés aux relations causales établies dans le cadre d'autres études. Les coefficients produits par un modèle de régression peuvent également être utilisés par d'autres chercheurs sans que ceux-ci aient besoin de données de « donneurs ». Un inconvénient éventuel de cette méthode tient au fait qu'elle peut fausser la distribution de la variable explicative et accroître l'association entre cette dernière et les autres variables du modèle. En outre, les valeurs imputées sont prédites plutôt qu'observées réellement dans une autre source de données, et il s'agit d'une approche paramétrique qui pourrait être sensible à l'erreur de spécification du modèle de régression (Durrant, 2005).

En deuxième lieu, nous avons examiné des méthodes d'imputation « hot-deck » qui consistent à attribuer des valeurs réelles de pression artérielle et de taux de cholestérol tirées de données de « donneurs » aléatoirement dans des classes d'imputation construites en se fondant sur le recoupement de variables communes entièrement observées (par exemple l'âge, le sexe et l'IMC). L'un des avantages de cette méthode est qu'elle est non paramétrique ou semi-paramétrique et qu'elle a pour objectif d'éviter de formuler des hypothèses concernant les distributions, propriété importante si la distribution des données qui doivent être imputées est asymétrique. Si l'on se sert de méthodes « hot-deck », les valeurs imputées des acteurs du modèle devraient avoir une distribution de même forme que celles des valeurs des répondants à l'enquête similaires figurant dans l'ensemble de données « donneur ». Cette propriété est importante parce que, si les valeurs extrêmes de la variable imputée représentent des personnes courant un risque élevé de maladie ou de décès dans la population, ce qui est le cas pour le cholestérol et la pression artérielle, il est important de recréer cette distribution afin de prédire correctement le risque de maladie ou de décès. De surcroît, comme les valeurs imputées sont utilisées dans une population de départ pour la simulation, les inexactitudes produites à cette étape deviendront additives à mesure que le modèle de simulation projettera les valeurs vers l'avenir. Un autre avantage de cette méthode est qu'un enregistrement « donneur » peut fournir un grand nombre de valeurs manquantes à un enregistrement « receveur ». Dans la présente étude, toutes les données sur les mesures physiques provenaient de la même base de données (ECMS), si bien qu'utiliser le même « donneur » pour fournir toutes les valeurs manquantes pourrait aider à préserver les relations existantes entre les mesures effectuées sur une même personne. Un inconvénient de ce modèle est qu'il dépend entièrement des données de « donneurs » qui peuvent ou non comprendre des variables dont les valeurs sont recueillies ou mesurées de la même façon que les données initiales. Qui plus est, les données sur d'importantes variables susceptibles d'être reliées causalement à la variable imputée et d'être utilisées dans l'imputation « hot-deck » pourraient ne pas figurer du tout dans l'ensemble de données « donneur ». Un autre inconvénient de cette méthode est que la taille de l'échantillon auprès duquel est obtenu l'ensemble de données « donneur » doit être suffisamment grande pour que la méthode donne de bons résultats (Durrant, 2005).

Nous avons évalué les méthodes d'imputation par la régression et d'imputation « hot-deck » en nous fondant sur trois critères. Premièrement, nous avons évalué l'exactitude de l'imputation en comparant aux données de l'ECMS les valeurs imputées pour l'ensemble de la population et pour diverses sous-populations. Ensuite, nous avons examiné le degré de discrimination de l'imputation (c'est-à-dire la capacité de discerner l'exposition entre des

personnes ou des groupes). Nous avons évalué le degré de discrimination en comparant la distribution et l'étendue des valeurs (centiles 1, 5, 10, *etc*.) des données imputées et des données de l'ECMS. Enfin, nous avons examiné les données imputées afin de déterminer si les relations connues entre les variables y étaient préservées, comme l'association entre l'âge avancé et l'hypertension. La relation entre les variables dans les données imputées devrait correspondre à celles dégagées de revues systématiques et de méta-analyses d'autres études (Khaw et Barrett-Connor, 1988; Strazzullo et coll., 2009).

L'objectif de l'étude était d'étudier diverses techniques permettant de créer des variables imputées pour l'hypertension et le cholestérol, en utilisant le cycle de 2007 à 2009 de l'ECMS comme ensemble de données donneur et le cycle 2.2 de l'ESCC comme ensemble de données receveur.

2. Méthodes

2.1 Sources des données

2.1.1 Source initiale : Enquête sur la santé dans les collectivités canadiennes, cycle 2.2

Le cycle 2.2 de l'ESCC a été utilisé comme source principale ou initiale des données. Il s'agit d'une enquête représentative de la population nationale menée auprès de 35 107 Canadiens (21 106 adultes de 18 ans et plus) en 2004. Cette enquête comprenait la collecte de données sur l'autoévaluation de l'état de santé, les problèmes de santé chroniques et l'activité physique, ainsi que de renseignements détaillés sur la consommation d'aliments au moyen d'un questionnaire de rappel alimentaire de 24 heures, soit la norme de référence en ce qui concerne les données sur la nutrition disponibles au Canada.

2.1.2 Source de donneurs : Enquête canadienne sur les mesures de la santé

L'ECMS de 2007-2009 a été utilisée comme source de donneurs de données. Cette enquête comprenait la collecte de données sur l'autoévaluation de l'état de santé, les maladies chroniques et l'activité physique de la même manière que dans l'ESCC. En outre, elle comportait la collecte de mesures physiques de la santé, dont l'IMC, la pression artérielle et les taux de cholestérol. Au cours du cycle 1, réalisé de 2007 à 2009, des données ont été recueillies auprès de 5 604 Canadiens (3 719 adultes de 18 ans et plus) à 15 emplacements au Canada.

2.2 Analyse statistique

2.2.1 Imputation par la régression

L'imputation par la régression comprend l'ajustement d'un modèle de régression qui relie la variable dépendante Y aux variables auxiliaires X (variables communes aux deux ensembles de données pour lesquelles les données sont entièrement observées). Les valeurs prédites, obtenues par modélisation, sont utilisées pour l'imputation des valeurs manquantes de Y. Plus simplement, soit $Y_{(\text{imputée})} = \mathrm{E}\{Y|X\}$ (Durrant, 2005).

Dans la présente étude, nous avons modélisé les niveaux mesurés de la pression artérielle systolique (PAS) et diastolique (PAD), ainsi que les taux de cholestérol total et de lipoprotéines de haute densité (HDL) (toutes les variables pour lesquelles des données sont disponibles dans l'ECMS) sur un ensemble de variables auxiliaires communes disponibles à la fois dans l'ECMS et le cycle 2.2 de l'ESCC. En nous servant de techniques de régression linéaire simples ainsi que des poids de sondage et des poids bootstrap, nous avons spécifié des modèles distincts pour les hommes et pour les femmes, et utilisé les valeurs prédites obtenues pour imputer les valeurs manquantes. Le tableau 2.3.1-1 résume les variables explicatives incluses dans les modèles finaux pour la pression artérielle et pour le cholestérol.

Tableau 2.3.1-1
Résumé des variables explicatives dans les modèles d'imputation de la PAS, la PAD, et les taux de cholestérol total et de HDL

	Âge (c)	IMC (c)	HTN connue (2)	Médicaments anti-HTN (2)	Maladie cardiaque (2)	Diabète (2)	Études (2)	État matrimonial (2)	Ethnicité (2)	Propriétaire (2)	Activité physique (2)	Santé générale (2)	Fume tous les jours (2)
PAS													
Hommes	√	√	√		√								
Femmes	√	√	√				√						
PAD													
Hommes	√	√	√	√	√	√		√			√		
Femmes	√	√	√	√	√								
Cholestérol total													
Hommes	√	√	√		√	√		√					
Femmes	√	√	√			√							√
HDL													
Hommes	√	√						√		√			
Femmes	√	√			√	√			√		√	√	√

Nota : (c) = variable continue, (2) = nombre de catégories de la variable, PAS = pression artérielle systolique, PAD = pression artérielle diastolique, HDL = lipoprotéines de haute densité, IMC = indice de masse corporelle, HTN = hypertension.

2.3.2 Imputation « hot-deck »

Dans l'imputation « hot-deck », on attribue la valeur d'un enregistrement contenant des données observées à un enregistrement dans lequel il y a des données manquantes. Selon cette méthode, les valeurs manquantes sont imputées d'après d'autres enregistrements de la base de données qui possèdent des attributs en commun avec la variable incomplète. On construit pour cela des classes d'imputation fondées sur des variables auxiliaires communes disponibles dans les deux ensembles de données et les valeurs provenant de « donneurs » sont sélectionnées à l'intérieur des classes d'imputation (Kalton et Kaspryzk, 1982; Durrant, 2005).

Dans la présente étude, nous nous sommes servis d'un sous-ensemble de variables explicatives les plus importantes, déterminées au moyen de la modélisation par la régression décrite plus haut, comme classes d'imputation dans l'imputation « hot-deck » (voir le tableau 2.3.1-1). Nous avons également procédé à l'imputation selon le sexe. Nous nous sommes servis d'une macro SAS pour l'imputation « hot-deck » itérative qui nous a permis d'imputer les valeurs manquantes pour la PAS, la PAD, le cholestérol total et le cholestérol HDL simultanément dans nos classes d'imputation (Ellis, 2007).

3. Résultats

3.1 Imputation par la régression – Exactitude

Pour évaluer l'exactitude de l'imputation, nous avons comparé les estimations de la PAS obtenues selon les données de l'ECMS et selon les données imputées dans l'ensemble, et par sous-groupes importants. Nous avons choisi l'imputation de la PAS comme exemple ici pour des raisons d'espace. Le tableau 3.1-1 montre que l'imputation par la régression donne des estimations exactes des valeurs moyennes et médianes de la PAS pour l'ensemble de la population et selon l'âge, le sexe et le niveau d'études. Remarquablement, l'étendue des valeurs estimées en utilisant l'imputation par la régression dans l'ensemble et pour les divers sous-groupes est fortement tronquée, comme en témoignent les valeurs maximales et minimales. Cette constatation est explorée plus en détail à la section suivante.

Tableau 3.1-1
Exactitude des valeurs imputées de la PAS dans l'ensemble et selon le sous-groupe

	Don.	Imp.	Don.	Imp.	Don.	Imp.	Don.	Imp.
	Moyenne	**Moyenne**	**Médiane**	**Médiane**	**Max.**	**Max.**	**Min.**	**Min.**
Dans l'ensemble	116,91*	116,55	114,63	115,90	194,61	150,08	80,22	94,72
Sexe								
Hommes	118,73	116,04	117,42	115,79	194,61	140,073	83,01	99,82
Femmes	115,28	116,96	111,84	116,07	189,96	150,08	80,22	94,72
Groupe d'âge								
18 à 44 ans	109,72	106,04	109,05	106,04	153,69	124,74	80,22	94,72
45 à 64 ans	120,31	118,34	118,35	117,54	185,31	134,02	86,73	108,20
65 ans et plus	129,04	129,70	127,65	129,01	194,61	150,08	84,87	114,24
Études <études secondaires	119,86	122,55	117,42	123,08	189,96	148,47	83,94	94,72
Études secondaires ou plus	115,55	114,30	113,70	113,65	194,61	150,081	80,22	96,62

Nota : PAS = pression artérielle systolique, Don. =donneur, Imp. = imputée, Max. = maximum, Min. = minimum.
*La PAS est mesurée en millimètres de mercure (mmHg).

3.2 Imputation par la régression – Discrimination

Pour évaluer le degré de discrimination de l'imputation par la régression, nous avons comparé la distribution et l'étendue des valeurs estimées. Les histogrammes comparant les valeurs mesurées et imputées de la PAS d'après l'ECMS et l'ESCC ont montré que l'étendue des valeurs imputées était tronquée comparativement à celle des valeurs fondées sur l'ECMS (résultats non présentés). Le tableau 3.2-1 donne une comparaison des centiles des données mesurées et imputées, ainsi que la différence entre les deux valeurs. À l'extrémité supérieure de la distribution en particulier (90^{e} centile et au-delà), les données imputées produisent des sous-estimations importantes comparativement aux valeurs de l'ECMS.

Tableau 3.2-1
Comparaison de la distribution des valeurs mesurées et imputées pour la PAS

	Don.	Imp.	Différence
Centiles			
1 %	91,38	95,69	-4,31
5 %	96,96	98,18	-1,22
10 %	99,75	101,56	-1,81
25 %	106,26	108,32	-2,06
50 %	114,63	115,90	-1,27
75 %	124,86	124,30	0,56
90 %	136,95	131,98	4,97
95 %	143,46	136,71	6,75
99 %	163,92	142,05	21,87

Nota : PAS = pression artérielle systolique, Don. =donneur, Imp. = imputée

3.3 Imputation hot-deck – Exactitude

Le tableau 3.3-1 montre que l'imputation « hot-deck » produit aussi des estimations exactes des valeurs moyennes et médianes de la PAS, dans l'ensemble et selon le sexe et le niveau d'études. Pour les groupes d'âge, il existe des écarts (jamais supérieurs à 4,05 unités) entre les données de donneurs et les données imputées. L'étendue des valeurs estimées en utilisant l'imputation hot-deck dans l'ensemble et selon le sous-groupe est très semblable à celle des données de donneurs, les données imputées produisant des minimums légèrement inférieurs à ceux observés avec les données de l'ECMS. Cette constatation est examinée plus en détail à la section suivante.

Tableau 3.3-1
Exactitude des valeurs imputées de la PAS, dans l'ensemble et selon le sous-groupe

	Don.	Imp.	Don.	Imp.	Don.	Imp.	Don.	Imp.
	Moyenne	**Moyenne**	**Médiane**	**Médiane**	**max.**	**Max.**	**Min.**	**Min.**
Dans l'ensemble	116,91	115,01	114,63	113,00	194,61	197,00	80,22	74,00
Sexe								
Hommes	118,73	116,25	117,42	115,00	194,61	197,00	83,01	77,00
Femmes	115,28	114,00	111,84	111,00	189,96	192,00	80,22	74,00
Groupe d'âge								
18 à 44 ans	109,72	106,10	109,05	105,00	153,69	153,00	80,22	74,00
45 à 64 ans	120,31	116,72	118,35	115,00	185,31	187,00	86,73	81,00
65 ans et plus	129,04	126,83	127,65	125,00	194,61	197,00	84,87	79,00
Études								
<études secondaires	119,86	119,73	117,42	118,00	189,96	197,00	83,94	74,00
Études secondaires ou plus	115,55	113,41	113,70	112,00	194,61	194,00	80,22	77,00

Nota : PAS = pression artérielle systolique, Don. =donneur, Imp. = imputée, Max. = maximum, Min. = minimum.

3.4 Imputation hot-deck – Discrimination

Les histogrammes comparant les valeurs mesurées et imputées de la PAS provenant de l'ECMS et de l'ESCC ont montré que l'étendue des valeurs imputées était fort semblable à celle des valeurs mesurées (résultats non présentés). Le tableau 3.4-1 donne une comparaison des centiles des valeurs mesurées et imputées, ainsi que la différence entre ces valeurs. Comparativement à l'imputation par la régression, la méthode « hot-deck » reconstruit plus exactement la distribution des valeurs imputées sur l'ensemble de la distribution. Contrairement à l'imputation par la régression, l'imputation « hot-deck » donne lieu à une sous-estimation des valeurs à l'extrémité inférieure de la distribution.

Tableau 3.4-1
Comparaison de la distribution des valeurs mesurées et imputées de la PAS

	Don.	Imp.	Différence
Centiles			
1 %	91,38	86,00	5,38
5 %	96,96	93,00	3,96
10 %	99,75	96,00	3,75
25 %	106,26	104,00	2,26
50 %	114,63	113,00	1,63
75 %	124,86	124,00	0,86
90 %	136,95	136,00	0,95
95 %	143,46	145,00	-1,54
99 %	163,92	165,00	-1,08

Nota : PAS = pression artérielle systolique, Don. =donneur, Imp. = imputée

3.5 Validité

Dans la présente étude, la validité des valeurs imputées a été difficile à évaluer, car nous ne disposions pas d'un ensemble de données de comparaison comprenant à la fois l'apport de sodium **mesuré** et la pression artérielle **mesurée**. Une approche consiste à vérifier si les relations connues, par exemple l'augmentation de la prévalence de l'hypertension avec l'âge, l'indice de masse corporelle (IMC) et l'autodéclaration de l'état d'hypertension, sont préservées dans l'ensemble de données imputées (Khaw et Barrett-Connor, 1988; Strazzullo et coll., 2009). Comme le montrent les tableaux 3.1-1 et 3.2-1 qui précèdent, les valeurs imputées par la régression et celles imputées par la méthode « hot-deck » augmentent avec l'âge. En outre, la relation prévue entre la PAS moyenne et l'accroissement

de l'IMC et de l'autodéclaration de l'état d'hypertension est également préservée dans l'ensemble de données imputées (résultats non présentés).

Nous avons aussi spécifié des modèles de régression de la PAS imputée selon le sexe, corrigés pour l'âge, l'IMC et l'autodéclaration/la connaissance d'un diagnostic d'hypertension. Les résultats de la comparaison des données obtenues par la méthode de régression et par la méthode « hot-deck » aux données de l'ECMS sont présentés au tableau 3.5-1. L'association significative entre l'âge, l'IMC et la PAS est préservée dans les deux ensembles de données imputées, et les coefficients de régression sont semblables. Il est intéressant de souligner que les intervalles de confiance (IC) des coefficients obtenus en recourant à l'imputation par la régression sont très étroits. Par contre, l'association significative entre l'état d'hypertension et la PAS n'est pas préservée lorsqu'on utilise la méthode « hot-deck », mais elle l'est si l'on applique la méthode d'imputation par la régression. Alors qu'il existe une relation statistiquement significative entre les valeurs de la PAS moyenne lorsque la comparaison est faite en se fondant sur l'état d'hypertension (résultats non présentés), la stratification selon le sexe et la correction pour l'âge et l'IMC semblent éliminer l'association. Ces résultats, particulièrement les intervalles de confiance étroits sous imputation par la régression et l'association non significative entre l'état d'hypertension et la PAS dans le cas de l'imputation « hot-deck », doivent faire l'objet d'une étude plus approfondie qui pourrait aboutir à certains ajustements de la façon dont l'imputation « hot-deck » est mise en œuvre, par exemple en ajoutant des classes d'imputation supplémentaires.

Tableau 3.5-1
Relation entre la PAS imputée et d'autres variables dans les données de donneurs et les données imputées

Variable	Coefficient		IC		valeur p	
	Hommes	**Femmes**	**Hommes**	**Femmes**	**Hommes**	**Femmes**
ECMS-âge (années)	0,26	0,45	(0,22;0,29)	(0,41;0,49)	<0,00	<0,00
Régr.	0,29	0,56	(0,28;0,29)	(0,55;0,56)	<0,00	<0,00
HD	0,26	0,50	(0,23;0,28)	(0,47;0,53)	<0,00	<0,00
ECMS-IMC (kg/m^2)	0,23	0,41	(0,08;0,38)	(0,23;0,58)	<0,00	<0,00
Régr.	0,25	0,13	(0,23;0,25)	(0,12;0,13)	<0,00	<0,00
HD	0,33	0,49	(0,39;0,58)	(0,23;0,42)	<0,00	<0,00
ECMS-état d'hypertension (non/oui)	5,05	9,07	(2,69;7,42)	(6,62;11,52)	<0,00	<0,00
Régr.	5,09	8,45	(4,92;5,25)	(8,38;8,52)	<0,00	<0,00
HD	0,27	0,65	(-0,94;1,48)	(-0,69;1,99)	0,34	0,67

Nota : PAS = pression artérielle systolique, IC = intervalle de confiance, Régr. = régression, HD= hot-deck, IMC = indice de masse corporelle

4. Conclusion

L'objectif de l'étude était d'évaluer les techniques d'imputation par la régression et d'imputation « hot-deck » comme méthodes d'imputation des facteurs de risque importants dans les profils des acteurs de modèles de microsimulation. Les deux méthodes ont permis de reproduire avec succès les valeurs moyennes et médianes des variables imputées, mais les méthodes « hot-deck » recréaient mieux la distribution des valeurs que la méthode par la régression. Nous avons aussi étudié la validité des valeurs imputées, quoique de façon limitée. Ces résultats ont montré que les associations prévues de la PAS avec l'âge et l'IMC étaient préservées dans les deux ensembles de données imputées, tandis que l'association de la PAS avec l'état d'hypertension n'était pas préservée par l'imputation « hot-deck ». Ces résultats doivent être examinés de manière plus approfondie et de futures analyses auront pour objectif de déterminer si les relations entre les valeurs imputées et les variables **non** utilisées dans le processus d'imputation sont préservées. En outre, d'autres méthodes d'imputation, dont l'imputation par la régression à partir d'une distribution conditionnelle et l'imputation multiple, seront étudiées. Enfin, il faudra également se pencher sur l'association ultime entre l'apport de sodium et la PAS imputée, afin d'établir la validité des valeurs imputées.

Dans l'ensemble, les résultats de l'étude donnent à penser que les méthodes d'imputation par la régression et « hot-deck » offrent des avantages différents et que le choix de la méthode pourrait dépendre, en dernière analyse, du rôle de la variable imputée dans le modèle de microsimulation. Pour conclure, une méthode d'imputation exacte qui produit des résultats valides est importante pour la microsimulation, parce que les données simulées ont éventuellement pour objet de fournir des projections des tendances de nombreux aspects de l'état de santé afin d'orienter les décisions en matière de politique. L'objectif de la présente étude était de présenter, de manière transparente et pertinente, des données sur la performance de deux techniques d'imputation qui pourraient être utilisées pour préparer des ensembles de données pour la microsimulation.

Remerciements

Les travaux de Deirdre Hennessy ont été menés grâce à une bourse de recherche postdoctorale offerte par l'équipe STAR (Simulation Technology in Advanced Research) financée par les Instituts de recherche en santé du Canada (IRSC). Douglas Manuel est titulaire d'une chaire en santé publique appliquée des IRSC et de l'Agence de la santé publique du Canada.

Bibliographie

Durrant, G.B. (2005), « Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review », document de travail, Royaume-uni : National Centre for Research Methods, Southampton Statistical Sciences Research Institute, University of Southampton.

Ellis, B. (2007), « A consolidated macro for iterative hot-deck imputation », document présenté au NorthEast SAS Users Group - 2007, Baltimore, Maryland.

Kalton, G. et D. Kasprzyk (1982), « Imputing for missing survey responses », *Proceedings of the Section on Survey Research Methods, American Statistical Association*, p. 22 à 31.

Khaw, K.T. et E. Barrett-Connor (1988), "The association between blood pressure, age, and dietary sodium and potassium: a population study", *Circulation*, vol. 77, n^{o} 1, p. 53 à 61.

Strazzullo, P, D'Elia, L., Kandala, N.B. et F.B. Cappuccio (2009), « Salt intake, stroke, and cardiovascular disease: meta-analysis of prospective studies », *BMJ*, 339, b4567.

SÉANCE 7A

MÉTHODES DE PROTECTION DE LA CONFIDENTIALITÉ ET OUTILS POUR L'ACCÈS AUX DONNÉES

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Méthodes de masquage de l'identité pour les fichiers de données sur la santé à grande diffusion

Khaled El Emam[1]

Résumé

Des quantités croissantes de données non nominatives sur la santé sont rendues publiques. Par exemple, aux États-Unis, les Centers for Medicare & Medicaid Services affichent maintenant en ligne les données sur les demandes de remboursement; en Californie, l'Heritage Provider Network a lancé un concours de portée mondiale dans le cadre duquel les concurrents devront développer un modèle pour prédire l'hospitalisation en utilisant des données sur la santé ayant trait à plus de 150 000 patients, et, au Canada, les Instituts de recherche en santé du Canada exigent maintenant la divulgation publique des données de niveau individuel provenant des essais cliniques qu'ils financent. La communication décrira une méthode quantitative de création de fichiers de données à grande diffusion sous diverses contraintes d'accès, pour la gestion des risques de divulgation de l'identité et des attributs. La méthode a été utilisée pour divulguer de grands ensembles de données sur la santé au cours des trois dernières années et l'on discutera d'exemples de problèmes qui se posent durant ces divulgations.

[1]Khaled El Emam, Université d'Ottawa, Canada.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Le système d'analyse des microdonnées du U.S. Census Bureau

Michael Freiman, Jason Lucero, Lisa Singh, Jiashen You, Michael DePersio et Laura Zayatz[1]

Résumé

L'article décrit un système d'analyse des microdonnées (MAS, *Microdata Analysis System*) développé à l'heure actuelle pour permettre aux utilisateurs de procéder à des analyses de données confidentielles du Census Bureau, telles que des totalisations croisées et des régressions, sans avoir accès aux microdonnées sous-jacentes. Ils peuvent effectuer des analyses portant sur un univers (sous-ensemble) de leur choix, sous certaines contraintes. Pour les univers acceptables, un sous-échantillon aléatoire d'observations comprises dans cet univers est supprimé de toutes les analyses ultérieures. Cette règle, appelée « règle de l'abandon de *q* observations » (*Drop* q *Rule*) joue un rôle crucial dans la protection des données tabulaires. Dans le cas de la régression, certaines autres règles concernent la façon dont les variables catégoriques sont traitées, ainsi que les interactions et les transformations permises. Toute régression proposée est vérifiée afin de s'assurer que la qualité de l'ajustement du modèle ne soit pas si élevée que cela crée un risque de divulgation. Des diagnostics respectant les règles de confidentialité sont également fournis, et nous sommes en train d'ajouter au système d'autres capacités, telles que la production de statistiques sommaires, d'histogrammes et de nuages de points, tous protégés comme il convient.

Mots clés : Divulgation ; système d'accès à distance ; totalisations ; régression ; données synthétiques.

1. Introduction

Le U.S. Census Bureau recueille ses données d'enquête et de recensement en vertu du Titre 13 du Code des États-Unis, qui lui interdit de diffuser des données qui permettraient de reconnaître celles fournies en vertu de ce Titre par tout établissement individuel ou tout particulier. La *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA) de 2002 exige aussi que soit protégée l'information recueillie ou acquise à des fins statistiques sous une promesse de confidentialité. Cependant, l'organisme a également la responsabilité de diffuser des données aux fins d'analyse statistique. Comme les autres organismes statistiques nationaux, notre objectif est de diffuser autant de données de haute qualité que possible sans enfreindre l'engagement de confidentialité.

Le présent article porte sur un système d'analyse de microdonnées (MAS, pour *Microdata Analysis System*) en cours de développement au U.S. Census Bureau. Le cadre du système a été décrit en grande partie dans Steel et Reznek (2005) et dans Steel (2006). Le système est conçu pour permettre aux utilisateurs des données d'exécuter diverses analyses statistiques (régressions, totalisations croisées, production de statistiques sommaires univariées ou bivariées, etc.) sur des microdonnées confidentielles provenant d'enquêtes et de recensements sans voir ou sans télécharger les microdonnées sous-jacentes.

À la section 2 nous donnons certains renseignements contextuels sur le MAS et sur les raisons qui ont motivé son développement. À la section 3, nous discutons de l'état actuel du système, y compris ses capacités et les règles qui protègent la confidentialité, en nous concentrant sur la régression et les tableaux. À la section 4, nous considérons certaines autres fonctions disponibles dans le système. À la section 5, nous examinons une autre approche de la création d'un système d'accès à distance. À la section 6, nous concluons par des commentaires sur les travaux de recherche à venir et la poursuite du développement du système.

[1]Michael Freiman, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 (michael.freiman@census.gov) ; Jason Lucero, Freddie Mac, 8200 Jones Branch Drive, McLean, VA 22102 ; Lisa Singh, Georgetown University Department of Computer Science, 329A St. Mary's Hall, Washington, DC 20057 ; Jiashen You, University of California-Los Angeles Department of Statistics, 8125 Math Sciences Bldg., Box 951554, Los Angeles, CA 90095 ; Michael DePersio, University of Delaware Department of Mathematical Sciences, 501 Ewing Hall, Newark, DE 19716; Laura Zayatz, U.S. Census Bureau. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et d'en favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau.

2. Contexte du MAS

Au U.S. Census Bureau et dans d'autres organismes statistiques à travers le monde, le problème de la confidentialité des données, ainsi que la demande croissante d'une gamme variée de produits de données personnalisables ont motivé la création de *systèmes d'accès à distance* en ligne qui permettent à l'utilisateur de demander une analyse statistique et de recevoir les résultats sans avoir un accès direct aux microdonnées sous-jacentes. Selon les particularités du système, le résultat peut être fondé sur des données perturbées et certaines demandes peuvent être rejetées afin de préserver la confidentialité des données. Le concept d'un système d'accès à distance n'est pas nouveau; en effet, l'idée d'un système permettant des demandes personnalisées a déjà été proposée à l'époque où le Système de stockage et d'extraction des données codées suivant une grille géographique a été décrit par Fellegi et Goldberg (1969).

Le système d'analyse des microdonnées (MAS) permettra au U.S. Census Bureau de donner accès à des renseignements plus complets et plus détaillés que ceux qui, autrement, seraient mis à la disposition de la plupart des utilisateurs. En outre, le système sera ouvert à quiconque souhaite l'utiliser, sans frais ni processus de demande, ce qui étendra l'accès aux personnes pour qui une visite à l'un de nos centres de données de recherche est impossible. Le MAS permettra, au départ, d'avoir accès aux données des enquêtes démographiques et des recensements décennaux, et nous espérons que, lorsqu'il sera développé davantage, le système sera également capable de traiter les données économiques. Nous commencerons par offrir des régressions et des totalisations croisées, et d'autres types d'analyses seront ajoutés dans l'avenir. Nous avons l'intention de tenir un registre de toutes les demandes entrées dans le système, mais non de l'identité des utilisateurs faisant les demandes. Ce registre n'aura pas d'incidence directe sur les données de sortie que fournit le système en réponse aux demandes enregistrées, mais il nous permettra de voir comment le système est utilisé, afin que nous puissions apporter de petites modifications au besoin en ce qui concerne la convivialité et la prévention de la divulgation.

Notre plan actuel, décrit dans Chaudhry (2007), consiste à offrir le MAS au moyen d'une interface Java dans le cadre du service en ligne gratuit DataFerrett du Census Bureau. Le MAS est doté d'une interface graphique qui permet aux utilisateurs de choisir les variables d'intérêt sur une liste. Dans le cas de la régression, les utilisateurs peuvent faire glisser les variables dans les modèles et, de quelques clics de souris, créer des interactions de variables et des transformations de certaines variables.

L'usage du MAS a notamment été suggéré comme moyen préliminaire d'examiner les données. Même si un chercheur souhaite exécuter une analyse plus complexe que ne le permet le MAS, une demande au MAS pourrait être faite afin d'obtenir certains renseignements élémentaires au sujet des variables d'intérêt. Cela pourrait aider le chercheur à décider s'il doit ou non entreprendre le processus plus coûteux et plus long d'obtention de l'accès à un centre de données de recherche. En outre, puisque l'accès à ces centres n'est accordé que sur présentation d'une candidature, l'utilisation du MAS pourrait permettre de soumettre une demande plus réfléchie.

3. Aperçu des règles de confidentialité du MAS

Plusieurs règles et procédures de protection de la confidentialité qui assurent le respect des normes de protection contre la divulgation sont programmées dans le logiciel du MAS. Nombre d'entre elles ont été conçues au Census Bureau, tandis que certaines ont été élaborées par le professeur Jerome Reiter à la Duke University. Ces règles et procédures ont pour objectif d'empêcher des intrus d'utiliser les résultats d'une ou de plusieurs demandes pour reconstruire les enregistrements de microdonnées individuels, entièrement ou partiellement.

3.1 Règles de confidentialité pour la formation de l'univers

Les utilisateurs du MAS peuvent choisir un univers, ou une sous-population, sur lequel seront exécutées les analyses. Dans le présent article, nous désignerons par U(n) un univers comportant n observations. Cette notation est ambiguë, mais l'univers qui présente un intérêt particulier sera généralement évident d'après le contexte. Le système donne un ensemble de variables et de niveaux de catégorie à partir desquels un utilisateur peut définir un univers en appliquant des énoncés de conditions aux variables. Par exemple, l'utilisateur peut choisir un univers constitué de la sous-population de l'ensemble des femmes dans une région géographique particulière. Un univers plus complexe pourrait

correspondre à l'ensemble des personnes de sexe masculin ou en chômage, ou à toutes les personnes dont le revenu exprimé en dollars se situe dans l'union de [9180,20155] et de [31662,43468], quoiqu'il faille admettre que le dernier de ces univers pourrait être d'une utilité douteuse (et, comme il est décrit plus loin, il est probablement inadmissible tel qu'il est présenté ici, à cause du rôle des seuils d'exclusion dans le système). Une des règles de confidentialité requiert que toutes les variables utilisées pour définir l'univers soient catégoriques.

Puisqu'un utilisateur peut vouloir définir un univers en se basant sur des variables qui sont intrinsèquement continues plutôt que catégoriques, chaque variable continue est accompagnée d'une liste des limites d'intervalle admissibles appelées *seuils*, qui sont décrites dans Lucero et coll. (2009). Quand un univers est défini en utilisant une variable numérique, les seules tranches de valeur de cette variable qui doivent être utilisées sont celles fondées sur cette liste prédéterminée. Par exemple, si la variable d'intérêt est le revenu brut corrigé et que les seuils sont 0 $, 10 000 $, 23 000 $, 35 000 $, 52 500 $ et 100 000 $, l'univers des personnes, dont le revenu est compris dans l'intervalle (23 000 $, 52 500 $] serait admissible, puisque les deux limites sont des seuils, tandis que l'univers des personnes, dont le revenu est compris dans l'intervalle (35 000 $, 38 000 $] ne le serait pas, car l'une des limites n'est pas un seuil. Cette contrainte vise à prévenir la divulgation due à une attaque par différence, qui est décrite plus en détail à la section 3.1.1. Ce genre de divulgation aurait lieu, par exemple, si un utilisateur demandait un tableau pour l'univers des personnes ayant un revenu de plus de 11 313 $ ainsi que le tableau correspondant pour l'univers des personnes ayant un revenu de plus de 11 314 $, puis comparait manuellement les deux tableaux. Si une seule personne de l'ensemble de données a un revenu exactement égal à 11 314 $, les autres attributs de cette personne pourraient être déduits facilement.

Les seuils sont attribués quand les données sont chargées dans le MAS et ils demeurent inchangés par après. Dans certains cas, le système attribuera à une variable plusieurs ensembles de seuils possibles, les seuils pertinents dans un cas donné étant déterminés par les autres caractéristiques de l'univers. En particulier, si la région géographique qui définit l'univers est élargie, la liste des seuils peut être étendue, afin que l'utilisateur ait accès à des intervalles plus fins, car les niveaux plus élevés d'agrégations géographiques diminuent le risque de divulgation d'enregistrements individuels.

D'autres exigences concernant les univers s'appliquent aussi aux variables intrinsèquement catégoriques ainsi qu'aux variables seuils, comme une taille minimale d'univers et les exigences que les cellules de certains tableaux associés à la définition de l'univers ne soient pas trop peu peuplées.

3.1.1 Divulgation due à une attaque par différence et suppression aléatoire d'enregistrements

Dans un système d'accès à distance, une préoccupation importante que soulèvent les tableaux tient à ce qu'un intrus pourrait identifier les attributs d'un répondant particulier (l'*observation cible*) en faisant la différence entre les résultats de l'analyse statistique obtenus pour deux demandes sur des univers similaires, méthode qui porte le nom d'*attaque par différence*. Le problème éventuel est qu'un intrus crée deux univers dans le MAS, $U(n)$ et $U(n-1)$, contenant tous deux les mêmes n observations à l'exception de l'observation cible qui manque dans le deuxième univers. Dans ce cas, $U(n) \setminus U(n-1) = U(1)$ est l'univers constitué de l'observation cible uniquement. Par exemple, supposons que l'intrus sache qu'il n'existe qu'un seul non-citoyen parmi les n résidents de la région d'intérêt. Il peut alors créer $U(n)$ et $U(n-1)$, où $U(n)$ est l'univers complet de personnes dans la région et $U(n-1)$ est l'univers constitué des citoyens qui vivent dans la région. Supposons que l'intrus demande alors deux totalisations croisées distinctes pour les mêmes variables de totalisation sous-jacentes; nous appelons ces deux tableaux T_n et T_{n-1}, comme l'illustre la figure 3.1.1-1. Les tableaux qui suivent montrent une attaque par différence fondée sur une totalisation selon l'âge (une classification binaire indiquant si la personne a au moins 45 ans ou non) en fonction du revenu (une classification binaire indiquant si le revenu est au moins égal à 50 000 $ ou non).

Figure 3.1.1-1
Exemple d'exécution d'une attaque par différence par soustraction de matrices

Tous les résidents		
T_n	<50 000 $	≥50 000 $
Âge<45	323	170
Âge≥45	45	58

-

Citoyens seulement		
T_{n-1}	<50 000 $	≥50 000 $
Âge<45	323	169
Âge≥45	45	58

=

Non-citoyens seulement		
T_1	<50 000 $	≥50 000 $
Âge<45	0	1
Âge≥45	0	0

L'intrus peut effectuer la soustraction des matrices T_n-T_{n-1}=T_1, où T_1 est un tableau à double entrée du sexe selon la situation d'emploi pour les non-citoyens, dont le nombre n'est que d'un. Comme le montre la figure 3.1.1-1, T_1 contient une cellule affichant une fréquence de 1 pour les personnes de moins de 45 ans ayant un revenu d'au moins 50 000 $, ce qui indique à l'intrus que le non-citoyen contenu dans U(1) possède ces deux caractéristiques. L'intrus a donc créé une divulgation au sujet du non-citoyen, en dépit du fait que les contraintes de création de l'univers dans le MAS ne donnent pas directement accès à l'univers des non-citoyens. En exécutant des attaques par différence semblables à celle qui vient d'être décrite, un intrus peut arriver à reconstruire l'enregistrement de microdonnées confidentielles au complet pour l'unique observation contenue dans U(1).

Une attaque par différence peut aussi être crainte lorsqu'il y a exactement deux unités qui ont une certaine caractéristique commune, par exemple être non-citoyen, et que l'intrus est l'une de ces deux unités. L'intrus peut alors construire l'univers complet $U(n)$ et la partie de l'univers constituée uniquement de citoyens $U(n-2)$. Puis, il peut se retrancher manuellement de $U(n)$ pour obtenir $U(n-1)$, c'est-à-dire l'univers de toutes les personnes sauf l'intrus. Autrement dit, $U(n-1)$ est l'univers constitué de tous les citoyens et du non-citoyen autre que l'intrus. Celui-ci peut alors procéder à une attaque par différence comme il est décrit plus haut en comparant $U(n-1)$ et $U(n-2)$ pour obtenir l'information sur l'autre non-citoyen.

Les vérifications effectuées par le système au moment de la sélection d'un univers aident à prévenir les attaques par différence réussies, mais le principal outil en vue de contrecarrer ce genre d'attaque est la *règle de l'abandon de q observations* (*Drop q Rule*). D'un univers défini par l'utilisateur qui a été accepté à toutes les vérifications, on retranche au hasard *q* enregistrements. Pour cela, le MAS commence par tirer une valeur entière aléatoire de *q* telle que $2 \leq q \leq k$ et que, quand l'univers est modifié en omettant *q* enregistrements, le nombre d'enregistrements restants soit un multiple de 3. Ici *k* est un nombre prédéterminé, qui dépend de la taille de l'univers. Ensuite, sachant *q*, le MAS sous-échantillonne l'univers $U(n)$ en supprimant au hasard *q* enregistrements de $U(n)$ pour produire un univers sous-échantillonné $U(n-q)$ qui sera utilisé pour toutes les analyses subséquentes. Le MAS ne produit pour chaque univers $U(n)$ qu'un seul univers sous-échantillonné $U(n-q)$, qui est utilisé pendant toute la durée de vie du système, afin qu'il soit impossible d'obtenir des renseignements supplémentaires au sujet de l'univers original en examinant à plusieurs reprises différents sous-échantillons.

Les attaques par différence les plus préoccupantes requièrent, entre autres, qu'il existe deux univers dont la taille diffère d'une ou de deux unités. Cependant, sous la règle de l'abandon de *q* observations décrite plus haut, tous les univers sous-échantillonnés ont une taille qui est un multiple de 3, et aucune paire de multiples de 3 (y compris les paires dont les deux chiffres sont les mêmes) ne peut contenir des chiffres qui diffèrent de 1 ou de 2. Donc, la règle d'abandon de *q* observations élimine la possibilité de cette forme de divulgation, voire même d'une de divulgation apparente de cette sorte. (Par « divulgation apparente », nous entendons une soustraction de matrices où la différence résultante ne contient aucune cellule ayant une valeur négative, et où la somme des valeurs sur l'ensemble des cellules est égale au nombre d'observations – dans le cas qui nous occupe 1 ou 2 – que l'intrus souhaite isoler. Dans ces conditions, l'intrus peut penser qu'il a obtenu une divulgation, même si celle-ci est inexacte.)

Un scénario moins probable est celui où un groupe de $j > 2$ personnes partage une caractéristique (ou une combinaison de caractéristiques) qui par ailleurs ne sont pas observées dans l'ensemble de données. Dans ce cas, $j-1$ de ces personnes pourraient conspirer contre l'unique autre personne pour créer deux tableaux, $\mathrm{U}(n)$ et $\mathrm{U}(n-j)$, puis se soustraire manuellement de $\mathrm{U}(n)$ pour obtenir $\mathrm{U}(n-(j-1))$, qui, combiné à $\mathrm{U}(n-j)$, peut être utilisé pour obtenir une divulgation au sujet de l'individu qui ne conspire pas. La règle de l'abandon de q observations rend peu probable le succès de ce genre d'entreprise. Cependant, la possibilité de réussite, voire même de réussite apparente, peut être éliminée en remplaçant l'exigence que la taille de l'univers sous-échantillonné soit un multiple de 3 par l'exigence qu'elle soit un multiple de m+1, où m est la valeur la plus grande de j pour laquelle ce genre de collusion est crainte. Il en est ainsi parce que le scénario de collusion requiert deux univers dont la taille diffère exactement d'une valeur j, mais qu'aucune paire de multiples de $m+1$ ne peut différer exactement de cette quantité. Cependant, les conditions nécessaires pour que ce genre d'attaque réussisse semblent tellement spécifiques que, pour le moment, nous sommes enclins à maintenir l'exigence que la taille de l'univers sous-échantillonné soit un multiple de 3, au lieu d'utiliser un multiple d'un autre nombre.

En plus des deux principaux types de divulgation par différence susmentionnés, une autre préoccupation tient à la possibilité qu'il existe un groupe relativement petit dont tous les membres partagent une même caractéristique qui diffère de celle(s) utilisée(s) pour définir le groupe. Par exemple, supposons qu'une attaque par différence indique que cinq anciennes combattantes de la guerre de Corée vivent dans une région donnée et que l'on examine leur état matrimonial. Si nous trouvons que deux d'entre elles sont mariées, deux sont divorcées et une est veuve, il n'existe aucune divulgation apparente au sujet d'une personne particulière. Cependant, si nous constatons que toutes les cinq sont divorcées, nous avons produit une divulgation apparente au sujet des cinq femmes. La règle de l'abandon de q observations est utile ici, en grande partie pour la même raison que les données d'échantillon sont intrinsèquement moins susceptibles de donner lieu à une divulgation que des données de recensement. Dans notre exemple, il n'est pas forcément vrai que toutes les anciennes combattantes de la Guerre de Corée soient divorcées, car une ou plusieurs d'entre elles non divorcées pourraient avoir été retranchées de l'ensemble de données en appliquant la règle de l'abandon de q observations. Les intrus devraient reconnaître qu'ils ne peuvent pas obtenir une divulgation jugée correcte en utilisant cette méthode, même dans les cas où les données de sortie du MAS indiquent une divulgation apparente.

3.2 Protection de la confidentialité pour les modèles de régression

Le MAS permet d'exécuter des régressions sur un univers admissible et comprend plusieurs règles additionnelles applicables spécifiquement à la régression. Le nombre de variables indépendantes dans une équation de régression est limité à 20. Les variables indépendantes et les variables réponses numériques peuvent être transformées, mais seules les transformations figurant sur une liste établie sont permises, afin d'empêcher l'utilisateur d'effectuer des transformations qui accordent délibérément trop d'importance aux valeurs aberrantes ou à d'autres observations particulières. À l'heure actuelle, les seules transformations permises sont le carré, la racine carrée et le logarithme naturel, mais cette liste sera probablement allongée.

Reznek (2003), ainsi que Reznek et Riggs (2004) décrivent un important risque de divulgation qui peut résulter de modèles de régression comprenant toutes les interactions entre variables dans lesquels seules des variables indicatrices sont utilisées comme variables indépendantes. Donc, nous avons imposé des contraintes sur les termes d'interaction : pas plus de trois variables ne peuvent interagir les unes avec les autres, et nous ne permettons pas la spécification de modèles comportant toutes les interactions possibles. L'utilisateur crée des interactions en cliquant sur les variables indépendantes déjà dans le modèle, de sorte que seules les interactions bidirectionnelles sont possibles si chaque variable mise en interaction apparait comme non mise en interaction dans le modèle. Une situation semblable existe pour les interactions tridirectionnelles, et quand le système crée une telle interaction, il crée également toutes les interactions bidirectionnelles correspondantes entre les paires de variables concernées. Les variables indépendantes catégoriques sont intégrées dans le modèle en utilisant des variables indicatrices pour toutes les catégories sauf la catégorie de référence, qui correspond à la catégorie la plus fréquente de la variable. Pour que des variables indicatrices soient associées à une catégorie d'une variable catégorique ou à ses interactions, cette catégorie doit posséder au moins un certain nombre m d'observations; si ce minimum n'est pas satisfait, toutes les variables indicatrices pour la catégorie en question sont omises du modèle. En fait, cela signifie que les catégories ne comportant que peu d'observations sont absorbées dans la catégorie de référence. Nous avons fixé au départ m=3, mais cette valeur peut être modifiée comme il est décrit plus bas.

Les régressions pour lesquelles la valeur de R^2 est élevée posent un autre risque de divulgation, car elles permettent d'estimer une variable d'après un enregistrement de microdonnées avec un haut degré d'exactitude si les valeurs des autres variables sont connues pour cet enregistrement. Par conséquent, le système ne produit pas ce genre de régression. Ce cas diffère quelque peu du contexte habituel de la régression, dans lequel une valeur élevée de R^2 est une situation souhaitable, alors qu'ici elle est considérée comme posant problème. Il se pourrait aussi qu'il existe une variable indicatrice telle que, chaque fois que sa valeur est égale à 1, la valeur de l'observation correspondante est prédite avec une très grande exactitude par la régression. Cette variable indicatrice peut provenir d'une variable indépendante catégorique ou de l'intersection de catégories de variables indépendantes catégoriques que l'on fait interagir. Les résultats des régressions présentant cette caractéristique ne seront pas fournis à l'utilisateur non plus; on pourrait se représenter cette situation comme une vérification de la qualité locale de l'ajustement pour compléter la vérification du R^2 sur la qualité globale de l'ajustement. En outre, les données de sortie ne seront pas fournies s'il existe une variable indicatrice qui prend très peu de fois une valeur de 1 dans l'ensemble de données.

Quand des variables catégoriques sont utilisées comme variables indépendantes, les règles susmentionnées peuvent être très contraignantes, surtout si elles sont appliquées à des ensembles de données relativement petits ou que l'on fait interagir des variables catégoriques, ce qui rend peu probable le fait que le système produise les données de sortie souhaitées. Puisque notre objectif est de fournir des données de sortie dans la mesure du possible, nous apportons une légèrement modification à la régression dans un tel cas. Pour cela, nous augmentons la borne inférieure *m* du nombre d'observations qu'une catégorie doit contenir pour ne pas être intégrée dans la catégorie de référence. En absorbant un plus grand nombre de catégories dans la catégorie de référence, nous espérons atténuer les conditions qui empêchent de produire les résultats de la régression. Le MAS continue d'augmenter la valeur de *m* jusqu'à ce qu'il trouve une régression dont les résultats peuvent être produits sans poser de risque, auquel cas la régression est ajustée, ou jusqu'à ce que la valeur de *m* soit suffisamment grande pour qu'une des variables indépendantes catégoriques soit réduite à ne posséder qu'un seul niveau, tous les autres niveaux étant absorbés dans le niveau de référence, ce qui mène le système à refuser la sortie des données.

Un inconvénient de notre approche courante est qu'elle donne parfois lieu à la combinaison de catégories de façon indésirable. En particulier, la méthode que nous avons décrite plus haut ne tient compte d'aucune structure ordinale éventuellement présente. Par exemple, si une variable indépendante est une variable catégorique décrivant le plus haut niveau d'études atteint, il est possible que la catégorie de référence contienne les personnes dont le diplôme de niveau le plus élevé est un diplôme d'école secondaire, un diplôme menant à un grade d'associé ou un diplôme de maîtrise, tandis que les personnes dont le diplôme de niveau le plus élevé est un baccalauréat se retrouvent dans une catégorie distincte, ce qui n'a intuitivement aucun sens. Nous espérons améliorer cet aspect du système dans l'avenir.

Si toutes les exigences en vue d'exécuter une régression sont satisfaites, avant ou après avoir ajusté le paramètre *m*, le MAS transmet les données de sortie à l'utilisateur. À l'heure actuelle, ces données de sortie comprennent les coefficients de régression, leurs erreurstypes, les statistiques t et les valeurs p, la statistique F pour la régression et sa valeur p, le coefficient R^2 pour la régression, ainsi qu'un tableau d'analyse de variance (ANOVA). Tous ces résultats sont arrondis afin de contrecarrer toute attaque fondée sur des valeurs exactes des coefficients de régression obtenues d'après un grand nombre de régressions. Pour la régression par les moindres carrés ordinaires (MCO), les coefficients, les erreurs types, les statistiques t et les valeurs p sont donnés à l'heure actuelle avec quatre chiffres significatifs.

Le MAS a également la capacité d'exécuter une régression logistique (binaire ou multinomiale) quand la variable réponse est catégorique, ou une régression de Poisson quand la variable réponse est un compte, et les règles qui s'appliquent à la régression par les MCO sont alors adaptées au nouveau contexte. Les contraintes appliquées aux interactions et l'approche concernant les variables indépendantes catégoriques sont les mêmes. Pour déterminer si une régression doit être refusée (ou si *m* doit être augmenté), nous utilisons des mesures du pseudo-R^2 pour déterminer si la qualité globale de l'ajustement est trop élevée; le cas échéant, les résultats de la régression ne seront pas fournis ou il faudra augmenter la valeur de *m*. Nos mesures de la qualité locale de l'ajustement sont un peu différentes. Pour la régression logistique, la régression peut être refusée (ou *m* augmenté) en se fondant sur les observations figurant dans l'ensemble de données dont les probabilités prédites d'être dans une catégorie particulière de réponse (selon le modèle) sont proches de 1. Une règle similaire s'applique à la régression de Poisson, et nous poursuivons nos évaluations afin de savoir si elle offre une protection suffisante. Comme auparavant, les coefficients estimés arrondis, ainsi que leurs erreurstypes, les statistiques de test et les valeurs p sont fournis, de même que le tableau d'analyse de la déviance dans le cas de la régression de Poisson et de la régression binaire ou multinomiale.

Bien que les diagnostics de régression soient utiles pour déterminer si une régression par les MCO décrit adéquatement les données, ce genre de diagnostic, en particulier les graphiques des résidus non perturbés, pose un risque de divulgation, car ils permettent à l'intrus de déterminer l'information au sujet de points individuels. Par conséquent, le système crée des graphiques des résidus fondés sur des résidus synthétiques, des variables indépendantes synthétiques et des valeurs prédites. Notre approche de la régression par les MCO suit de près la méthode décrite dans Reiter (2003). Pour la régression logistique, binaire ou multinomiale, le système produit également des graphiques diagnostics conformément à la méthode décrite dans Reiter et Kohnen (2005). Nous fournissons aussi les diagrammes quantile-quantile et les statistiques de test pour évaluer la normalité des résidus. Les diagrammes quantile-quantile sont fondés sur des données synthétiques, car autrement, les points individuels des diagrammes pourraient donner lieu à la divulgation des données. Toutefois, la nature de la méthode de synthèse des données devrait produire des résidus synthétiques qui paraissent plus normaux que les données réelles. Cependant, les statistiques des tests sont fondées sur les données réelles, de sorte que nous recommandons aux utilisateurs d'évaluer la normalité en se basant sur les statistiques de test, puis d'utiliser les diagrammes quantile-quantile pour évaluer la nature de tout écart par rapport à la normalité. L'un des défauts de nos diagrammes diagnostiques est que, étant donné que les contraintes appliquées aux analyses qui peuvent être exécutées dans le système, il pourrait n'exister aucun moyen de corriger les problèmes révélés.

4. Fonctions additionnelles

Les fonctions additionnelles en cours de développement comprennent la production d'histogrammes et de nuages de points pour les données numériques. Il existe, dans les deux cas, un risque de divulgation si les données ne sont pas perturbées.

4.1 Histogrammes

Les histogrammes ne semblent pas poser de risque de divulgation important, sauf en présence de valeurs aberrantes. La règle d'abandon de *q* observations offre déjà une certaine protection contre la divulgation, mais nous adaptons la méthode utilisée pour créer les histogrammes afin d'accroître la protection. Dans le cas d'un histogramme, la principale crainte est qu'il pourrait être utilisé pour trouver les valeurs aberrantes de la variable représentée graphiquement.

Nous commençons par éliminer de la distribution toutes les valeurs extrêmes. Ensuite, nous utilisons une fonction de densité estimée par la méthode du noyau pour trouver une estimation lissée de la distribution de la variable. Puis, nous tirons un échantillon de la distribution lissée égale en taille à l'ensemble de données original. Notons qu'en raison du lissage, les bornes de la densité estimée se situent au-delà des bornes des données observées, de sorte qu'il est possible d'étendre l'histogramme au-delà des données originales. Cependant, si une observation est un peu plus extrême que les autres, mais peut-être pas extrême au point d'être exclue en tant que valeur aberrante, il se peut, lorsque l'on effectue le tirage à partir de la distribution lissée, qu'aucune partie de l'échantillon ne provienne de la région autour de cette observation, de sorte que l'histogramme peut également s'étendre moins loin que les données réelles dans cette direction. Le caractère discret des classes de l'histogramme joue aussi le rôle de perturbation *de facto* des données.

Afin de protéger encore davantage les valeurs inhabituelles, nous exigeons que toute classe d'un histogramme contienne au moins trois observations. Des observations sont ajoutées aux classes qui en contiennent moins de trois afin qu'elles en contiennent trois. Les classes contenant trois observations (après l'ajout) sont colorées en rouge quand l'histogramme est tracé, tandis que les autres classes sont colorées en gris.

Nous continuons de mettre la méthode de création des histogrammes à l'essai et à la modifier afin d'être certain qu'elle ne crée pas de risque de divulgation. Nous avons également testé la procédure de production d'histogrammes pour déterminer dans quelle mesure les histogrammes synthétiques imitent les histogrammes réels. L'une des préoccupations est de savoir ce qui constitue une « valeur aberrante extrême ». Nous omettons tous les points qui dévient de plus de quatre écarts-types de la moyenne. Cependant, dans les distributions asymétriques, cela pourrait éliminer certains points qui ne sont pas des valeurs aberrantes.

4.2 Nuages de points et boîtes à moustaches côte à côte

Nous envisageons diverses approches en vue d'arriver à mettre à l'épreuve de la divulgation un nuage de points de deux variables numériques.

Une approche consiste à utiliser la même méthode que pour les résidus synthétiques. Un inconvénient éventuel est que cette méthode traite les deux variables de manière asymétrique, de sorte qu'un diagramme synthétique de *y* en fonction de *x* ne doit pas nécessairement ressembler à un diagramme synthétique de *x* en fonction de *y*. Dans le cas d'un diagramme des résidus, cette asymétrie entre les variables est naturelle, mais dans un nuage de points plus général, nous pourrions souhaiter que les deux variables soient traitées de la même manière.

Une autre approche consiste à appliquer une méthode qui a pour point de départ le nuage de points réel – ou, si celui-ci contient un trop grand nombre de points, un sous-ensemble de points – puis à déplacer chaque point d'une distance aléatoire dans une direction aléatoire. Ceux qui ne nécessitent qu'une protection relativement faible ne sont déplacés que d'une petite distance, tandis qu'un point aberrant, même modestement, sera déplacé davantage. Nous considérons deux variantes de cette approche, dont une est décrite dans You (2010).

Une préoccupation éventuelle en ce qui concerne cette méthode concerne les valeurs extrêmement aberrantes, car elles peuvent être déplacées considérablement dans n'importe quelle direction. Dans une simulation d'une des méthodes envisagées, nous avons utilisé un ensemble de données de 1 001 points. Les 1 000 premiers avaient une distribution $X \sim \text{iid } N(0,1)$ et $Y=X^2$, tandis que le 1 001^{e} point était une valeur aberrante, pour laquelle $X=4$ et $Y=16$. Ce point se situait dans le coin tout à fait en haut et à droite du graphique et était assez éloigné de tout autre point, particulièrement verticalement. Par conséquent, il pouvait être déplacé considérablement. Cela a donné des résultats peut-être inattendus : parfois, le point demeurait un point aberrant dans le coin supérieur droit, son caractère extrême étant accru ou réduit, mais parfois, il se déplaçait dans une partie entièrement différente du graphique, devenant, par exemple, un point aberrant dans la région moyenne supérieure ou même dans la région gauche supérieure. Donc, si aucune modification n'est apportée à cette méthode, il pourrait être nécessaire d'avertir les utilisateurs que la présence d'un point très éloigné des autres dans un graphique est une preuve qu'il s'agit d'une valeur aberrante, mais qu'elle pourrait ne donner aucun renseignement utile au sujet de la position de cette valeur aberrante.

Sparks et coll. (2008) utilisent une méthode fondée sur des boîtes à moustaches côte à côte pour remplacer les diagrammes des résidus et les nuages de points ordinaires, et nous considérons également cette approche pour le MAS. Lorsque l'on utilise cette méthode, la variable x est divisée en classes et un nuage de points de la variable y pour chaque classe x est produit, puis les nuages de points sont tracés côte à côte. Si certaines précautions sont prises, comme la winsorisation des données pour protéger les valeurs aberrantes, il est possible de réduire au minium le risque de divulgation. Sparks et coll. soutiennent que, dans de nombreux cas, les boîtes à moustaches côte à côte présentent non seulement moins de risque de divulgation que les nuages de points, mais sont également plus utiles pour l'utilisateur.

4.3 Statistiques descriptives et tests

Le MAS calcule quelques statistiques descriptives de base, et nous prévoyons augmenter leur nombre à mesure que le développement du système progressera. Nous avançons relativement lentement à ce sujet parce que nous voulons être certains que personne ne puisse manipuler les statistiques descriptives pour arriver à une divulgation. Cela paraît toutefois peu probable, car dans la plupart des cas les statistiques descriptives (à l'exception des quantiles) en disent peu au sujet des observations individuelles quand l'ensemble de données n'est pas très petit. Les statistiques sont arrondies de manière appropriée. Le MAS exécute aussi des tests t sur la moyenne d'une variable, et fournit les intervalles de confiance à 95 % pour la moyenne. Nous examinons encore la question de savoir dans quelle mesure l'utilisateur devrait être capable de déterminer le niveau de confiance.

5. Une autre approche – La *Luxembourg Income Study*

Des systèmes d'accès à distance offrant parfois plus de versatilité que le MAS ont été développés, mais au prix d'être plus difficiles à maintenir et à protéger. Un exemple important est le système mis en œuvre par le groupe de la *Luxembourg Income Study* (LIS), un institut de recherche qui recueille des données sur le revenu, la richesse et diverses autres mesures, fondé en 1983 (voir Luxembourg Income Study, 2009a). Les données de la LIS sont une agrégation des données d'enquêtes-ménages menées par divers pays contributeurs. Le système d'accès à distance de la LIS – appelé LISSY – permet aux utilisateurs inscrits de soumettre, par courriel ou au moyen d'un formulaire en ligne, leur propre code qui peut être écrit en SAS, SPSS ou Stata. Les données de sortie, si elles sont jugées admissibles, sont envoyées par courriel et peuvent être visionnées sur le formulaire. Le système interdit certaines commandes qui pourraient être utilisées pour obtenir une divulgation concernant un particulier ou un ménage. Sont également interdites les séquences de commandes et/ou de variables qui aboutiraient à la violation des règles de confidentialité des données; ces séquences, ainsi que les demandes produisant des sorties excessivement longues, sont signalées pour l'analyse manuelle ou sont tout bonnement refusées. D'autres renseignements sont donnés dans Luxembourg Income Study (2009b). Schouten et Cigrang (2003) mentionnent aussi que la LIS contient un répertoire des tâches soumises, qui peuvent ainsi être évaluées plus en profondeur afin de s'assurer que les données sont utilisées correctement.

6. Travaux à venir

Nous testerons bientôt les règles de confidentialité mises en œuvre dans le prototype bêta du MAS pour confirmer qu'elles sont suffisamment strictes pour fournir la protection requise de la confidentialité.

Le système actuel ne contient aucun mécanisme pour traiter les valeurs manquantes, qui sont souvent présentes dans une enquête telle que l'American Community Survey, pour laquelle les répondants sont parfois disposés à répondre à une partie du questionnaire, mais ne souhaitent pas divulguer la réponse à chaque question.

Une autre préoccupation tient à la différence entre les enquêtes et les recensements. Puisque la plupart des ensembles de données sur lesquels est exécuté le MAS sont fondés sur des enquêtes, des méthodes supplémentaires doivent être mises en place pour tenir compte de ce fait. Un avantage de l'utilisation de données d'enquête est que le simple fait d'échantillonner la population fournit une protection considérable des données, puisqu'une unité qui est unique dans l'enquête à certains égards peut avoir plusieurs correspondants dans l'ensemble de la population. Cependant, jusqu'à présent, notre méthodologie n'a pas pris en compte les poids de sondage qui sont nécessaires pour faire des inférences correctes d'après une analyse et d'autres travaux seront nécessaires afin d'intégrer ces poids dans le système et de s'assurer que les analyses, surtout les totalisations, sont faites de manière que les poids ne puissent pas être déterminés, puis utilisés pour révéler des renseignements sensibles au sujet des entités pondérées. Dans le cas des enquêtes auprès des établissements, il existe aussi le risque qu'un établissement dominant se distingue du lot de telle façon que cela puisse causer une divulgation. Cela pourrait se produire si un établissement est beaucoup plus grand que les autres dans une cellule d'un tableau, ou dans une moindre mesure, mais en pouvant encore poser problème, dans une régression dans laquelle un point de ce genre pourrait être un point influent.

En outre, nous prévoyons créer un ensemble de règles de confidentialité pour les totalisations croisées et ajouter différents types d'analyses statistiques dans le système, dont un ensemble élargi de statistiques descriptives et de tests de signification. Nous aimerions aussi ajouter des variantes de la régression, telles que la régression pas-à-pas ascendante ou descendante. Cependant, cela posera certains défis, car le choix des variables à chaque pas devra être intégré aux règles sur les variables indépendantes catégoriques.

Bibliographie

Chaudhry, M. (2007), « Overview of the Microdata Analysis System », Statistical Research Division internal report, Washington DC: U.S. Census Bureau.

Fellegi, I.P. et S.A. Goldberg (1969), *Some Aspects of the Impact of the Computer on Official Statistics*, Ottawa: Dominion Bureau of Statistics.

Keller-McNulty, S. et E. Unger. (1998), « A Database Prototype System for Remote Access to Information Based on Confidential DataConference of European Statisticians », *Journal of Official Statistics*, 14, p. 347 à 360.

Luxembourg Income Study (2009a). *LIS Micro-data Access*, http://www.lisproject.org/data-access/lissy.htm. Accessed December 24, 2011.

Luxembourg Income Study (2009b). *LIS Micro-data Access – Job Syntax*, http://www.lisproject.org/data-access/lissy-syntax.htm. Accessed December 24, 2011.

Lucero, J., Zayatz, L. et L. Singh (2009), « The Current State of the Microdata Analysis System at the Census Bureau », *Proceedings of the American Statistical Association, Government Statistics Section*.

Reiter, J. (2003), « Model Diagnostics for Remote Access Regression Servers », *Statistics and Computing*, 13, p. 371 à 380.

Reiter, J. et C. Kohnen (2005), « Categorical Data Regression Diagnostics for Remote Access Servers », *Journal of Statistical Computation and Simulation*, 75, p. 889 à 903.

Reznek, A. (2003), « Disclosure Risks in Cross-Section Regression Models » *Proceedings of the Section on Government Statistics, JSM*.

Reznek, A. et T. Riggs (2004), « Disclosure Risks in Regression Models: Some Further Results », *Proceedings of the Section on Government Statistics, JSM*.

Schouten, B. et M. Cigrang (2003), « Remote Access Systems for Statistical Analysis of Microdata », *Statistics and Computing*, 13, p. 381 à 389.

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T. et D. McAullay (2008), « Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics® », *Computer Methods and Programs in Biomedicine*, 91, p. 208 à 222.

Steel, P. et A. Reznek (2005), « Issues in Designing a Confidentiality Preserving Model Server », *Monographs of Official Statistics*, 9, p. 29 à 36.

Steel, P. (2006), « Design and Development of the Census Bureau's Microdata Analysis System: Work in Progress on a Constrained Regression Server », présenté à la Federal Committee on Statistical Methodology Policy Seminar, Washington DC.

You, J. (2010), « Data-Driven Quality-Preserving Methods for Synthesizing Microdata on a Remote-Access Regression Server », unpublished report, Washington DC: U.S. Census Bureau.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Accorder l'accès aux microdonnées à des fins statistiques – Expérience de l'Australian Bureau of Statistics concernant les serveurs d'analyse à distance

James O. Chipperfield, Frank Yu et Melissa Gare[1]

Résumé

De nombreux organismes statistiques nationaux cherchent à améliorer leur stratégie de diffusion des données qu'ils produisent en facilitant l'accès aux microdonnées à l'aide de serveurs d'analyse à distance. Sous cette approche, les résultats des analyses statistiques ou des totalisations de données sont diffusés sous une forme qui ne permet pas de relier les microdonnées à des individus ou organismes particuliers. L'Australian Bureau of Statistics (ABS) élabore un service d'accès à distance qui lui permettra de répondre aux demandes de tableaux de fréquences et de résultats d'analyses au moyen de modèles statistiques soumises par les utilisateurs, tout veillant au respect strict de la confidentialité des renseignements contenus dans les microdonnées fournies par les répondants individuels. Le présent article donne un aperçu d'une méthode examinée à l'heure actuelle en vue d'assurer le respect de cette confidentialité. Elle comprend le contrôle des demandes et la perturbation des données de sortie.

Mots clés : Confidentialité ; serveur d'analyse à distance ; perturbation.

1. Introduction

Les organismes statistiques recueillent de très grandes quantités de microdonnées provenant de recensements, d'enquêtes par sondage et de sources administratives qui peuvent être utilisées pour élaborer et évaluer des politiques bénéfiques ou utiles à la société. Par conséquent, la demande d'accès à ce genre de microdonnées est importante chez les analystes tant du secteur public que des universités. Lorsqu'il accorde l'accès à ces microdonnées, l'organisme statistique est souvent tenu légalement de veiller à ce que le risque de divulgation de renseignements au sujet d'un particulier ou d'un organisme soit acceptablement faible. La gestion du risque de divulgation est ordinairement appelée contrôle de la divulgation statistique (CDS). Même après avoir supprimé des microdonnées les renseignements d'identification personnels, tels que le nom et l'adresse, le risque de divulgation persiste (voir, par exemple, Willenborg et De Waal, 2001).

Les méthodes de CDS applicables aux microdonnées comprennent la réduction du niveau de détail, le remplacement de valeurs réelles par des valeurs synthétiques (voir par exemple Reiter, 2002), le sous échantillonnage, la microagrégation, la permutation d'attributs entre enregistrements, ainsi que la perturbation des valeurs catégoriques. Mathews et Harel (2011), ainsi que Duncan et Pearson (1991) ont bien résumé nombre de ces méthodes, ainsi que quelques autres.

Un moyen possible d'améliorer le compromis entre l'utilité et le risque de divulgation consiste à utiliser un serveur d'analyse à distance. Voici un modèle simple de ce genre de serveur :

A. Un analyste soumet une demande, par Internet, au serveur d'analyse de l'organisme statistique.
B. Le serveur d'analyse exécute la demande de l'analyste sur les microdonnées délicates. Le produit statistique (par exemple coefficients de régression) résultant de la demande est modifié aux fins du CDS. La diffusion de certaines données de sortie est parfois restreinte parce qu'elles pourraient permettre à un analyste de reconstruire les attributs d'un enregistrement particulier.
C. Le serveur d'analyse envoie les données de sortie modifiées, par Internet, à l'analyste.

Les serveurs d'analyse à distance donnent aux utilisateurs le contrôle sur les données de sortie particulières qu'ils souhaitent extraire d'un ensemble de données. Il s'agit d'une évolution fondamentale du processus en vue de passer

[1]James O. Chipperfield, Frank Yu et Melissa Gare, Australian Bureau of Statistics, ABS House, Belconnen, ACT 2614, Australie, james.chipperfield@abs.gov.au , frank.yu@abs.gov.au et m.gare@abs.gov.au.

du paradigme classique, en vertu duquel ce sont les organismes statistiques nationaux qui décident de toutes les données de sortie qui pourront être diffusées, à un paradigme où les utilisateurs peuvent spécifier ce dont ils ont besoin, et quand et comment ils en ont besoin. Le défi des organismes statistiques est d'assurer le CDS pour les divers produits possibles.

Voici certains avantages offerts par un serveur d'analyse à distance :

- bien que le produit statistique soit modifié, il est fondé sur des microdonnées réelles, ce qui signifie que les relations complexes entre les microdonnées sont essentiellement préservées ;
- la mesure dans laquelle un produit est modifié peut dépendre du produit proprement dit. Ainsi, des estimations à un niveau très agrégé peuvent nécessiter proportionnellement moins de modifications que des estimations à un fin niveau de détail, par exemple au niveau du petit domaine. Puisqu'un analyste ne peut voir les attributs d'aucun enregistrement, moins de modifications qu'il n'en faudrait autrement sont requises ;
- l'effet de la modification sur le produit peut être indiqué de façon générale à l'analyste. Si l'effet est important, l'analyste peut décider d'ignorer tout bonnement les résultats ;
- une fois que le serveur est configuré, il peut traiter de multiples analyses en temps réel ;
- tous les programmes soumis peuvent être journalisés et vérifiés. Si une vérification mène à la conclusion qu'une tentative de divulgation est faite, l'organisme peut annuler l'accès de l'analyste au serveur et intenter une action en justice.

Voici certains inconvénients associés à un serveur d'analyse à distance :

- certains produits statistiques sont parfois agrégés (par exemple, les diagrammes des résidus au niveau de l'enregistrement peuvent être remplacés par des boîtes à moustaches) ou perturbés (par exemple coefficients de régression), et d'autres peuvent être tout bonnement restreints ;
- l'analyste peut être contraint de n'utiliser que les techniques d'analyse prises en charge par le serveur ;
- l'analyse peut prendre plus de temps que si l'analyste disposait des microdonnées sur son ordinateur personnel.

Certaines études ont été menées sur la gestion des risques de divulgation des données de sortie des analyses et des totalisations (c'est-à-dire le point B susmentionné). En ce qui concerne les résultats d'analyse, voir Gomatam et coll. (2008), Lucero et Zayatz (2010), Bleninger et coll. (2010) ainsi que Sparks et coll. (2008), et en ce qui concerne les résultats de totalisation, voir Shlomo (2007). L'objectif, dans ces études, est de protéger les données contre les attaques, qui comprennent le cas où un analyste se sert des données de sortie d'un serveur d'analyse pour reconstruire les attributs d'un ou de plusieurs enregistrements, lesquels, en cas de réussite, pourraient être utilisés pour essayer de divulguer des renseignements personnels par appariement avec d'autres microdonnées.

La section 2 décrit l'expérience et les plans de l'ABS en ce qui concerne les serveurs d'accès à distance. La section 3 décrit la méthode de gestion des risques de divulgation pour les tableaux de fréquences de l'ABS et la section 4, l'approche adoptée pour protéger les données de sortie des analyses. La méthode de protection des tableaux de mesures continues est en cours de mise en œuvre et ne sera pas discutée dans le présent article.

2. Expérience de l'ABS concernant les serveurs d'accès à distance

En 2002, l'ABS a lancé le Remote Access Data laboratory (RADL), qui est un service sécurisé de demande de données en ligne auquel les clients autorisés ont accès par la voie du site Web de l'ABS. Dans le RADL, les utilisateurs soumettent leurs demandes dans les langages statistiques SAS, STATA ou SPSS. Ces demandes sont exécutées sur un fichier de microdonnées dont la confidentialité est fortement protégée, appelé Confidentialised Unit Record File (CURF). Les résultats des demandes sont vérifiés automatiquement par le système et les données de sortie acceptées sont mises à la disposition des utilisateurs sur leur poste de travail. Les microdonnées sous-jacentes sont gardées en sécurité dans l'environnement de l'ABS et ne peuvent pas être visualisées. Contrairement aux CURF de base, qui sont diffusés sur cédérom aux utilisateurs approuvés qui les utilisent sur leur propre ordinateur, l'introduction du RADL a permis à l'ABS de mettre des microdonnées plus détaillées à la disposition des chercheurs.

Le Census Table Builder (CTB), un serveur à distance qui diffuse des tableaux de fréquences, a été lancé en 2009 et conçu principalement pour le Recensement de la population de 2006. Le CTB utilise la suite de programmes SuperSTAR de Space-Time Research (STR) et intègre la routine de protection de la confidentialité par perturbation dynamique de l'ABS. Cette routine, qui est exécutée sur les microdonnées anonymisées durant la production des tableaux demandés, est décrite à la section 3.1 du présent article, ainsi que dans Fraser et Wooton, 2005.

L'ABS a commencé à développer un nouveau serveur d'accès à distance appelé Remote Execution Environment for Microdata (REEM). Les composantes de base du REEM seront des générateurs de tableaux (*table builders*) qui s'appuient sur la méthode de perturbation élaborée pour le CTB dans lesquels sont intégrées d'autres améliorations apportées à la suite SuperSTAR, ainsi que sur un serveur d'analyse. Plusieurs facteurs dictent ces nouveaux travaux de développement. L'ABS doit répondre aux utilisateurs de plus en plus nombreux qui demandent un accès souple à de riches ensembles de microdonnées sur les ménages et sur les entreprises, y compris à un besoin croissant d'analyses d'ensembles de données administratives et d'ensembles de données appariées. Ces demandes ne peuvent être traitées par le RADL, qui est le service d'accès à distance existant. Le RADL s'appuie sur un fichier de microdonnées rendu confidentiel au préalable, appelé Confidentialised Unit Record File (CURF), et ne fournit pas toujours les renseignements à un niveau de détail suffisamment fin. En outre, l'effet que le CDS peut avoir sur les données de sortie analytiques obtenues d'après les CURF préoccupe les utilisateurs. La croissance du risque d'identification due à une augmentation de la puissance informatique (matérielle ainsi que logicielle) et la prolifération des ensembles de données externes détaillés menacent aussi la viabilité de l'approche fondée sur le RADL existant.

Il existe d'autres bonnes raisons de développer un serveur d'accès à distance pour remplacer le RADL. Premièrement, le projet répond à l'objectif de l'ABS d'accroître sa capacité et d'améliorer continuellement son efficacité. Notamment, l'évaluation et la production des CURF pour le RADL présentent un certain nombre d'inefficacités. Le remplacement du RADL par le REEM devrait, en principe, réduire considérablement ou éliminer le temps consacré par les employés à l'une et l'autre de ces activités. En particulier, le REEM comprendra des routines de protection de la confidentialité adaptées à chaque type d'analyse pris en charge par le REEM. Ces routines seront conçues pour être exécutées dynamiquement sur les microdonnées anonymisées, ce qui éliminera le besoin d'évaluer et de produire des CURF pour le futur serveur d'accès à distance.

Deuxièmement, l'un des facteurs qui motive le projet est d'accroître l'accès aux données de sortie de l'ABS. Celui-ci adoptera pour le REEM les normes concernant les métadonnées internationalement reconnues, y compris l'utilisation des normes DDI/SDMX et d'interfaces machine-machine (API), pour faciliter la découverte des éléments de données et la diffusion de données de sortie au moyen de services en ligne.

La première version du générateur de tableaux pour les microdonnées d'enquête sociale pondérées, appelée Survey Table Builder, a été diffusée en décembre 2011 pour un ensemble restreint de microdonnées. Le lancement limité d'une version d'essai du serveur d'analyse est prévu au milieu de 2012.

3. Fréquences

Les cellules d'un tableau sont soit *internes* soit *marginales*. La fréquence d'une cellule marginale est égale à la somme de deux autres fréquences ou plus qui figurent dans le tableau. Si une cellule n'est pas marginale, il s'agit d'une cellule interne. Nous allons maintenant décrire la méthode appliquée par l'ABS pour perturber les fréquences non pondérées (section 3.1) et pondérées (section 3.2) pour les cellules internes et les cellules marginales d'un tableau.

3.1 Générateur de tableaux de données du recensement (*Census Table Builder*)

Nous décrivons ici la méthode de perturbation des fréquences non pondérées mises en œuvre dans le *Census Table Builder* (CTB), un serveur d'accès à distance de l'ABS qui permet aux analystes de demander à distance le calcul de tableaux de contingence d'après les microdonnées du recensement de l'Australie. Les tableaux perturbés sont envoyés automatiquement à l'analyste, en général sans aucune intervention des employés de l'ABS. L'analyste peut définir les dimensions du tableau et les attributs des enregistrements intervenant dans le tableau, en n'étant soumis

qu'à des contraintes limitées (seuls les tableaux contenant un pourcentage élevé de cellules dont la fréquence est égale à 0 ou à 1 ne sont pas diffusés).

Désignons la i^{e} fréquence d'échantillon non pondérée pour une cellule *interne* d'un tableau de contingence par $n_i = \sum_{j=1}^{n} \delta_{ij}$, où $i = 1, \ldots, \cdots \mathrm{C}, \delta_{ij} = 1$ si le j^{e} enregistrement de microdonnées appartient à la i^{e} cellule et $\delta_{ij} = 0$ autrement, $j = 1, 2, ..., n$ et $n = \sum_{i=1}^{n} n_i$. Le CTB communique n_i^* à l'analyste au lieu de n_i, où

$$n_i^* = n_i + e_i^* + a_i^*,$$

$n_i^* \geq 0$, $\left|e_i^*\right| \leq L_e$, $\left|a_i^*\right| \leq L_a$, et L_e et L_a sont des entiers positifs spécifiés par l'organisme. Clairement, la différence entre n_i et n_i^* est contrainte d'être inférieure à $L = L_a + L_e$. Les e_i^* représentent la perturbation entière aléatoire de la i^{e} fréquence de cellule. Les a_i^* sont dérivés de manière que les fréquences internes et marginales soient cohérentes et que les variations des fréquences marginales soient bornées (pour des renseignements détaillés, voir l'annexe).

Soit $Var_*(\ )$ et $E_*(\ )$ la variance et l'espérance par rapport à la distribution de la perturbation e_i^*, qui satisfont aux critères suivants :

a) $E_*\left(e_i^*\right) = 0$

b) $Var_*\left(e_i^*\right) = \sigma^2$

c) $\mathrm{Cov}_*\left(e_i^*, e_j^*\right) = 0 \quad si\ i \neq j$

d) chaque fois que le même ensemble d'enregistrements contribue à une fréquence de cellule, la valeur de e_i sera la même (voir Fraser et Wooton, 2005);

e) e_i^* est un entier.

Le critère a) assure que les fréquences sont sans biais sur la distribution des perturbations. Le critère b) signifie que toute fréquence de cellule a une variance de perturbation fixe. Le critère c) fait en sorte que le calcul de la différence entre deux fréquences de cellule n'élimine pas l'effet de la perturbation. Le critère d) fait en sorte que l'effet de la perturbation ne soit pas éliminé par la demande répétée de la même fréquence de cellule.

Le tableau 1 donne un exemple de fréquence tabulaire avant et après perturbation. Les fréquences perturbées sont marquées d'un astérisque, tandis que les fréquences originales ne le sont pas. Par exemple, la fréquence réelle de 1 est perturbée pour devenir 3.

Tableau 1
Exemple de fréquences tabulaires avant et après perturbation

	Traitement A				Traitement B			
	Succès	*Essais*	*Succès**	*Essais**	*Succès*	*Essais*	*Succès**	*Essais**
Clinique 1	1	5	3	6	10	20	9	17
Clinique 2	9	10	9	11	5	20	4	18
Totaux	10	15	12	17	15	40	13	35

* Fréquences perturbées

3.2 Générateur de tableau de données d'enquête (*Survey Table Builder*)

Le *Survey Table Builder* (STB) applique le CDS aux fréquences pondérées par les poids de sondage. Désignons la i^{e} fréquence pondérée dans un tableau de contingence par $N_i = \sum_j d_j \delta_{ij}$, où d_j est le poids de sondage pour le j^{e} enregistrement. La fréquence perturbée correspondante est $N_i^* = \left[\tilde{d}_i n_i^*\right] + A_i^*$, où $\tilde{d}_i = n_i^{-1} N_i$ est le poids moyen

des enregistrements appartenant à la i^{e} cellule, n_i^* est la fréquence d'échantillon perturbée décrite antérieurement, $[x]$ arrondit x à l'entier le plus proche, et A_i^* remplit une fonction analogue à a_i^*, mais pour des fréquences pondérées (pour des renseignements détaillés voir l'annexe). Le STB ne diffuse aucune information au sujet de $\tilde{d}_i$, e_i^*, n_i^* ou N_i à l'analyste. Si $\tilde{d}_i$ =1 pour tout *i*, les méthodes de CDS du CTB et du STB sont équivalentes. Marley et Leaver (2011) ont étudié les mesures du risque et l'utilité associées au STB.

4. Serveur d'analyse

4.1 Sans contrôle de la divulgation statistique (cas standard)

Premièrement, nous examinons le cas standard de l'estimation des coefficients d'un modèle de régression. Considérons les microdonnées à partir desquelles un analyste spécifie une variable dépendante *y* et K covariable **x**, où les données sont $\mathbf{d} = \{(y_j, x_j) : j = 1,...n\}$. Ajustons un modèle de régression dont le paramètre est $\boldsymbol{\beta}$ en utilisant une fonction d'estimation sans biais $H(\boldsymbol{\beta})$ (voir Chambers et Skinner, 2003). En particulier, nous considérons l'équation d'estimation

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} G_j(\boldsymbol{\beta})\{y_j - f_j(\boldsymbol{\beta})\},$$

où $f_j(\boldsymbol{\beta}) = E(y_j | x_j)$ et $G_j(\boldsymbol{\beta})$ est un vecteur d'ordre *K* dont le k^{e} élément $G_{jk}(\hat{\boldsymbol{\beta}})$ est une fonction de $\boldsymbol{\beta}$ et x_j, mais non de y_j. La solution de $H(\boldsymbol{\beta}) = \mathbf{0}$ donne l'estimation standard, $\hat{\boldsymbol{\beta}}$, des coefficients de régression.

Les tentatives d'attaque des données comprennent l'obtention de $\hat{\theta}$ pour plusieurs demandes afin de reconstruire les attributs d'un enregistrement individuel. Ces attaques peuvent consister à calculer des différences, à tirer parti d'un enregistrement unique, à isoler un enregistrement avec une covariable ou à faire des inférences au moyen d'un modèle extrêmement précis. Une bonne discussion de ces situations figure, par exemple, dans Gomatam (2008). D'autres données de sortie, telles que les diagrammes, les statistiques diagnostiques ou les valeurs p peuvent naturellement être utilisées dans une attaque de données.

Lorsque l'on conçoit un ensemble de perturbations et de contraintes en vue de les appliquer aux données de sortie d'une analyse, il devient vite évident qu'une série de régressions conçues pour trouver un modèle optimal pourrait être impossible à distinguer d'une attaque de données sophistiquée. C'est en ça que réside le défi : ne pas restreindre la première action tout en contrecarrant la seconde.

4.2 Avec contrôle de la divulgation statistique

Nous allons maintenant discuter de l'approche que l'ABS envisage de mettre en œuvre dans son serveur d'analyse à distance.

4.2.1 Estimation des paramètres

Au lieu de résoudre $H(\boldsymbol{\beta}) = \mathbf{0}$ et de diffuser $\hat{\boldsymbol{\beta}}$, le serveur résout

$$H(\boldsymbol{\beta}) = \mathbf{E}^* \quad (1)$$

et diffuse l'estimateur résultant $\hat{\boldsymbol{\beta}}^*$, où $\mathbf{E}^* = (E_1^*, E_k^* ..., E_K^*)'$ sont les perturbations introduites aux fins du CDS, $E_k^* = u_k^* e_k$, u_k^* étant la distribution uniforme sur l'intervalle (-1,1), et $e_k = \max_j \left\{ G_{jk}(\hat{\boldsymbol{\beta}}) \left(y_j - f_j(\hat{\boldsymbol{\beta}}) \right) \right\}$ étant l'influence maximale que peut avoir un enregistrement sur la k^{e} équation d'estimation. Par exemple, dans le cas de variables binaires et du modèle logistique $e_k = 1$, les distributions des perturbations, E_k^*, sont indépendantes et si le

même modèle est ajusté, on utilise la même valeur de $\mathbf{E}^*$ – ce qui empêche un analyste d'estimer $\hat{\boldsymbol{\beta}}$ en ajustant le même modèle plusieurs fois et en calculant la moyenne sur les paramètres de régression obtenus en résolvant (1).

La grandeur de la perturbation est conçue de manière qu'elle soit suffisante pour masquer la contribution de tout enregistrement à l'équation d'estimation. L'application de la perturbation à la fonction de score est importante, car c'est là que $\hat{\boldsymbol{\beta}}$ impose une contrainte aux valeurs des données.

4.2.2 Inférence

Afin de faire une inférence valide à l'aide de $\hat{\boldsymbol{\beta}}^*$, un analyste doit tenir compte de la variance du modèle ainsi que de la perturbation de l'équation d'estimation. La variance de $\hat{\boldsymbol{\beta}}^*$ est

$$\mathbf{V}_{m*}\left(\hat{\boldsymbol{\beta}}^*\right) = V_m\left(\hat{\boldsymbol{\beta}}\right) + V_*\left(\hat{\boldsymbol{\beta}}^*\right)$$

où $V_m\left(\hat{\boldsymbol{\beta}}\right)$ est la variance de $\hat{\boldsymbol{\beta}}$ due au modèle (c'est-à-dire en l'absence de toute perturbation) et $V_*\left(\hat{\boldsymbol{\beta}}^*\right)$ est la variance de $\hat{\boldsymbol{\beta}}^*$ due à la perturbation. Nous proposons d'estimer $V_m\left(\hat{\boldsymbol{\beta}}\right)$ en utilisant la méthode du jackknife avec suppression d'un groupe (Rao et Wu, 1988). Un avantage de la méthode du jackknife est que les calculs sont simples et qu'elle est sans biais quand les microdonnées ont été recueillies auprès d'un échantillon issu d'un plan de sondage complexe (par exemple échantillonnage en grappes), comme cela est le cas de nombreuses enquêtes réalisées par l'ABS. La méthode du jackknife comprend l'affectation de toutes les unités sélectionnées à un seul groupe réplique de la même façon que l'échantillon a été sélectionné à partir de la population. En utilisant une approche semblable à celle de l'estimateur sandwich de la variance (voir Chambers et Skinner, 2003, p. 105), nous calculons $V_*\left(\hat{\boldsymbol{\beta}}^*\right) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{D}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$, où $\mathbf{D} = Var_*(\mathbf{E}^*)$.

Nous soutenons que l'incertitude présente dans l'estimateur de la variance par le jackknife (voir p. 196 Shao et Tu, 1996) suite à l'affectation des unités sélectionnées aux groupes répliques est telle que la variance totale, $\mathbf{V}_{m*}\left(\hat{\boldsymbol{\beta}}^*\right)$, ne peut pas être utilisée pour attaquer les données.

4.2.3 Contraintes générales

Plusieurs auteurs ont mentionné qu'une perturbation à distribution fixe (telle que celle utilisée plus haut) ne suffit pas à elle seule à protéger les données de sortie d'une analyse dans le contexte de demandes multiples. Les approches en vue de gérer les risques supplémentaires comprennent l'imposition de contraintes dans le serveur d'analyse (voir Gomatam et coll., 2005; Sparks et coll., 2008). Par ailleurs, lorsque l'on cherche à établir un ensemble de contraintes en vue de gérer le risque de divulgation, il devient vite évident qu'une série de régressions conçues pour trouver le modèle optimal pourrait être impossible à distinguer d'une attaque sophistiquée des données. C'est en ça que réside le défi, ne pas contraindre la première action tout en contrecarrant la seconde. À la présente sous-section, nous mentionnons un ensemble de contraintes qui ne défendent pas contre une attaque particulière des données, mais sont plutôt conçues pour gêner considérablement les personnes qui essayent d'attaquer les données tout en ne réduisant que légèrement l'utilité de ces dernières. Ces contraintes générales sont :

- $n > 50$;
- $n/K > 10$;
- $K>5$;
- ajuster des modèles à un sous-ensemble des enregistrements, où le sous-ensemble est défini par au plus 4 (toujours moins que K) variables binaires se trouvant au départ dans les microdonnées ;
- créer de nouvelles variables binaires uniquement à partir d'autres variables binaires présentes au départ dans les microdonnées ;
- créer de nouvelles variables continues uniquement en utilisant certaines transformations ;

- contraindre les variables à être non nulles dans au moins 15 enregistrements ;
- pour les modèles ne contenant que des covariables binaires, contraindre le nombre des combinaisons de covariables dans **x** à être supérieur à 50 ;
- imposer que $\mathbf{X}'\mathbf{X}$ soit de plein rang, où $\mathbf{X} = (\mathbf{x}_1, ...\mathbf{x}_j, ...\mathbf{x}_n)'$ et $\mathbf{x}_j$ est le vecteur colonne de dimension *K* des covariables pour le j^{e} enregistrement.

Les valeurs (par exemple 50) utilisées dans les contraintes susmentionnées sont données à titre d'exemple et peuvent naturellement être modifiées.

4.2.4 Contraintes supplémentaires selon le type d'attaque

Comme nous l'avons mentionné plus haut, certains types d'attaques des données sont bien décrits (voir par exemple, Gomatam 2008). Il est donc logique d'imposer, en plus de celles mentionnées à la section 3.2.3, des contraintes en vue de se défendre explicitement contre ces attaques. Nous ne discutons pas de ces contraintes ici, faute d'espace, mais elles peuvent être consultées dans Chipperfield et O'Keefe (2011). Par contre, nous décrivons brièvement trois attaques pour lesquelles sont construites des défenses explicites.

L'une de ces attaques est appelée *attaque par différence*. Une telle attaque comprend l'ajustement du même modèle à deux ensembles d'enregistrements identiques à l'exception d'un seul, que l'on élimine de l'un des deux ensembles. Les différences entre les coefficients de régression des deux modèles pourraient être utilisées pour essayer de reconstruire les attributs de l'enregistrement supprimé. Par exemple, si les covariables de l'enregistrement supprimé sont connues de l'attaquant, la différence entre les coefficients de régression permettrait de dériver une variable dépendante binaire pour l'enregistrement supprimé.

Une autre forme d'attaque consiste à *ajuster des modèles différents au même ensemble d'enregistrements* et de leurs attributs (c'est-à-dire même ensemble de données) en procédant comme il suit :

1. permuter le choix pour la variable dépendante ;
2. utiliser une fonction de lien différente (par exemple linéaire, logistique ou probit) ;
3. utiliser des variables qui sont des transformations différentes des mêmes attributs.

Chaque modèle impose *K* contraintes à un ensemble d'attributs des enregistrements, qui sont inconnus de l'analyste. Le but de cette attaque est d'imposer suffisamment de contraintes afin qu'il soit possible de trouver les valeurs dans l'ensemble de données sous-jacent.

4.2.5 Diagnostics

Une gamme de statistiques de test (voir Hosmer et Lemeshow, 2000) sont disponibles pour évaluer les hypothèses d'un modèle (par exemple normalité des résidus) et l'ajustement d'un modèle (par exemple AIC, R-carré). De nouveau, lorsqu'il diffuse ce genre de statistiques, l'organisme statistique doit trouver un compromis entre le risque de divulgation et l'utilité. Idéalement, le choix d'un modèle par l'analyste ne devrait pas être influencé par le contrôle de la divulgation statistique.

L'approche du CDS pour l'estimation du paramètre de dispersion ou des statistiques diagnostiques suit de près celle adoptée pour les coefficients de régression. Désignons ce genre de paramètre ou de statistique par $t^* = t(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$. Au lieu de diffuser $t^* = t(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$, nous diffusons

$$t^{**} = t^* + u^* s(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$$

où u^* est une variable aléatoire dans l'intervalle (-1,1) et $s(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$ borne l'influence maximale qu'un seul enregistrement dans **d** peut avoir sur la statistique t^* sachant $\hat{\boldsymbol{\beta}}^*$.

Les diagnostics qui comprennent la représentation graphique des valeurs d'enregistrements individuels (par exemple diagrammes des résidus) sont agrégés d'une certaine façon, en s'inspirant de Sparks et coll. (2008). Par exemple, les diagrammes Q-Q sont remplacés par une droite de régression non paramétrique lissée et les diagrammes des résidus sont remplacés par des boîtes à moustaches parallèles ou par des graphiques à barres parallèles.

Annexe

Désignons les cellules internes et marginales d'un tableau par $t = 1, 2, ..., C, C+1, ..., T,$ où $t = 1, 2, ..., C$ désigne les cellules internes. Désignons la t^e fréquence de cellule par n_t. Au lieu de diffuser n_t, le programme de génération de tableaux Table Builder diffuse la fréquence $n_t^* = n_t + e_t^* + a_t^*$ qui est obtenue en deux étapes. La première étape consiste à calculer les fréquences préliminaires $m_t^* = n_t + e_t^*,$ où e_t^* possède les propriétés a) à e) de la section 2.2. Les fréquences préliminaires du tableau ne sont pas cohérentes : les sommes des fréquences préliminaires pour les cellules internes ne sont pas forcément égales aux fréquences marginales préliminaires correspondantes. La deuxième étape consiste à trouver la valeur de a_t^* de sorte que le tableau contenant les fréquences n_t^* soit cohérent et que $\left|a_t^*\right| \le L_a$ pour tout $l = 1, ..., L.$ Cela signifie qu'aucune fréquence préliminaire d'une cellule marginale ou d'une cellule interne n'est modifiée de plus que L_a.

Bibliographie

Bleninger, P., Drechsler, J. et G. Ronning (2010), « Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study », *Privacy in Statistical Databases*, Springer.

Chambers, R.L. et C.J. Skinner (2003), *Analysis of Survey Data*, John Wiley & Sons.

Chipperfield, J.O. et M.C. O'Keefe (2011), « Disclosure-Protected Inference using Generalised Linear Models », rapport non publié, Canberra, Australie : Australian Bureau of Statistics.

Fraser, B. et J. Wooton (2005), « A proposed method for confidentialising tabular output to protect against differencing », UNECE work session on Statistical Data Confidentiality.

Gomatam, S., Karr, A.F., Reiter, J.P. et A.P. Sanil (2008), « Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk –Utility Framework for Remote Access Analysis Servers », *Statistical Science*, n° 20, p.163 à 177.

Hosmer, D.W. et S. Lemeshow (2000), *Applied Regression Analysis*, John Wiley and Sons.

Little, R.J.A. (1993), « Statistical Analysis of Masked Data », *Journal of Official Statistics*, n° 2, p. 407 à 426.

Marley, J.K. et V.L. Leaver (2011), « A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis », Proceedings of the International Statistics Institute.

Mathews, G.J. et O. Harel (2011), « Data Confidentiality: A Review of methods for statistical disclosure limitation and methods for assessing privacy », *Statistical Surveys*, 5, p. 1 à 29.

Rao, J.N.K. et C.F.J. Wu (1988), « Resampling Inference with Complex Survey Data », *Journal of the American Statistical Association*, 83, p. 231 à 241.

Reiter, J.P. (2002), « Satisfying Disclosure Restrictions with Synthetic Data Sets », *Journal of Official Statistics*, no. 18, p. 531 à 543.

Shao, J. et D. Tu (1996), *The Jackknife and Bootstrap*, Springer.

Shlomo, N. et C. Skinner (2010), « Assessing the Protections provided by Missclassification-based Disclosure Limitation », *The Annals of Applied Statistics*, p. 1291 à 1310.

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T. et D. McAullay (2008), « Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM », *Computer Methods and Programs in Biomedicine*, 91, p. 208 à 222.

Willenburg. L. et T. de Waal (2000), *Elements of Disclosure Control*, Springer.

SÉANCE 7B

CONTENU ET COLLECTE

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Certaines conséquences de la normalisation des méthodes de surveillance de la qualité des interviews d'enquête

Doug Currivan, Derek Stone, Kristin Fuller, Susan Kinsey et Howard Speizer[1]

Résumé

La normalisation des méthodes et des outils pour l'évaluation de la qualité des interviews d'enquête, pour les divers modes et études, représente un objectif de plus en plus important pour de nombreux organismes d'enquête. RTI a élaboré le système QUEST, un système normalisé d'évaluation pour la surveillance de la qualité des interviews, peu importe le mode. Ce système permet d'évaluer les comportements dans le cadre des interviews sur place et téléphoniques, à partir d'un ensemble commun de mesures de la qualité qui sont conservées dans une base de données unique partagée. Le système appuie l'évaluation de la qualité des interviews, tant pour la surveillance sur place en temps réel que pour l'examen des fichiers d'interviews enregistrées assistées par ordinateur (IEAO). QUEST remplace un ensemble varié de processus et d'outils de surveillance de la qualité de RTI, qui sont utilisés pour une vaste gamme de projets d'interviews sur place et téléphoniques, et permet l'évaluation de la qualité des données des intervieweurs au fil du temps, pour les divers modes et enquêtes. La communication porte sur des questions méthodologiques importantes découlant de la mise en œuvre de processus normalisés de surveillance de la qualité au moyen de QUEST. De façon plus particulière, nous examinons : 1) les demandes d'adaptation des protocoles QUEST aux différentes techniques d'interview et tâches spécialisées d'interview, 2) l'effet des multiples changements de protocole sur les résultats de la surveillance, comme les modules d'évaluation du rendement utilisés et les erreurs décelées, et 3) les premiers résultats de l'évaluation des variations entre les surveillants en ce qui concerne la détection des erreurs d'une séance de surveillance QUEST à l'autre. La présente communication aborde les prochaines étapes prévues pour améliorer ces procédures et outils, afin de rehausser de façon continue cet effort de surveillance de la qualité normalisé.

Mots clés : Interviews d'enquête ; contrôle de la qualité ; variabilité entre les surveillants.

1. Contexte et introduction

Il y a près de 20 ans, Couper, Holland et Groves (1992) notaient qu'il arrive souvent que les protocoles de surveillance : 1) suivent des procédures non systématiques et subjectives, et 2) comprennent uniquement des impressions générales des interactions téléphoniques, plutôt que des mesures objectives du comportement. Ces dernières années, la normalisation des méthodes et outils pour l'évaluation de la qualité des interviews d'enquête, pour les divers modes et études, a représenté un objectif de plus en plus important pour de nombreux organismes d'enquête. RTI a élaboré le système QUEST, un système normalisé d'évaluation pour la surveillance de la qualité des interviews, peu importe le mode (Speizer et coll., 2009; Speizer et coll., 2010). Ce système permet d'évaluer les comportements dans le cadre des interviews sur place et téléphoniques, à partir d'un ensemble commun de mesures de la qualité qui sont conservées dans une base de données unique partagée. Le système appuie l'évaluation de la qualité des interviews, tant pour la surveillance sur place en temps réel que pour l'examen des fichiers d'interviews enregistrées assistées par ordinateur (IEAO).

QUEST fournit un système normalisé pour la surveillance du rendement des intervieweurs, du point de vue de l'authenticité des interviews menées et de la pertinence des protocoles de collecte des données et d'administration des interviews. De façon plus particulière, ce système appuie la qualité des interviews, grâce : 1) à un ensemble uniforme de compétences/comportements des intervieweurs pour l'évaluation des interviews sur place et téléphoniques ; 2) un formulaire d'évaluation commun, une rubrique de notation, un suivi du rendement et un processus de rétroaction ; 3) l'utilisation accrue des IEAO pour améliorer la rétroaction des intervieweurs et suivre le rendement au niveau des questions de l'enquête (Biemer, Herget, Morton et Willis, 2000 ; Thissen et coll., 2008) ; 4) des données organisées sur le rendement des interviews et des intervieweurs dans l'ensemble des enquêtes, pour surveiller et améliorer la qualité ; et 5) une efficacité accrue, en vue de contrôler les coûts de surveillance de la qualité.

[1]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709.

La surveillance des interviews au moyen de QUEST nécessite le recours à un formulaire d'évaluation normalisé pour consigner les cas observés de comportements incorrects ou inappropriés de la part des intervieweurs. Le formulaire permet aussi de noter les comportements exceptionnels observés qui ont contribué de façon positive à la qualité des données. Le formulaire d'évaluation combine les comportements des intervieweurs dans 12 « modules d'évaluation du rendement » qui représentent des ensembles particuliers de compétences ou de résultats d'interview. Pour chaque séance de surveillance, QUEST produit un score global et un score de module pour chaque module examiné par le surveillant. Les scores sont fondés sur le nombre d'erreurs observées et leur « caractère critique ». Le caractère critique a trait à la gravité d'une erreur et, par conséquent, a des répercussions sur le score global de chaque séance. Tous les éléments QUEST sont définis comme non critiques, critiques ou extrêmement critiques. Le système utilise ensuite les scores de module et le nombre total d'erreurs observées pour produire un des trois scores globaux de séance suivants : « a dépassé les attentes », « a répondu aux attentes », ou « n'a pas répondu aux attentes ».

La présente communication porte sur trois questions méthodologiques importantes découlant des efforts de RTI pour mettre en œuvre des processus normalisés de surveillance de la qualité : 1) la nécessité d'adapter QUEST pour répondre à diverses techniques d'interview et tâches spécialisées d'interview, 2) l'effet des multiples changements de protocole sur les résultats de la surveillance, comme l'utilisation des modules de compétences et les erreurs décelées, et 3) les premiers résultats de l'évaluation des variations entre les surveillants en ce qui concerne la détection des erreurs d'une séance de surveillance QUEST à l'autre. Même si le système QUEST est utilisé à la fois pour les efforts de collecte sur le terrain (sur place) et par téléphone, la présente communication met uniquement l'accent sur la collecte des données par téléphone effectuée dans un centre d'appel de RTI. Les sections qui suivent décrivent la justification et les résultats de chacun de ces trois éléments, ainsi que les travaux à venir en vue d'améliorer les procédures et les outils, afin de rehausser de façon continue cet effort normalisé de surveillance de la qualité.

2. Adaptation du système en fonction des diverses techniques et tâches spécialisées

La présente section décrit les efforts récents en vue d'adapter QUEST à d'autres techniques d'interview et tâches spécialisées d'interview qui n'étaient pas prises en charge auparavant par le système. Les formulaires et l'algorithme de notation de QUEST ont été conçus pour permettre la création et la mise en œuvre de modules d'évaluation du rendement additionnels ou de rechange, lorsque cela est considéré comme nécessaire pour les différents projets de collecte de données. Les modules et les questions de base sont utilisés le plus souvent possible pour uniformiser la surveillance de la qualité, mais le système est suffisamment souple pour répondre aux besoins de projets spéciaux.

2.1 Prise en compte des techniques d'interview non normalisées

La majorité des enquêtes administrées par les intervieweurs à RTI sont conçues pour suivre des procédures d'interview conventionnelles normalisées (Fowler et Mangione, 1990). Le développement initial de QUEST a été axé sur l'évaluation de la qualité des interviews pour les protocoles normalisés. Des rajustements ont dû être apportés pour tenir compte des interviews prenant davantage la forme d'une « conversation » mises en œuvre pour certaines études (Conrad et Schober, 2000). Ces rajustements comprennent la création d'un module de rechange de compétences de lecture et la mise en œuvre d'une exigence, afin que les projets précisent si un module de compétences de lecture « conventionnel » ou « de type conversation » devait être utilisé. Les éléments relatifs à l'articulation et à la prononciation appropriées demeurent dans le module des interviews de type conversation, mais ceux axés sur l'administration normalisée ont été remplacés par des éléments des techniques de type conversation. La ***figure 2.1.1*** montre ces deux modules.

Figure 2.1.1
Module de compétences de lecture de QUEST pour les interviews normalisées conventionnelles par rapport aux interviews de type conversation

COMPÉTENCES DE LECTURE – INTERVIEW CONVENTIONNELLE
Articulation peu claire
Prononciation incorrecte
Ajout improvisé d'un mot/d'une phrase important
Ajout improvisé d'un mot/d'une phrase mineur
Omission d'un mot/d'une phrase important
Omission d'un mot/d'une phrase mineur
Catégories de réponse non lues quand il le fallait
Questions/instructions entièrement omises

COMPÉTENCES DE LECTURE – INTERVIEW DE TYPE CONVERSATION
Articulation peu claire
Prononciation incorrecte
Interview de type conversation non utilisée ou utilisée incorrectement
Éléments clés de la question omis
Usage inapproprié/excessif du paraphrasage
Questions/instructions entièrement omises
N'a pas utilisé une bonne grammaire

2.2 Prise en compte des tâches spécialisées d'interview

Même si l'objectif premier de QUEST est de faciliter la surveillance de la qualité des enquêtes normalisées administrées par des intervieweurs, pour certaines études, les intervieweurs s'acquittent d'autres tâches qui nécessitent une surveillance de la qualité. Parmi elles figure la reprise de contact avec les membres du ménage en vue de récupérer les données manquantes de leurs interviews. Le rajustement qui en découle a nécessité la création d'un module additionnel, *reprise de contact et suivi/extraction d'éléments de données manquants*. Ce module optionnel est montré dans la ***figure 2.2.1***.

Une deuxième tâche spécialisée d'interview du centre d'appel de RTI est le service d'assistance à l'appui des efforts de collecte des données. Les activités du service d'assistance sont uniques, du fait que le rôle premier de l'intervieweur est de fournir de l'assistance aux personnes et aux institutions échantillonnées, plutôt que de mener une interview. Les tâches du service d'assistance présentent des problèmes uniques qui n'étaient pas pris en compte dans le formulaire et les critères d'évaluation originaux de QUEST. C'est pourquoi un élément d'évaluation additionnel a été créé, en vue d'être ajouté aux multiples modules de QUEST, pour les projets faisant intervenir des tâches de service d'assistance. La ***figure 2.2.2*** montre les éléments ajoutés pour faciliter l'évaluation des tâches du service d'assistance.

Figure 2.2.1
Module QUEST pour la reprise de contact et l'extraction d'éléments de données manquants

REPRISE DE CONTACT ET SUIVI/EXTRACTION D'ÉLÉMENTS DE DONNÉES MANQUANTS
A codé incorrectement l'information sur le point d'appel
Raison du suivi incorrect
N'a pas recueilli les éléments de données manquants exactement
Explication incorrecte concernant les éléments de données manquants
Procédures de télécopieur/courriel administrées incorrectement
Message sur les mesures à prendre suivies incorrectement

Figure 2.2.2
Éléments additionnels de QUEST pour les tâches du service d'assistance

COMPÉTENCES DE LECTURE – INTERVIEW DE TYPE CONVERSATION
N'a pas utilisé la grammaire appropriée
COMPÉTENCES DE RÉTROACTION
A fourni une solution inappropriée
COMPÉTENCES DE PRÉSENTATION
Temps d'attente non communiqué
COMPORTEMENT PROFESSIONNEL
N'a pas manifesté d'empathie pour les préoccupations du demandeur
PROTOCOLE D'INTERVIEW
Expérience du contenu de l'étude non manifestée/n'a pas donné de preuve de connaissance de l'étude
Appel non consigné de manière complète/rapide
Assistance supplémentaire non offerte

3. Effet des multiples changements de protocole sur les résultats de la surveillance

Le 1er avril 2011, deux changements de protocole ont été effectués dans QUEST. Tout d'abord, l'intervalle d'échantillonnage a été réduit de 15 à 12 minutes pour les séances durant lesquelles un intervieweur est observé. En deuxième lieu, des éléments et des critères de notation supplémentaires ont été ajoutés dans les formulaires d'évaluation QUEST, y compris une nouvelle définition du caractère critique de certains comportements des intervieweurs. En outre, des séances de compte rendu se sont tenues avec le personnel de surveillance, afin de résoudre des questions comme celles à savoir à quel moment des modules d'évaluation du rendement particuliers de QUEST devraient être envisagés durant les séances de surveillance. L'objectif de ces changements de protocole et séances de compte rendu des surveillants était d'améliorer la quantité et la qualité des données de surveillance recueillies au centre d'appel de RTI. Afin d'évaluer l'effet possible des deux changements de protocole et des séances de compte rendu des surveillants sur les résultats de QUEST, nous avons examiné : 1) la proportion moyenne de modules d'évaluation du rendement pris en considération dans les séances de surveillance, et 2) les taux moyens d'erreur observés pour trois modules d'évaluation du rendement particuliers.

Tableau 3.1
Fréquence des modules d'évaluation du rendement non considérés avant et après le 1er avril 2011

Module d'évaluation du rendement	Séances avant le 1er avril 2011		Séances le 1er avril 2011 et après	
	En direct (n=54 528)	Enregistrée (n=264)	En direct (n=13 191)	Enregistrée (n=112)
Gestion de cas	13,8 %	68,6 %	9,3 %	50,0 %
Contact initial	55,1 %	37,9 %	48,4 %	30,4 %
Compétences de saisie clavier	56,6 %	73,5 %	52,2 %	50,9 %
Compétences de lecture	60,1 %	42,1 %	51,0 %	13,4 %
Compétences d'approfondissement	67,3 %	48,1 %	57,9 %	17,9 %
Compétences de rétroaction	70,6 %	58,3 %	60,0 %	24,1 %
Compétences de présentation	57,5 %	39,0 %	49,4 %	9,8 %
Comportement professionnel	55,8 %	39,0 %	46,3 %	9,8 %
Protocole d'interview	73,2 %	65,2 %	58,1 %	22,3 %

Le ***tableau 3.1*** présente les fréquences des séances dans lesquelles des modules particuliers de QUEST n'ont pas été considérés avant et après le 1er avril, tant pour les séances en direct qu'enregistrées. Globalement, ces données montrent que les surveillants ont évalué une proportion plus grande de modules dans les séances de surveillance en direct et enregistrées après les changements de protocole, le 1er avril 2011. Même si cette constatation semble

généralement s'appliquer à la fois aux séances en direct et à celles enregistrées, le nombre de séances enregistrées pour les deux périodes a été assez limité (n=376).

Le ***tableau 3.2*** présente les taux moyens de détection des erreurs pour trois modules d'évaluation du rendement sélectionnés, avant le 1er avril 2011 et après, tant pour les séances en direct que celles enregistrées. Pour les séances en direct et enregistrées, les taux de détection des erreurs ont augmenté légèrement pour les trois modules d'évaluation du rendement. La seule exception est qu'aucune erreur de gestion des cas n'a été observée dans les 112 séances enregistrées le 1er avril ou après. Au moment de la comparaison des résultats avant et après le 1er avril 2011, nous n'avons pas pu déterminer exactement comment les deux changements de protocole ou séances de compte rendu des surveillants peuvent avoir affecté les taux de détection d'erreurs particulières par les surveillants.

Tableau 3.2
Taux d'erreur pour trois modules d'évaluation du rendement avant et après le 1er avril 2011

Module d'évaluation du rendement	**Séances avant le 1er avril 2011**		**Séances le 1er avril 2011 et après**	
	En direct (n=54 528)	Enregistrée (n=264)	En direct (n=13 191)	Enregistrée (n=112)
Gestion des cas	0,013	0,023	0,015	0,000
Compétences de lecture	0,009	0,030	0,011	0,089
Compétences d'approfondissement	0,005	0,042	0,009	0,169

4. Résultats initiaux de l'évaluation de la variabilité dans la détection des erreurs

En théorie, les méthodes et outils très uniformisés de QUEST devraient favoriser un niveau élevé de cohérence entre les surveillants qui évaluent les séances en direct et enregistrées. Le fait que QUEST soit axé sur la collecte d'indicateurs objectifs du comportement des intervieweurs, par opposition à des impressions plus subjectives quant à la qualité des interviews, semble appuyer cette hypothèse. Diverses approches peuvent être adoptées pour examiner la fiabilité des évaluations entre surveillants, selon l'objectif de l'évaluation (Hicks et coll., 2010). Notre évaluation initiale de la fiabilité des évaluations entre surveillants dans QUEST a été axée sur la cohérence globale de la détection des erreurs entre les surveillants et de la détection des erreurs dans des modules d'évaluation du rendement particuliers. Ces données nous ont permis d'examiner ce qui suit : 1) la variabilité globale de la détection des erreurs entre l'ensemble des surveillants, et 2) si des surveillants particuliers semblent détecter des erreurs sensiblement plus/moins fréquemment que les taux moyens. L'analyse de la variabilité du rendement des surveillants repose sur le principe que les séances de surveillance à l'intérieur de projets de collecte des données sont attribuées aléatoirement aux surveillants du centre d'appel, ce qui semble être une hypothèse généralement fiable, compte tenu des procédures actuelles du centre d'appel de RTI.

Afin d'évaluer la variabilité entre les surveillants, nous avons organisé les données des séances sur les erreurs observées par tous les surveillants et par certains, pour trois modules d'évaluation du rendement d'un projet particulier, sur une période de trois mois. L'organisation des données par mois appuie l'hypothèse que l'ensemble des séances de surveillance ont été généralement attribuées de façon aléatoire aux surveillants, par opposition aux périodes plus courtes, dans lesquelles l'attribution aléatoire des séances a pu être limitée, à tout le moins pour certains surveillants. Ces données nous ont permis de déterminer la proportion moyenne des séances pendant lesquelles des erreurs avaient été observées et la proportion moyenne des séances pendant lesquelles des erreurs avaient été observées pour chacun des trois modules. Nous avons par la suite examiné comment les proportions pour chaque surveillant se comparaient à la moyenne et à l'écart-type pour toutes les séances menées par ces surveillants.

Tableau 4.1
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées (mai 2011)

Catégorie en rapport avec la proportion moyenne	Nombre de surveillants	Proportion de surveillants
Deux écarts-types ou plus au-dessus de la proportion moyenne	2	15,4 %
À deux écarts-types près (au-dessus ou en dessous) de la proportion moyenne	11	85,6 %
Deux écarts-types ou plus en dessous de la proportion moyenne	0	0,0 %
TOTAUX	13	100 %

Les ***tableaux 4.1 à 4.3*** inclusivement montrent le nombre et la proportion de surveillants regroupés en rapport avec la proportion moyenne de toutes les séances de surveillance au cours desquelles des erreurs ont été observées pour les mois de mai, juin et juillet 2011. Dans l'ensemble, la variabilité entre les surveillants en ce qui a trait à la détection des erreurs pour ce projet et cette période semble faible. Pour les trois mois, seulement un ou deux surveillants se trouvaient à deux écarts-types ou plus au-dessus de la proportion moyenne pour tous les surveillants. Aucun surveillant ne se situait à deux écarts-types ou plus au-dessus de la proportion moyenne pour l'un et l'autre des trois mois. Pendant ces trois mois, les surveillants semblent avoir affiché une probabilité globale similaire de détecter des erreurs durant les séances.

Tableau 4.2
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées (juin 2011)

Catégorie en rapport avec la proportion moyenne	Nombre de surveillants	Proportion de surveillants
Deux écarts-types ou plus au-dessus de la proportion moyenne	1	10,0 %
À deux écarts-types près (au-dessus ou en dessous) de la proportion moyenne	9	90,0 %
Deux écarts-types ou plus en dessous de la proportion moyenne	0	0,0 %
TOTAUX	10	100 %

Tableau 4.3
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées (juillet 2011)

Catégorie en rapport avec la proportion moyenne	Nombre de surveillants	Proportion de surveillants
Deux écarts-types ou plus au-dessus de la proportion moyenne	1	8,3 %
À deux écarts-types près (au-dessus ou en dessous) de la proportion moyenne	11	91,7 %
Deux écarts-types ou plus en dessous de la proportion moyenne	0	0,0 %
TOTAUX	12	100,0 %

Les ***tableaux 4.4 à 4.***6 inclusivement fournissent des données similaires sur la détection des erreurs pour trois modules d'évaluation du rendement particuliers de QUEST, compétences de lecture, compétences d'approfondissement et gestion de cas, pour mai, juin et juillet 2011. Ces tableaux montrent la proportion moyenne de séances pendant lesquelles des erreurs ont été détectées pour ces trois modules à la deuxième colonne. La troisième et la

quatrième colonnes indiquent le nombre de surveillants dont le taux d'erreur se situait à deux écarts types ou plus de la moyenne ou à deux écarts types ou plus en dessous de la moyenne. Comme c'est le cas pour les taux de détection globaux des erreurs de ces trois mois qui figurent dans les ***tableaux 4.1 à 4.3*** inclusivement, un ou deux surveillants semblent être plus susceptibles que la norme de détecter des erreurs des intervieweurs à l'intérieur de chacun de ces modules d'évaluation du rendement. Mis à part ces cas, les surveillants semblent être relativement constants dans la détection des erreurs à l'intérieur de ces trois modules.

Tableau 4.4
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées pour certains modules (mai 2011)

Module d'évaluation du rendement	Proportion moyenne de séances peu importe le nombre d'erreurs	Nombre de surveillants se situant à deux écarts-types ou plus au-dessus de la moyenne	Nombre de surveillants se situant à deux écarts-types ou plus en dessous de la moyenne
Compétences de lecture	0,017	1	0
Compétences d'approfondissement	0,007	2	0
Gestion de cas	0,006	1	0

Tableau 4.5
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées pour certains modules (juin 2011)

Module d'évaluation du rendement	Proportion moyenne de séances peu importe le nombre d'erreurs	Nombre de surveillants se situant à deux écarts-types ou plus au-dessus de la moyenne	Nombre de surveillants se situant à deux écarts-types ou plus en dessous de la moyenne
Compétences de lecture	0,009	2	0
Compétences d'approfondissement	0,006	1	0
Gestion de cas	0,011	1	0

Tableau 4.6
Variabilité entre les surveillants dans les séances pendant lesquelles une ou plusieurs erreurs ont été observées pour certains modules (juillet 2011)

Module d'évaluation du rendement	Proportion moyenne de séances peu importe le nombre d'erreurs	Nombre de surveillants se situant à deux écarts-types ou plus au-dessus de la moyenne	Nombre de surveillants se situant à deux écarts-types ou plus en dessous de la moyenne
Compétences de lecture	0,015	1	0
Compétences d'approfondissement	0,012	2	0
Gestion de cas	0,013	1	0

5. Conclusions et prochaines étapes

QUEST a été conçu pour assurer la normalisation de la surveillance de la qualité. Au fil du temps, les besoins liés aux divers efforts de collecte des données ont fait en sorte que des améliorations ont dû être apportées au système pour tenir compte des différentes procédures et tâches spéciales. En outre, des changements multiples de protocole ont été requis de façon périodique pour améliorer la quantité et la qualité des données de surveillance recueillies. Sans pouvoir isoler l'effet des changements particuliers de protocole, les résultats de la surveillance par QUEST ont fait ressortir une proportion plus forte de modules d'évaluation du rendement considérés pendant les séances et des taux de détection des erreurs légèrement accrus pour trois modules d'évaluation du rendement, par suite de la mise en œuvre de ces changements. Les données initiales sur la variabilité des surveillants au chapitre de l'évaluation des séances montrent que la cohérence entre les surveillants est assez élevée en ce qui concerne le taux global d'erreurs

détectées et les erreurs détectées dans des modules particuliers. Pour l'avenir, RTI prévoit maintenir des protocoles normalisés pour la surveillance de la qualité des interviews sur le terrain et par téléphone, tout en tenant compte des techniques et tâches spéciales, au besoin. Nous prévoyons poursuivre l'analyse des effets possibles sur les résultats de QUEST des changements importants de protocole qui sont apportés. Une évaluation plus poussée de la variabilité entre les surveillants dans la détection des erreurs comprendra l'analyse de la variabilité entre les surveillants sur des périodes plus longues et l'évaluation des degrés de concordance au niveau des erreurs entre les surveillants pour des interactions particulières enregistrées.

Remerciements

Les auteurs souhaitent remercier les autres membres actuels de l'équipe QUEST : Richard Heman-Ackah, Sridevi Sattaluri, Curry Spain, Dave Foster, Melissa Cominole et Nicole Tate. Nous tenons en outre à souligner la contribution des anciens membres de l'équipe : Rita Thissen, Orin Day, Mai Nguyen, Mary Allen et Courtney Gainey.

Bibliographie

Biemer, P., Herget, D., Morton, J. et W.G. Willis (2000), « The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI) », dans *Proceedings of the American Statistical Association's Section on Survey Research Methods*, p. 1068 à 1073.

Conrad, F. et M. Schober (2000), « Clarifying question meaning in a household telephone survey », *Public Opinion Quarterly*, 64, 1 à 28.

Couper, M., Holland, L. et R. Groves (1992), « Developing systematic procedures for monitoring in a centralized telephone facility », *Journal of Official Statistics*, 8, 63 à 76.

Fowler, F.J. et T. Mangione. (1990), *Standardized Survey Interviewing: Minimizing Interviewer-related Error*, Sage: Newbury Park, CA.

Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kotejin, L. et A. Moss (2010), « Using CARI Tools to Understand Measurement Error », *Public Opinion Quarterly*, 74, 985 à 1003.

Speizer, H., Kinsey, S., Heman-Ackah, R. et R. Thissen (2009), « Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance », présenté à la conférence du *Federal Committee on Statistical Methodology* Research, Washington, D.C.

Speizer, H., Currivan, D., Heman-Ackah, R. et S. Kinsey (2010), « Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance: lessons learned », présenté à la conférence annuelle *American Association for Public Opinion Research*, Chicago, IL.

Thissen, M.R., Sattaluri, S., McFarlane, E. et P. Biemer (2008), « The Evolution of Audio Recording in Field Surveys », *Survey Practice*. http://surveypractice.org/2008/ 12/19/audio-recording/ (accédé au lien le 2 février 2010).

Enquête santé européenne par examen : la perspective de l'échantillonnage et du recrutement

Johan Heldal[1], Susie Jentoft[1],
Kari Kuulasmaa[2], Päivikki Koponen[2] et Sanna Ahonen[2]

Résumé

En 2009, le projet European Health Examination Survey (EHES) - Enquête santé européenne par examen - a été lancé, avec le soutien de la Commission européenne, en vue de recueillir des données comparables de grande qualité sur la santé et les risques pour la santé de la population adulte de l'Europe. L'enquête comprend une interview et des mesures physiques de base. Les divers pays peuvent inclure leur propre contenu. L'objectif de l'EHES est de fournir des données pour la planification et l'évaluation des politiques en matière de santé, la promotion de la santé et la recherche, tant au niveau national qu'à l'échelle de l'Europe.

Le Centre de référence de l'EHES coordonne les activités et établit des normes communes pour tous les aspects de la collecte des données. Au total, 12 pays ont participé à une étape pilote. Cinq d'entre eux ont mené des enquêtes régulières, et les autres sont prêts à entreprendre leurs enquêtes nationales.

L'établissement de normes communes qui peuvent être appliquées pour créer des enquêtes comparables dans 30 pays, dans lesquels la disponibilité des bases de sondage, la culture et les dispositions législatives varient, représente un défi de taille. La communication est axée sur les différences dans l'échantillonnage, le recrutement et la réponse aux enquêtes entre les pays européens, et les approches pour les traiter.

Mots clés : Plans de sondage; enquête par examen; taux de participation; enquêtes pilotes; enquêtes régulières.

1. Introduction

L'histoire des enquêtes santé par examen en Europe remonte aux années 1960, au moment où la Finlande et la Suède ont mené leurs premières enquêtes nationales. Plus tard, plusieurs pays ont effectué de telles enquêtes. Toutefois, toutes ces enquêtes se tiennent au niveau national ou régional et comportent leurs propres protocoles, sans uniformisation ni coordination d'un pays à l'autre. Nombre d'entre elles sont des enquêtes sur échantillon représentatives au niveau national (EHRM, 2002, FEHES, 2008a). L'étude MONICA de l'Organisation mondiale de la Santé, à laquelle plusieurs pays d'Europe ont participé, utilisait des échantillons statistiques, mais ceux-ci n'étaient pas représentatifs au niveau national (MONICA, 2003).

La section 2 décrit l'organisation du projet de l'EHES. La section 3 donne un bref aperçu des recommandations du Centre de référence (CR) de l'EHES. La section 4 décrit les expériences des enquêtes pilotes et des enquêtes régulières menées jusqu'à maintenant. La section 5 présente des conclusions.

2. Organisation

Le projet pilote d'Enquête santé européenne par examen (EHES) a été lancé en 2009, pour une période de deux ans, dans le cadre du Programme Santé de l'Union européenne. L'objectif du projet était *d'élaborer et de planifier une Enquête santé par examen pilote au sein de l'Union européenne et des états membres de l'Association européenne de libre-échange/Accord sur l'Espace économique européen, en préparation pour l'essai des modules d'examen et des procédures sur le terrain de cette enquête*. Le projet pilote de l'EHES est coordonné par le CR de l'EHES, qui a été créé conjointement par le National Institute for Health and Welfare de la Finlande, Statistics Norway

[1]Statistics Norway, Kongens Gate. 6, N-0153 Oslo, Norvège (johan.heldal@ssb.no and susie.jentoft@ssb.no).
[2]National Institute for Health and Welfare (THL), Mannerheimintie 166, FIN-00300 Helsinki, Finlande (kari.kuulasmaa@thl.fi et hanna.tolonen@thl.fi).

(méthodologie statistique) et l'Istituto Superiore di Sanità en Italie (questions juridiques et éthiques). Le CR de l'EHES est aussi responsable de l'établissement de normes communes pour tous les aspects de la collecte des données. Il a été financé par la Commission européenne, dans le cadre d'un contrat de service[3]. Une action collective, bénéficiant d'un financement conjoint de l'Union européenne (UE), a été lancée pour se préparer en vue des enquêtes dans 12 pays (République tchèque, Allemagne, Grèce, Finlande, Italie, Malte, Pays-Bas, Norvège, Pologne, Portugal, Slovaquie et Royaume-Uni/Angleterre) et pour mener un essai pilote. La taille recommandée pour les enquêtes pilotes a été établie à environ 200 participants.

En décembre 2011, cinq de ces pays avaient mené des enquêtes régulières (Allemagne, Italie, Pays-Bas, Slovaquie et Royaume-Uni/Angleterre). Trois autres prévoient entreprendre leurs enquêtes nationales en 2012 (Grèce, Finlande et Portugal). Même si aucun financement au niveau de l'Europe n'est disponible pour les enquêtes régulières jusqu'à maintenant, les pays partagent un intérêt commun à l'égard de la comparabilité de leurs résultats et du partage des données avec le CR de l'EHES, pour l'évaluation de la qualité et la production de rapports conjoints.

L'objectif à long terme est d'établir un système d'ESE (Enquêtes santé par examen) fondées sur un échantillon national en Europe, afin de recueillir des données comparables de grande qualité sur la santé et les risques pour la santé de la population adulte en Europe. Ces données serviront à la planification et à l'évaluation des politiques en matière de santé, à la promotion de la santé et à la recherche, à l'échelle de l'Europe. La décision de lancer le projet de l'EHES a été fondée sur les recommandations de l'étude de faisabilité (FEHES), qui a conclu, en 2008, que l'objectif était atteignable. Le projet pilote de l'EHES devrait être complété d'ici la fin d'avril 2012. Le financement au niveau européen de la prochaine étape de l'EHES est toujours en attente.

3. Recommandations

Un manuel européen pour l'EHES est en voie d'élaboration. Le manuel doit être publié en trois parties :

A. Planification et préparation des enquêtes (16 chapitres, 194 pages)
B. Procédures pour les travaux sur le terrain (7 chapitres, 124 pages)
C. Collaboration et coordination au niveau de l'Europe (7 chapitres).

Les parties A et B ont été publiées (EHES, 2011). Elles étaient fondées sur les recommandations de la FEHES (FEHES, 2008b). La partie C sera finalisée d'ici février 2012.

3.1. Population cible et échantillonnage

La population cible de base de l'EHES, qui doit être utilisée dans tous les pays, est constituée de tous les habitants de 25 à 64 ans. Certains pays ont élargi l'intervalle d'âge. Toutefois, on a noté une certaine variation dans la couverture des bases de sondage disponibles (tableau 3.1-1).

Selon le manuel (partie A, chapitre 3), un plan d'échantillonnage à deux degrés stratifié au niveau géographique doit être utilisé dans la plupart des pays. Les unités primaires d'échantillonnage (UPE) sont utilisées comme régions de collecte des données, chacune comprenant un lieu d'examen. Les UPE devraient être sélectionnées selon la probabilité proportionnelle à la taille. Si possible, les personnes peuvent être stratifiées selon le sexe et l'âge au moment de l'échantillonnage de deuxième degré. Les procédures à cette fin sont décrites et mises en œuvre dans une application R qui est offerte (RcmdrPlugin.EHESsampling, voir CRAN). La taille d'échantillon recommandée pour chaque pays est de 500 personnes au minimum dans chacun des quatre groupes d'âge de dix ans (25-34, 35-44, 45 54 et 55-64) pour chaque sexe. Elle est fondée sur un effet de plan de sondage présumé de 1,5. Les UPE devraient être suffisamment petites pour que les participants puissent se rendre facilement au lieu d'examen et participer à l'enquête. Les détails relatifs aux recommandations d'échantillonnage se trouvent dans le manuel de l'EHES, partie A (EHES, 2011). Les enquêtes régulières devraient couvrir toutes les saisons, et les UPE devraient être visitées de façon aléatoire, afin d'éviter les effets confusionnels de la saison et de la géographie.

[3]Avertissement : Le projet pilote de l'EHES a reçu du financement de la Commission européenne/DG SANCO. Les opinions exprimées ici sont celles des auteurs et ne représentent pas la position officielle de la Commission.

Pour le deuxième degré de l'échantillonnage, une base de sondage de personnes comportant une couverture élevée a été recommandée. Dans la Health Survey of England, on a utilisé une base d'adresses postales et on a invité tous les résidents admissibles des adresses sélectionnées à participer. Dans l'enquête pilote en Grèce, les ménages ont été sélectionnés à partir du recensement de 2001, et on a retenu une personne dans chaque ménage.

Selon le manuel de l'EHES, il est approprié de reprendre les enquêtes, y compris les mesures de base, environ tous les cinq ans, alors que certaines mesures supplémentaires peuvent être répétées moins fréquemment. On peut aussi mettre en œuvre, comme solution de rechange, un système de collecte continue des données permettant l'agrégation des données d'enquête de plusieurs années, en vue de fournir des estimations de qualité. La Health Survey of England, la US National Health and Nutrition Survey (NHANES) et l'Enquête canadienne sur les mesures de la santé (ECMS) sont des enquêtes à collecte continue.

Tableau 3.1-1
Bases de sondage principales

Pays	Base de sondage	Couverture de la base
République tchèque	Registre national des résidents permanents	Tous les résidents
Finlande	Registre central des résidents permanents	Tous les résidents
Allemagne	Registre des populations locales	Tous les résidents
Grèce	Enquête pilote : Recensement de 2001, enquête régulière : Recensement de 2011	Ménages privés
Italie	Registre des résidents locaux	Tous les résidents
Malte	Registre central de la population	Tous les résidents
Pays-Bas	Registres de la population	Citoyens seulement
Norvège	Registre central de la population	Tous les résidents
Pologne	Registre national de la population	Tous les résidents
Portugal	Liste nationale des services de santé	Toutes les personnes inscrites dans le système national de santé
Slovaquie	Registre de la population	Tous les résidents
Royaume-Uni/ Angleterre	Liste des adresses postales	Tous les ménages privés

3.2 Recrutement

Le processus de recrutement devrait être planifié de la façon la plus pratique dans chaque pays. Les personnes ou ménages sélectionnés pour l'enquête devraient d'abord être contactés au moyen d'une lettre d'invitation comprenant une brochure. La brochure devrait fournir les renseignements les plus importants, de façon concise et intéressante, comme les objectifs de l'enquête, le questionnaire et les mesures, et inciter les personnes à participer. Elle devrait souligner l'importance de l'enquête et de la participation, décrire brièvement le processus de sélection, la confidentialité stricte des données d'enquête, les avantages pour la santé publique et pour les participants qui reçoivent les résultats et la possibilité de bénéficier d'un bilan de santé gratuit et d'autres renseignements pertinents. Un numéro de téléphone sans frais devrait être fourni pour permettre de poser des questions. On recommande de une à trois reprises de contact pour les participants qui ne se manifestent pas ou n'appellent pas. Dans la plupart des pays, les prises de contact par téléphone ou les visites à domicile semblent être plus efficaces que le simple envoi d'invitations par la poste. Le remplacement des non participants n'est pas autorisé. Le manuel aborde des facteurs qui peuvent affecter les taux de participation et suggère des mesures qui peuvent servir à augmenter la participation, y compris de longues heures d'ouverture, des appels aux employeurs, un profil dans les médias locaux et par les dirigeants locaux, des incitatifs et le remboursement des frais de déplacement. L'utilisation d'incitatifs dépend de la culture et de la législation, ainsi que de l'organisation de l'enquête et des ressources disponibles dans le pays. Il peut s'agir d'une compensation monétaire ou de petits cadeaux.

3.3 Interview et mesures

L'enquête comprend un questionnaire, des mesures physiques de base, y compris la taille, le poids, la circonférence de la taille, la tension artérielle et des échantillons de sang pour la mesure des lipides sanguins et de la glycémie à jeun ou HbA1c. Certains pays ont inclus des mesures supplémentaires dans l'enquête, par exemple, des tests de la fonction respiratoire, des électrocardiogrammes, des examens de la santé buccodentaire ou des tests de capacité fonctionnelle. Les questions de base sont fondées sur le questionnaire de l'EHIS. Celui ci est déjà traduit dans la plupart des langues européennes. L'utilisation des mêmes questions permet d'assurer la comparabilité avec les enquêtes de l'EHIS. Les mesures cliniques ont été sélectionnées sur la base d'un certain nombre de critères, l'élément clé étant de s'occuper des problèmes de santé publique. Pour un examen complet, voir le manuel de l'EHES, partie A, chapitre 5.

3.4 Gestion des données, estimation et accès

Toutes les données sur les mesures de base de l'enquête pilote et des enquêtes régulières effectuées à l'intérieur de l'EHES seront transférées au CR de l'EHES, de façon à préserver l'anonymat, et un protocole de partage des données est en voie d'élaboration, afin de permettre l'accès par les chercheurs. Le CR de l'EHES estimera un certain nombre d'indicateurs de base pour chaque pays. Ces indicateurs seront principalement des estimations uniformisées selon l'âge et le sexe pour des groupes d'âge de dix ans, selon des normalisations sur un an standard. Cette procédure permet de contrôler les répartitions différentes selon l'âge entre les pays et est nécessaire pour que les estimations soient comparables au niveau international au fil du temps. Des détails sur les indicateurs, ainsi que les procédures d'imputation, de pondération et d'estimation, seront inclus dans le manuel, partie C.

4. Défis

4.1 Expériences des enquêtes pilotes

De nombreux aspects des recommandations ont été vérifiés dans les enquêtes pilotes. Dans la plupart des pays, seulement un ou deux UPE/lieux d'examen ont été sélectionnés. Toutefois, la sélection des UPE a souvent été faite à dessein et non pas comme un test de plan de sondage complet pour l'échantillonnage de premier degré. Dans certains des pays menant des enquêtes régulières, les enquêtes pilotes ont été intégrées dans l'enquête principale. Des personnes ont été sélectionnées de façon aléatoire à l'intérieur des UPE retenues, habituellement stratifiées selon le sexe et le groupe d'âge. Tous les pays ont été visités par des représentants du CR de l'EHES, afin de vérifier comment les recommandations avaient été mises en œuvre et de discuter des problèmes. Toutes les enquêtes pilotes nationales étaient conformes au contenu de l'enquête de base et à l'intervalle d'âge. Même si des écarts mineurs ont été observés, on a noté peu de problèmes du point de vue de l'uniformisation des interviews ou des mesures physiques et cliniques. Le défi le plus important consistait à obtenir les taux de participation souhaités. La FEHES recommandait un taux de participation cible d'au moins 70 %, un objectif qu'aucune des enquêtes pilotes n'a permis d'atteindre. Le tableau 4.1-1 présente les taux de participation pour neuf des enquêtes pilotes, sans divulgation des noms des pays où elles se sont déroulées. Les chiffres sont fondés sur une analyse préliminaire des données au niveau individuel sur le statut au chapitre de la participation, et peuvent rendre compte en partie des erreurs possibles dans le codage des données. Le calcul des taux de participation est décrit dans le manuel de l'EHES, partie A, chapitre 13.

Tableau 4.1-1
Taux de participation aux enquêtes pilotes

Pays	Tailles d'échantillon		Taux de participation (%)	
	Hommes	Femmes	Hommes	Femmes
1	200	200	40	54
2	370	391	38	44
3	125	125	54	71
4	198	202	57	54
5	1 600	1 600	42	49
6	1 311		23	
7	245	245	41	43
8	300	300	34	47
9	124	126	44	67

Dans la plupart des enquêtes pilotes, les personnes ont été sélectionnées directement à partir des registres de la population (tableau 3.1-1). La qualité des bases de sondage variait. Les taux de non contact ont été élevés dans certaines études pilotes, parce que la base de sondage était périmée et/ou à cause du faible niveau de déclaration lié au recours au registre principal. Dans certains pays, les attitudes négatives à l'égard des enquêtes ou des autorités peuvent avoir contribué aux faibles taux de réponse. Les contacts personnels par téléphone ont été difficiles dans certains pays, étant donné que peu de personnes ont un téléphone fixe et que les numéros de téléphone mobiles ne sont pas toujours disponibles. Dans un pays, le comité d'éthique a empêché les organisateurs de l'enquête de communiquer à nouveau avec les participants d'aucune façon après la première lettre d'invitation, ce qui a eu un effet désastreux sur la participation. Peu de pays ont procédé au recrutement et/ou aux examens de santé au moyen de visites à domicile, et certains ont déclaré que les visites à domicile n'étaient pas appréciées dans leur pays. Il semble y avoir de grandes différences culturelles à cet égard. Nous croyons que les organisateurs de l'enquête et les comités d'éthique ne devraient pas hésiter à motiver les personnes de façon positive à participer, grâce aussi à des contacts personnels.

4.2 Échantillonnage pour les enquêtes régulières

Comme il est indiqué au début de la section 4.1, l'échantillonnage de la population n'a pu se faire que partiellement dans les petites enquêtes pilotes. La plupart des pays ont utilisé la même base de sondage pour sélectionner les personnes dans leurs enquêtes pilotes que pour les enquêtes régulières. Le rôle de ces bases de sondage consiste à fournir les tailles des UPE au premier degré du plan de sondage, ainsi qu'une base d'échantillonnage de deuxième degré dans les UPE sélectionnées. La plupart des bases de sondage nationales couvrent, au moins en principe, tous les résidents du pays, mais comme le montre le tableau 3.1-1, il existe des écarts à ce chapitre.

Les UPE ont besoin de divisions qui peuvent être clairement définies sur une carte et pour lesquelles il existe des données sur le nombre de personnes ou de ménages. Cela est habituellement possible pour les divisions administratives, comme les municipalités, les districts ou les régions (la signification de ces termes peut différer d'un pays à l'autre). Il est aussi possible d'utiliser les régions de code postal. Malheureusement, les divisions administratives sont parfois trop grandes ou trop petites, du point de vue de la population ou de l'étendue, pour être utilisées comme UPE. Dans ce cas, il est nécessaire de répartir ou de combiner des divisions pour obtenir des unités de taille convenable. Les nouvelles divisions devraient aussi pouvoir être facilement identifiables dans la base de sondage. Les possibilités qui s'offrent dépendent aussi du fait que les bases individuelles sont organisées ou non de façon centralisée ou qu'elles existent uniquement au niveau local.

De nombreux pays souhaitent utiliser les installations existantes, comme les centres régionaux de santé ou les hôpitaux, comme lieux d'examen. Toutefois, dans certains pays, la densité de ces installations est trop faible, et nombre d'entre elles couvrent un district comportant des distances de déplacement trop longues pour convenir comme UPE. La collaboration avec ces installations peut aussi poser un défi. D'autres pays ont évité ce problème en n'utilisant pas les installations existantes. En Allemagne, des petites fourgonnettes sillonnent le pays avec l'équipement nécessaire, et des lieux d'examen temporaires sont aménagés dans des emplacements loués, une semaine à la fois, avant de passer au lieu suivant. Dans le cadre de la Health Survey of England, on procède aux

examens au domicile des personnes, mais cette solution peut créer d'autres problèmes de comparabilité entre les pays. La solution de l'Allemagne pourrait être envisagée par d'autres pays, soit comme une approche générale ou comme un complément des installations existantes qui sont trop éloignées.

La NHANES (États Unis) et l'ECMS (Canada) utilisent des unités d'examen mobiles. De tels véhicules seront trop coûteux pour de nombreux pays européens. Ils doivent être utilisés de façon continue pendant de nombreuses années pour que leur coût se justifie économiquement. Un investissement dans des unités d'examen mobiles pour les enquêtes de l'EHES pourrait éventuellement se faire au niveau européen, avec des véhicules pouvant être utilisés dans de nombreux pays.

Tableau 4.2-1
Plan de sondage et période de certaines enquêtes régulières en Europe

Pays	Plan de sondage	Période d'enquête
Allemagne (DEGS)	180 UPE avec probabilité proportionnelle à la taille → 42 personnes par UPE → 7 560 personnes	Novembre 2008 à novembre 2011
Grèce (planifié)	UPE avec probabilité proportionnelle à la taille → 5 500 ménages → une personne par ménage.	Lorsque le Recensement de 2011 sera disponible comme base de sondage
Finlande (FINRISK, planifié)	87 strates → 10 000 personnes au premier degré. Couverture régionale seulement.	Janvier à avril 2012
Italie (OED)	Un centre par région (20) avec échantillonnage non probabiliste → m x 220 personnes examinées par centre → 9 020 personnes examinées. Seules les personnes à proximité du centre sont invitées en raison des distances de déplacement trop grandes.	Septembre 2008 à mars 2012
Malte (planifié)	Stratifié à un degré → 3 600 personnes	2014
Pays-Bas (NL de Maat)	7 UPE (15 planifiées) → 15 000 participants, 4 000 personnes examinées	Première phase : Mai à décembre 2009 Deuxième phase : Octobre à décembre 2010
Pologne (WOBASZ (pas une enquête de l'EHES*)	104 UPE dans 48 strates. Sélection avec probabilités égales des UPE à l'intérieur des strates. 100 hommes + 100 femmes sélectionnés dans chaque UPE → 20 800 participants	2003 à 2005
Slovaquie	Échantillon de 4 000 personnes de 36 districts comptant des instituts de santé publique régionaux. 43 districts sans instituts de santé ne sont pas couverts en raison des distances de déplacement trop grandes.	Octobre à décembre 2011
Royaume-Uni/ Angleterre (HSE 2011) (Pas une enquête de l'EHES*)	576 UPE avec probabilité proportionnelle à la taille → 16 adresses postales par UPE → 9 216 logements → tous les résidents admissibles	Janvier à décembre 2011

* Une enquête est définie comme une enquête de l'EHES si les données sont partagées avec le CR de l'EHES. Dans le cas du Royaume-Uni/Angleterre, seules les données de l'enquête pilote seront partagées.

Les enquêtes régulières menées jusqu'à maintenant étaient planifiées avant que les recommandations de l'EHES soient publiées, et l'inclusion dans l'EHES vient du fait qu'elles partagent un intérêt commun quant à la comparabilité de leurs résultats. Ces enquêtes ont adopté les recommandations de l'EHES dans la mesure du possible, tard à l'étape de la planification, et n'ont pu se conformer complètement. Cela concerne la couverture et le plan de sondage en particulier. Le tableau 4.2-1 présente deux aspects des différences entre ces enquêtes : les plans de sondage et les périodes d'enquête.

Dans l'une des enquêtes, les UPE n'ont pas été sélectionnées avec probabilité proportionnelle à la taille. Autrement, le nombre d'UPE montre une variation importante. Du point de vue d'un échantillonnage pur, il est souhaitable de compter de nombreuses UPE, avec quelques participants dans chacune, plutôt que quelques UPE, comptant de nombreux participants, afin de réduire la variation de l'échantillonnage. Toutefois, il peut être difficile au niveau opérationnel et coûteux de faire participer un grand nombre de centres de santé ou de cliniques. Il s'agit aussi d'une question d'organisation de l'enquête. Les deux pays comptant le nombre le plus important d'UPE sont les deux pays qui mènent leurs enquêtes de façon indépendante des installations fixes existantes.

Le tableau 4.2-1 montre aussi une grande variation dans les périodes d'enquête, qui vont de trois mois à presque trois ans et demi. Certains pays ont l'intention d'utiliser des périodes d'enquête plus courtes, pour assurer la comparabilité avec les enquêtes antérieures. Toutefois, pour permettre la comparabilité entre les pays, toutes les enquêtes futures régulières de l'EHES devraient couvrir la même saison. De préférence, toutes les saisons devraient être couvertes également. La NHANES et l'ECMS tirent leurs échantillons pour deux ans à la fois, afin de couvrir toutes les saisons. Dans l'EHES, des intérêts concurrents doivent être rapprochés.

5. Conclusion

Est il possible d'établir un système de grande qualité d'enquêtes santé par examen uniformisées comparables en Europe? Nous croyons que c'est le cas, mais il faut un engagement complet des pays et de l'UE pour cibler les défis méthodologiques mentionnés dans le présent article. Toute solution intermédiaire représentera un gaspillage d'argent. Il faut aussi élaborer des stratégies améliorées pour augmenter les taux de participation. Nous prévoyons aussi qu'une enquête régulière recevant une attention nationale, plutôt que seulement locale, suscitera davantage d'intérêt qu'une enquête pilote locale.

Au moment de la rédaction du présent article, seulement quelques unes des enquêtes mentionnées dans le tableau 4.2-1 étaient terminées et avaient fait l'objet d'un rapport. Au moment de la publication de l'article dans le recueil, nous en saurons davantage au sujet des résultats, des succès et des échecs, ainsi que de l'avenir de l'EHES. Des résultats fondés sur une analyse plus poussée des données seront publiés ultérieurement.

Remerciements

Des listes des employés clés qui contribuent au projet pilote de l'EHES sont disponibles à l'adresse suivante : http://www.ehes.info/contact.htm. L'action collective de l'EHES a reçu du financement de la Commission européenne (entente de subvention n° 2009-23-01). Le Centre de référence de l'EHES est financé par la Commission européenne, grâce à un contrat de service (SANCO/2008/C2/02-SI2.538318 EHES).

Bibliographie

European Health Examination Survey (EHES) (2011), Manuel de l'EHES, partie A et B, http://www.ehes.info/manuals/EHES_manual/EHES_manual.htm.

European Health Interview & Health Examination Surveys Database, https://hishes.iph.fgov.be/index.php?hishes=home.

European Health Risk Monitoring (EHRM) Project (2002), « Review of surveys for risk factors of major chronic diseases and comparability of the results », http://www.ktl.fi/publications/ehrm/product1/title.htm.

Feasibility of a European Health Examination Survey (FEHES) (2008a), « Review of Health Examination Surveys in Europe », *B18/2008, Publications of the National Public Health Institute*, Helsinki, disponible à http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2008/2008b18.pdf.

FEHES (2008b), Recommendations for the Health Examination Surveys in Europe *B21/2008, Publications of the National Public Health Institute*, Helsinki, disponible à http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2008/2008b21.pdf.

MONICA (2003), Monograph and Multimedia Sourcebook, World Health Organization, Geneva.

RcmdrPlugin.EHESsampling (2011), Software. *The Comprehensive R Archive Network*. http://www.r-project.org/, User manual available from authors.

Planification préliminaire de la collecte : Guichet de la collecte

Anie Marcil[1]

Résumé

Le programme de la collecte de Statistique Canada offre maintenant un point de contact organisationnel unique pour toutes les divisions clientes concernant les activités de collecte : un guichet de la collecte. Le guichet de la collecte est responsable de l'étape initiale d'évaluation de la faisabilité de la collecte. Pour ce faire, on examine les spécifications de l'enquête, on détermine le déroulement du processus de collecte des données, on évalue la capacité de collecte requise et l'on prépare les estimations budgétaires en collaboration avec les divers partenaires de la collecte.

Le guichet de la collecte vise à fournir des services de consultation sur la collecte des données aux divisions clientes; jouer un rôle de coordination quant aux services de collecte des données entre les partenaires de la collecte et les divisions clientes durant l'étape initiale d'évaluation de la faisabilité de la collecte ; jouer un rôle d'intégration et clarifier les rôles et responsabilités des partenaires de la collecte et des divisions clientes.

Ce nouveau service permet de normaliser et centraliser les activités de la collecte, assurer la cohérence pour l'ensemble des activités de collecte et utiliser des paradonnées d'enquêtes antérieures afin de bien guider la planification préliminaire des enquêtes de Statistique Canada.

Mots clés : Guichet de la collecte ; planification ; estimations budgétaires ; activités de la collecte.

1. Introduction

Depuis plusieurs années, les services de collecte et les procédures liées à la prestation de ces services ne cessent d'évoluer et de se complexifier. Auparavant, il était de la responsabilité des gestionnaires d'enquête de déterminer les services de collecte requis par leur enquête et, pour chacun de ces services, de trouver la personne-ressource qui pourrait fournir les estimations de coûts préliminaires. Ils pouvaient parfois consacrer une bonne part de leur temps pour chercher cette personne afin d'être conseillés ou encore redirigés vers les services dont ils avaient besoin. L'information fournie à ces gestionnaires pour chacun des services de collecte pouvait varier et ne pas être constante. Par conséquent, les estimations budgétaires n'étaient pas toujours comparables. D'ailleurs, ces gestionnaires pouvaient ne pas être au courant de tous les nouveaux services de collecte qui étaient à leur disposition, de sorte que les procédures suivies pour leur enquête n'étaient pas nécessairement les plus efficaces.

En 2007, la nouvelle vision des services de collecte de Statistique Canada proposa la création d'un point de contact organisationnel unique pour les divisions clientes qui recherchent des services de collecte. Deux années plus tard, soit en 2009, le guichet de la collecte fut inauguré. Le guichet de la collecte est donc devenu le point unique et central de la planification préliminaire des activités de la collecte pour toutes les enquêtes de Statistique Canada.

2. Clients et partenaires de la collecte

2.1 Clients

Les clients du guichet de la collecte sont les gestionnaires d'enquête responsables des enquêtes (entreprises, ménages, agricoles et institutionnelles).

[1]Anie Marcil, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (anie.marcil@statcan.gc.ca).

Pour les nouvelles enquêtes, le guichet de la collecte permet de déterminer certaines spécifications d'enquête, d'en apprendre davantage sur les nouveaux services de collecte disponibles et d'identifier le processus de collecte et les partenaires de la collecte qui seront concernés. Pour les enquêtes existantes qui subissent plusieurs changements, c'est une façon de revoir ces spécifications d'enquête afin de déterminer les changements qui sont requis.

Les évaluations du guichet de la collecte s'adaptent à tous les modes de collecte (interview téléphonique assistée par ordinateur (ITA0), interview sur place assistée par ordinateur (IPAO), papier ou encore collecte électronique).

2.2 Partenaires de la collecte

Les principaux partenaires de la collecte à Statistique Canada et leurs responsabilités sont brièvement présentés dans cette section :

La Division de la planification et de la gestion de la collecte est responsable :

- des procédures de la collecte de l'enquête ;
- de la formation des intervieweurs ;
- du suivi du progrès de la collecte ;
- de la liaison avec les clients et les bureaux régionaux ;
- de la coordination de la capacité avec les bureaux régionaux ; *etc*.

Les bureaux régionaux sont responsables :

- de l'embauche des intervieweurs ;
- du plan de production ;
- de l'horaire des intervieweurs ;
- de la collecte de données auprès des répondants ;
- des comptes rendus après la collecte ; *etc*.

La Division des systèmes et de l'infrastructure de collecte est responsable :

- de la planification, du développement, du soutien ainsi que de la maintenance des systèmes de collecte de Statistique Canada pour tous les modes de collecte.

La Division des opérations et de l'intégration est responsable :

- de la conception de produits d'enquête et services d'impression ;
- des services du centre de distribution ;
- de la saisie des données ;
- de l'imagerie ;
- du codage ;
- du traitement des données administratives ; *etc*.

La Division des communications est responsable :

- de la conception des brochures ;
- de la mise à l'essai des enquêtes électroniques ;
- du Centre d'assistance pour les répondants d'enquêtes électroniques ; *etc*.

La Division de la diffusion et la Division de l'ingénierie des systèmes sont responsables :

- de la mise à l'essai des enquêtes électroniques ; *etc*.

3. Portée et avantages

3.1 Portée

Dans le cas d'une nouvelle enquête et du remaniement d'une enquête existante, le guichet de la collecte est responsable de l'étape d'évaluation préliminaire de la collecte des données de l'enquête. C'est le point de contact entre les divisions clientes et les partenaires de la collecte. Lors de cette étape préliminaire, les spécifications de l'enquête sont examinées, le déroulement du processus de la collecte des données est déterminé, la capacité requise est évaluée, les estimations budgétaires sont préparées et les risques ainsi que les recommandations sont partagés avec le client.

Le guichet de la collecte fournit uniquement ses services durant la planification préliminaire de la collecte. Une fois la confirmation reçue que l'enquête ira de l'avant, une rencontre de transition est organisée entre les partenaires de la collecte et le client afin de faire une synthèse de l'évaluation préliminaire et donner un coup d'envoi au développement de l'enquête. L'apport du guichet de la collecte est par la suite terminé.

3.2 Avantages

Les avantages de ce service sont les suivants :

- Un seul point de contact pour toutes les divisions clientes qui souhaitent obtenir des services de collecte. Les gestionnaires d'enquête ont seulement un endroit à contacter pour obtenir les informations concernant les activités de la collecte.
- Une approche plus normalisée qui donne une vision globale et harmonisée de l'ensemble des activités de collecte. Ce service fait en sorte que les hypothèses principales utilisées sont les mêmes d'un partenaire de la collecte à l'autre et que l'intégration de ces informations dans les études de faisabilité globale est uniforme et cohérente à l'ensemble des enquêtes de Statistique Canada.
- Une meilleure communication et une meilleure planification des activités de collecte réduisent les risques, les répercussions négatives, les possibilités d'erreurs ou de surprises lors du développement et du déroulement de l'enquête.
- Ce service permet d'avoir et de maintenir un centre d'expertise en ce qui concerne les activités de la collecte à Statistique Canada.
- De plus, différents scénarios ou options peuvent être évalués permettant aux gestionnaires d'enquête de prendre une décision plus éclairée sur la meilleure stratégie de collecte tout en respectant les objectifs principaux de l'enquête.

4. Processus et Rapport de l'évaluation de la faisabilité de la collecte

4.1 Processus du guichet de la collecte

Le processus du guichet de la collecte est le suivant :

- Les clients doivent soumettre une demande par un système central de la gestion des demandes corporatives accompagnées d'un formulaire des spécifications de l'enquête.
- Ensuite, le guichet de la collecte rencontre le client afin de passer en revue les objectifs et les spécifications de l'enquête. Les rôles et responsabilités sont précisés et une ou plusieurs stratégies de collecte sont définies. À cette étape initiale de la planification, les spécifications de l'enquête sont généralement vagues et peuvent changer rapidement. Il est donc normal d'avoir plusieurs discussions entre le guichet de la collecte et les divisions clientes afin d'éclaircir certains points ou questions en suspens.
- Les spécifications recueillies lors de ces rencontres sont par la suite communiquées aux partenaires de la collecte. Ceux-ci doivent fournir leurs estimations budgétaires, la disponibilité de leurs ressources ainsi que des recommandations ou risques associés à ces stratégies de collecte selon les spécifications fournies.
- La préparation du Rapport d'évaluation de la faisabilité de la collecte détaillé est rédigée en collaboration avec les partenaires de la collecte et ce rapport est remis au client.
- Une rencontre de transition est par la suite organisée lorsque le guichet de la collecte reçoit la confirmation que l'enquête peut aller de l'avant. L'intervention du guichet de la collecte se termine à ce moment.

Comme mentionné ci-dessus, le guichet de la collecte travaille avec des paramètres qui changent constamment et avec des terminologies qui peuvent porter à confusion. Il est ainsi important de prévoir suffisamment de temps pour permettre au guichet de la collecte d'effectuer ces évaluations et d'établir des hypothèses de travail fiables et réalistes. De plus, il est fréquent de devoir refaire les rapports/évaluations avec de nouvelles spécifications d'enquête.

Avec l'expérience acquise pendant les deux dernières années, le guichet de la collecte a démontré une flexibilité à répondre rapidement aux besoins des clients tout en respectant les étapes du processus.

4.2 Rapport de l'évaluation de la faisabilité de la collecte

Le Rapport d'évaluation de la faisabilité de la collecte comprend les sections suivantes :

- un sommaire de l'ensemble des activités de collecte pour l'enquête en question ;
- une liste des spécifications et hypothèses utilisées pour l'évaluation ;
- un survol des activités de chaque partenaire de la collecte ;
- une évaluation de la capacité de chaque partenaire ;
- une section qui explique les risques et recommandations associés à certains aspects de la collecte de cette enquête ;
- une description détaillée des coûts préliminaires de chaque partenaire fondée sur des spécifications préliminaires ;
- et quelques dates importantes à respecter, dont celle de la confirmation que l'enquête ira de l'avant.

5. Les succès du guichet de la collecte

Depuis la création du guichet de la collecte, il y a plusieurs réalisations importantes à souligner :

- Le guichet de la collecte est maintenant reconnu comme un centre d'expertise concernant les activités de collecte. Les membres du guichet de la collecte ont acquis (et continuent d'acquérir) une expertise depuis les deux dernières années, ce qui rend les consultations avec les divisions clientes plus productives et plus rapides. Le guichet de la collecte peut répondre directement aux questions des divisions clientes sans l'intervention des partenaires de la collecte, ce qui permet à ceux-ci de se concentrer sur les autres priorités de l'organisation.
- À ce jour, le guichet de la collecte a évalué plus de 140 enquêtes utilisant différents modes et stratégies de collecte, par exemple : enquêtes multimodes (ITAO/IPAO), enquêtes électroniques, enquêtes unimodes, enquêtes post-censitaires, enquêtes longitudinales.
- Les rôles et responsabilités sont clairement déterminés au moment de l'évaluation préliminaire de la collecte, ce qui facilite la gestion des activités de développement de l'enquête. Le client connait les activités qui seront effectuées par les partenaires, ce qui lui permet de se concentrer sur d'autres aspects du développement et de la mise en œuvre de son enquête. Par le fait même, les relations entre les partenaires de la collecte et les divisions clientes se sont améliorées.
- Les rapports d'évaluation de la faisabilité de la collecte permettent de bien cerner les répercussions de certaines stratégies de collecte, ce qui aide les gestionnaires d'enquête à prendre une meilleure décision.
- Le guichet de la collecte contribue au projet de l'Architecture opérationnelle du Bureau en assignant les activités de collecte aux bons centres d'expertise, ce qui aide à l'élimination du dédoublement des activités de collecte à Statistique Canada.
- Le formulaire normalisé de spécifications permet de recueillir les informations de base sur l'enquête en question. Par ailleurs, ce formulaire s'est avéré être une liste de contrôle fort utile pour rappeler aux divisions clientes toutes les étapes du processus de collecte ainsi que les ressources nécessaires qui devront être comprises dans le projet.
- La création d'un outil de planification préliminaire aide les gestionnaires d'enquête de Statistique Canada lors de discussions préliminaires avec les clients externes selon des scénarios prédéfinis.

5.2 Succès pour la méthodologie

Pour l'univers de la méthodologie, ce nouveau service permet :

- d'obtenir de l'information sur les résultats d'enquêtes similaires ;
- de se renseigner sur les services de collecte disponibles ;
- de connaître les fichiers/informations qui doivent être fournis aux partenaires afin de débuter le développement de l'enquête ;
- de préciser les rôles et responsabilités de chacun ;
- d'avoir une vision globale et harmonisée des activités de collecte et de recueillir le point de vue des experts en collecte ;

- le guichet de la collecte permet finalement d'effectuer des évaluations préliminaires fondées sur plusieurs scénarios. Les méthodologistes proposent aux gestionnaires d'enquête quelques stratégies de collecte qui sont par la suite évaluées et comparées en terme d'incidence sur les activités de collecte (capacité, coûts, échéances, *etc.*).

6. Conclusion

En conclusion, le guichet de la collecte a été conçu pour :

- fournir des services de consultation en matière de collecte des données aux divisions clientes ;
- coordonner les services de collecte des données entre les divisions clientes et les partenaires de la collecte durant l'étape initiale d'évaluation de la faisabilité de la collecte ;
- préparer des rapports d'évaluation de la faisabilité de la collecte, y compris des estimations de coûts préliminaires, des analyses de la capacité et l'identification des risques liés aux stratégies de collecte proposées ;
- jouer un rôle d'intégration :
    - s'assurer de la participation de tous les partenaires de la collecte concernés ;
    - chercher à établir des spécifications intégrées de haut niveau pour les projets de collecte des données ;
    - veiller à la normalisation des estimations de coûts et de la planification de la capacité ainsi qu'à la constance des services offerts aux clients ;
    - servir de point de contact unique pour les divisions clientes qui sont à la recherche de services de collecte ;
- réduire le plus possible les chevauchements et combler les lacunes dans les rôles et responsabilités de collecte actuels.

Le guichet de la collecte joue un rôle très important et primordial dans la planification préliminaire des enquêtes de Statistique Canada.

Remerciements

L'auteure remercie Milana Karaganis pour son encadrement et ses encouragements lors de la création de ce nouveau service, ainsi qu'Edward Joseph et Stuart McFarlane pour leurs efforts continus afin d'améliorer le processus et de répondre aux besoins des clients.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Mise en œuvre de procédures de contrôle de la qualité au centre national des opérations du NASS

Jeffrey M. Boone, Joseph L. Parsons, Shari R. Feld, Jenna N. Levy et Kristie L. Flaherty[1]

Résumé

En août 2011, le National Agricultural Statistics Service (NASS) a centralisé la plupart de ses activités de collecte de données d'enquête dans un nouveau centre national des opérations (National Operations Center ou NOC). Celui-ci est responsable de la collecte téléphonique des données, du traitement des questionnaires papier, de la tenue à jour de la base de sondage du NASS et d'autres activités de collecte connexes. Le NOC a été créé dans le but de réduire les sources d'erreurs inhérentes aux activités de collecte, d'améliorer la qualité des données et de réduire les coûts des opérations. L'analyse de rentabilité justifiant la création du NOC mentionnait la nécessité de mettre en place un programme complet de contrôle de la qualité au nouveau centre pour favoriser l'excellence des résultats. Le présent article décrit la situation actuelle du programme de contrôle de la qualité au NOC, y compris les méthodes de mesure et de surveillance. Le processus de mise en œuvre de mesures de contrôle de la qualité dans diverses fonctions est également discuté, de même que les futures possibilités d'amélioration et les leçons apprises.

Mots clés : Contrôle de la qualité ; assurance de la qualité ; centre d'appels ; mesures ; surveillance.

1. Introduction

1.1 Vue d'ensemble

Le contrôle et l'assurance de la qualité représentent un important aspect des activités de tout organisme. L'Organisation internationale de normalisation définit la qualité comme « le degré auquel un ensemble de caractéristiques inhérentes satisfait les exigences » (2005). L'assurance de la qualité consiste à s'assurer que les cibles ou les objectifs établis sont atteints. Le contrôle de la qualité consiste à surveiller ou à évaluer un produit ou un processus pour s'assurer que les normes souhaitées sont satisfaites. Bien que nous utilisions le terme « contrôle de la qualité » dans la suite du présent article, l'assurance de la qualité et le contrôle de la qualité sont tous deux examinés.

Le présent article traite de la mise en œuvre de procédures de contrôle de la qualité au centre national des opérations du National Agricultural Statistics Service (NASS). Les progrès réalisés à ce jour et les objectifs futurs sont présentés, ainsi qu'une analyse documentaire des articles traitants du contrôle de la qualité et de son application à la collecte des données d'enquête.

Un programme de contrôle de la qualité efficace permet de s'assurer qu'un processus donne des résultats de qualité pendant toute sa durée. Selon le Guide to the Project Management Body of Knowledge, satisfaire aux exigences de qualité a notamment pour avantages de réduire le nombre de reprises, d'augmenter la productivité, de diminuer les coûts et d'accroître la satisfaction des clients et des intervenants (2008). Dans le contexte des méthodes d'enquête, cela signifie moins d'appels de suivi, moins de temps consacré à la vérification d'enquêtes et à la recherche de données inhabituelles, un plus grand nombre d'enregistrements contactés (un plus grand nombre d'interviews), des coûts globaux plus faibles, ainsi qu'une plus grande assurance d'obtenir des données de haute qualité.

Outre les économies de coûts qu'il procure, l'établissement d'un programme de contrôle de la qualité peut permettre de déterminer les volets d'un processus qui fonctionnent bien et sur lesquels il convient d'insister. Cela pourrait aboutir à des suggestions quant à de futures activités de formation et de futures mesures incitatives et récompenses à l'intention des employés. Le système permettra aussi de déceler les problèmes dans le processus, ce qui mènera aussi

[1]Jeffrey M. Boone, Joseph L. Parsons, Shari R. Feld, Jenna Levy et Kristie Flaherty, National Agricultural Statistics Service, United States Department of Agriculture, 3251 Old Lee Hwy., Rm. 305, Fairfax, Virginie 22030, jeff.boone@nass.usda.gov.

à des suggestions concernant l'accroissement de la formation, ainsi qu'à d'éventuelles mesures disciplinaires ou d'autres remèdes. La maintenance d'un programme de contrôle de la qualité peut également aider à reconnaître les possibilités d'amélioration du processus, telles que Lean Six Sigma (par exemple, voir George [2003]). Cela, à son tour, peut conduire à des possibilités d'amélioration continue tout au long de la vie du programme. Bien que des améliorations puissent être appliquées à n'importe quelle étape d'un processus, les mesures de paramètres déjà existantes permettront la mise en œuvre rapide et sans heurts de l'amélioration du processus.

1.2 Analyse documentaire

Le contrôle de la qualité est une question qui a été abordée dans de nombreux travaux relevant de la méthodologie d'enquête. Laflamme, Mayden et Miller décrivent la gestion active (contrôle de la qualité) comme un ensemble de plans et d'outils utilisés pour recueillir de manière efficace et pleinement intégrée des renseignements sur les processus de collecte des données afin de mieux améliorer les pratiques d'administration des enquêtes. La gestion active s'entend de « la surveillance des progrès, l'analyse rapide des indicateurs, la découverte des problèmes, la mise en œuvre et la communication de mesures correctives et l'évaluation des réussites » (2008).

Les « paradonnées » sont souvent utilisées pour décrire l'information au sujet du processus de collecte des données. Ces paradonnées sont équivalentes aux mesures de contrôle de la qualité lorsqu'elles sont appliquées au processus de collecte de données d'enquête. Selon Bates et ses collaborateurs, les divers types de paradonnées comprennent les enregistrements d'appel, les observations des intervieweurs et des répondants, les enregistrements audio des interactions entre les intervieweurs et les répondants, les éléments de données produits par les questionnaires assistés par ordinateur, tels que les temps de réponse et les touches frappées, *etc*. Les paradonnées sont utilisées pour mesurer la qualité d'une enquête dans un environnement de production et pour gérer cette dernière en vue d'optimiser la qualité et de réduire les coûts au minimum. Les paradonnées peuvent être utilisées pour surveiller le travail sur le terrain, analyser la non-réponse, élaborer des plans de collecte adaptatifs, faciliter l'évaluation de l'erreur de mesure, corriger la non-réponse, et améliorer la vérification et le codage (Bates, Dalhammer, Phipps, Safir et Tan, 2010).

Lepkowski et ses collaborateurs décrivent quatre paradigmes ayant trait aux paradonnées, à savoir l'effort, l'échantillon actif, la productivité et l'équilibre de l'ensemble de données. Dans le cas de la National Survey of Family Growth, en ce qui concerne l'effort, plusieurs sources de données sont mesurées : nombre d'intervieweurs qui travaillent, nombre d'heures, pourcentage d'heures productives, nombre d'appels par jour, nombre d'appels par heure, pourcentage d'appels durant les heures de pointe, et appels préliminaires par opposition aux appels principaux. Les auteurs ont également examiné le pourcentage d'appels avec signal occupé, le pourcentage d'appels admissibles, le pourcentage de numéros hors service, le nombre de cas de non-contact, le nombre moyen d'appels, le pourcentage de cas nécessitant huit appels ou plus, le pourcentage d'immeubles verrouillés, le pourcentage de cas résistants, le pourcentage de rendez-vous fermes et la propension à répondre. Pour ce qui est de la productivité, ils ont mesuré le nombre d'interviews, le nombre cumulé d'interviews, le nombre d'heures par interview, et le nombre d'appels par interview. Enfin, pour l'équilibre de l'ensemble de données, ils ont mesuré le taux de réponse, le pourcentage de familles avec enfants, le pourcentage de répondants sexuellement actifs et les taux de groupe (2010).

Lyberg explique que les paradonnées peuvent fournir des mises à jour continues des progrès et des vérifications de la stabilité (surveillance, renseignements permettant d'améliorer les processus de longue durée), de la qualité des produits (analyse des variations ayant une cause spéciale ou une cause commune) et des renseignements en vue d'apporter des modifications méthodologiques (découverte et élimination des causes fondamentales des problèmes). Les paradonnées sont également essentielles à l'élaboration de plans de collecte adaptatifs et à la fourniture de données en vue de procéder à des changements organisationnels. Elles peuvent être utilisées pour comprendre les variations et pour déterminer le coût de la mauvaise qualité et du gaspillage. Les paradonnées sont de nature multivariée et doivent parfois être combinées pour être pertinentes. La création d'archives de paradonnées permet de reprendre les analyses afin de mieux comprendre quels sont les éléments clés (2009).

2. Le NASS et le centre national des opérations

2.1 À propos du NASS et du centre national des opérations

Le NASS fournit des statistiques à jour, exactes et utiles en vue de servir l'agriculture aux États-Unis. Le NASS réalise des centaines d'enquêtes par année, effectue le Recensement de l'agriculture tous les cinq ans, et fournit des données sur les produits agricoles des États Unis ainsi que des données qui servent à déterminer les prix des produits de base. À l'heure actuelle, l'approche de collecte des données, relativement décentralisées, s'appuie sur 46 bureaux locaux répartis à travers les États-Unis. Le NASS a commencé à centraliser ses procédures de collecte des données en ouvrant un centre national des opérations (National Operations Center ou NOC) à St. Louis, dans le Missouri, en août 2011. L'un des objectifs de ce projet est de réduire les sources d'erreurs inhérentes aux activités de collecte des données, d'améliorer la qualité des données et de réduire les coûts des opérations. Le centre remplira diverses fonctions, y compris l'interview téléphonique, le traitement des questionnaires papier, la tenue à jour de la base de sondage du NASS, la formation, l'élaboration des enquêtes, la programmation Blaise et la programmation du système d'enquête en ligne.

Le NASS a déjà mis en place certaines procédures de contrôle de la qualité. Pour le dénombrement par téléphone, elles comprennent les rapports quotidiens sur les taux de réponse, les rapports après enquête du rendement des agents recenseurs, les rapports après enquête sur les coûts, les primes à l'intention des intervieweurs contractuels, la surveillance des agents recenseurs, les appels de suivi et les évaluations du rendement des agents recenseurs. La surveillance des agents recenseurs et les appels de suivi, qui sont effectués par les superviseurs, comprennent le remplissage à la main d'un questionnaire imprimé et son archivage. En ce qui concerne le traitement des questionnaires, nombre d'entre eux sont vérifiés à la main. Pour le moment, le NASS ne procède pas au suivi du nombre de vérifications ou du temps nécessaire pour vérifier les questionnaires. Certains de ces questionnaires, comme ceux du recensement de l'agriculture et de la Cash Rents Survey, sont traités au centre national des traitements (National Processing Center ou NPC) du Bureau of the Census, à Jeffersonville, dans l'Indiana. Le NPC utilise un système électronique de suivi qui fournit de nombreux rapports, dont des renseignements sur le suivi des documents, tels que l'emplacement d'un questionnaire ou le nombre de questionnaires aux diverses étapes du processus.

2.3 Contrôle de la qualité au centre national des opérations

Un objectif central de la création du NOC est l'amélioration de la qualité des données. En réussissant à mettre en œuvre un programme de contrôle de la qualité, le NASS pourra continuer à évoluer en se fondant sur les renseignements tirés des processus de collecte des données du NOC. Comme nous l'avons mentionné plus haut, le NASS suit certaines procédures de contrôle de la qualité, mais celles-ci pourraient être inefficaces dans le contexte du NOC. En premier lieu, à l'heure actuelle, la surveillance des évaluations et des appels de suivi est faite sur papier. Étant donné le grand nombre d'agents recenseurs au NOC, il faudra créer de nombreux rapports de surveillance qui rendront le processus de surveillance inefficace; autrement dit, les méthodes informelles utilisées à l'heure actuelle par le NASS ne pourront pas être simplement étendues afin de répondre à la quantité de travail au NOC. En outre, un système normalisé garantira la qualité des données. Le traitement des questionnaires comprend aussi un programme de contrôle de la qualité qui décèlera les problèmes que pose le système. Afin de s'assurer de la normalisation des bases de sondage du NASS, les procédures de tenue à jour de ces bases devront également faire l'objet d'un contrôle de la qualité afin de s'assurer que les bases de sondage soient cohérentes et de haute qualité.

3. Progrès

3.1 Étapes

Les étapes en vue de créer un système de contrôle de la qualité sont simples. Premièrement, il convient de déterminer les mesures à utiliser pour surveiller la qualité du processus de collecte des données. Tant la productivité que la qualité des données doivent être prises en considération. Ensuite, il faut élaborer une méthode en vue de saisir ces mesures. Cela comporte à la fois l'élaboration d'une interface ou d'un système pour enregistrer l'information nécessaire pour le calcul de ces mesures et la détermination d'un emplacement pour l'enregistrement de cette

information. Enfin, il faut établir une méthode pour afficher ces mesures. Cette sorte de « tableau de bord » fournira aux utilisateurs du système l'information nécessaire pour prendre rapidement des décisions fondées sur des données. Ce processus semble simple, mais dans les environnements opérationnels, de nombreuses difficultés se posent. Ces problèmes seront examinés plus loin dans le présent article.

3.2 Mesures

La première étape en vue de créer un système de contrôle de la qualité consiste à déterminer les mesures qui seront utilisées pour surveiller la qualité du processus de collecte des données. Les mesures devraient donner la capacité d'obtenir une réponse à une question qui permettra de prendre une décision objective, fondée sur des données. Les mesures devraient fournir une évaluation de la production des membres du personnel, tels que les intervieweurs par téléphone, ainsi que de la qualité des données. Les mesures relatives à la productivité, telles que le nombre d'appels, la durée des appels, et le nombre de refus, ainsi que le coût, sont assez simples à déterminer; par contre, celles ayant trait à la qualité des données sont plus difficiles à obtenir. Ces mesures offriront un moyen de détecter les problèmes concernant le processus de collecte des données ainsi que la qualité des données recueillies. Si les problèmes sont décelés, des mesures peuvent être prises pour les corriger ou pour réduire au minimum leur manifestation.

Lorsque l'on choisit des indicateurs de productivité, il est important de tenir compte des caractéristiques qui suivent. Les indicateurs doivent être : 1) facilement compris, 2) mesurables et comparables à n'importe quel stade de la collecte des données, 3) mis à jour systématiquement au cours de la collecte des données, 4) pertinents, interprétables et comparables à divers niveaux d'agrégation. Ces indicateurs peuvent être répartis en trois catégories en se fondant sur les durées, les nombres de tentatives/d'appels, la durée par unité ou la durée par interview achevée (Laflamme, 2009).

De nombreuses mesures peuvent être utilisées au NOC. Le nombre d'interviews achevées utilisables par heure travaillée, le pourcentage de refus, la durée moyenne d'un appel et le nombre de questionnaires papier traités pendant une période donnée en sont quelques exemples. Certaines mesures peuvent être considérées à la fois comme des mesures de la productivité et de la qualité des données, telles que la durée moyenne d'un appel. Ainsi, la durée de l'appel renseigne sur la quantité de travail effectué et est donc une mesure de productivité. Cependant, si la durée moyenne des appels donnant lieu à des interviews complètes est anormalement faible, la qualité des données recueillies peut être compromise.

Une question fréquente lorsque l'on détermine les mesures à utiliser est celle du nombre de mesures à surveiller, ainsi que l'importance des mesures par rapport à la question posée. La surveillance d'un trop grand nombre de mesures peut non seulement demander trop de temps et d'effort, mais aussi signaler un problème alors qu'il n'en existe pas en réalité. Le nombre de fausses alarmes sera d'autant plus élevé que les domaines dans lesquels on recherche des problèmes sont nombreux. Une méthode fréquente pour pallier ce problème consiste à créer un indice combinant l'information provenant de plusieurs variables. Cette approche permet au superviseur d'observer un plus petit nombre de variables sans perdre aucune information.

Un autre problème que posent les mesures dans les centres d'appels est inhérent à la difficulté de certains cas. Par exemple, certaines régions géographiques produisent des taux de réponse naturellement plus faibles que d'autres. Dans de telles circonstances, il serait erroné de tenir compte uniquement des taux de réponse des agents recenseurs dans les diverses régions, puisque ceux auxquels sont attribués les cas plus difficiles auront probablement des taux de réponse plus faibles. Donc, des études ont été menées en vue de créer un indice de « difficulté » ou de « propension à répondre » (voir, par exemple, Laflamme et St-Jean, 2011). Un tel indice permettrait au superviseur de noter la difficulté des cas durant l'examen de la productivité des agents recenseurs. Le NASS effectue à l'heure actuelle des travaux de recherche en vue d'élaborer un indice de propension à répondre pour classer les cas dont la probabilité de non-réponse pourrait être élevée.

3.3 Stockage des données

Un aspect important d'un système de contrôle de la qualité est le stockage des données. L'espace physique peut parfois poser problème, mais au-delà du simple concept de taille il y a celui de la maintenance de la base de données : il s'agit notamment de veiller à ce que les éléments de données qu'il est permis de stocker dans la base de données soient choisis de façon à ce qu'aucun renseignement ne figure en double et qu'il soit possible de faire référence à

chaque variable de manière normalisée. Le NASS élabore à l'heure actuelle une nouvelle base de données centralisée pour le stockage des données sur les travaux en cours. Statistique Canada utilise aussi une base de données centralisée comme entrepôt de paradonnées pour stocker tous les renseignements nécessaires. Bien que la création d'une telle base de données soit une très bonne pratique, cela cause un goulot d'étranglement dans le développement du système qui s'appuie sur la base de données pour l'entreposage et l'accès aux données. Cet obstacle a été très difficile à surmonter dans la création du système de contrôle de la qualité.

Un autre aspect important de l'utilisation de la base de données centralisée qui vient d'être déployée consiste à s'assurer que les données provenant de divers systèmes se trouvent sous la même forme normalisée quand elles sont stockées dans la base de données. Par exemple, si les différents systèmes n'utilisent pas le même identificateur d'intervieweur, il n'est pas possible de fusionner correctement les données provenant de ces systèmes et de calculer correctement les mesures résultantes. Il convient de souligner qu'en raison de l'architecture de la base de données, il n'est pas toujours possible de transférer toute l'information nécessaire dans cette base de données. Donc, l'application de contrôle de la qualité pourrait devoir extraire certaines données d'autres bases de données.

3.4 Surveillance

Après avoir déterminé les mesures nécessaires ou souhaitées, il faut choisir une méthode pour obtenir ces mesures. Dans le cas du dénombrement par téléphone, cela comprend l'enregistrement des mesures de production, comme la durée d'un appel téléphonique ou le taux de refus, dans une base de données. Nombre de ces mesures peuvent être obtenues à partir de systèmes d'interview téléphonique assistée par ordinateur, tels que le système Blaise. Les mesures concernant le rendement des agents recenseurs qui ne sont pas saisies automatiquement, par exemple les évaluations données par le personnel de surveillance, doivent être enregistrées. D'autres renseignements, comme les compétences, la disponibilité, la fiche de présence et la durée de l'emploi de chaque agent recenseur, feront l'objet d'un suivi. Pour les autres fonctions du NOC, le suivi des mesures peut être effectué au moyen de logiciels, tels que le système de suivi et de contrôle et le système de saisie à partir d'images ou de questionnaires papier qui sont développés pour le traitement des questionnaires, et les systèmes internes de tenue à jour des bases de sondage, l'assistant de tenue à jour améliorée des listes et les opérations de tenue à jour améliorée des listes. Les données provenant de ces systèmes devront également être disponibles dans une base de données.

Une fois que les données seront accessibles, une application informatique sera utilisée pour illustrer les données à l'intention des membres du personnel qui doivent les utiliser. Ce tableau de bord présentera les données sous forme de tableaux descriptifs, ainsi que sous forme de graphiques et de figures. Il est important de veiller à ce que les gestionnaires des opérations sur le terrain puissent « creuser dans les fichiers de suivi pour examiner les renseignements détaillés sur les interviews » (O'Reilly, 2010). De nombreuses applications logicielles ont la capacité de créer des tableaux de bord. Quelque chose d'aussi simple d'une page Web interactive pourrait également suffire. Le choix du logiciel dépendra des ressources disponibles, y compris les licences existantes, le coût des nouveaux logiciels et les connaissances des employés.

4. Problèmes

À l'heure actuelle, le NASS développe plusieurs nouveaux systèmes. Nombre de ceux-ci sont très demandés et sont nécessaires à l'élaboration du NOC. Un système de contrôle de la qualité est très important, mais n'est pas souvent considéré comme une exigence en vue de lancer les opérations. Donc, les employés spécialisés nécessaires pour le développement du système sont affectés à ces autres systèmes ayant une « plus grande priorité » et, comme ces employés sont très demandés, ils représentent une ressource souvent difficile à obtenir. Malheureusement, certaines étapes doivent être achevées avant que les étapes subséquentes puissent débuter, autrement dit certaines étapes en vue d'achever le projet font partie du chemin critique.

Les données provenant de différents systèmes doivent être normalisées avant d'être stockées dans la base de données centralisée. Puisque de nombreux systèmes du NASS ont été créés par des groupes différents (à l'interne et à l'externe) et à différentes périodes, les identificateurs uniques d'intervieweur n'ont pas le même format dans tous les systèmes. Donc, lorsque les données sont entrées dans la base de données centralisée, il faut veiller à ce qu'elles soient fusionnées correctement au moyen d'une certaine transformation ou d'un certain script en vue d'apparier les données sur les intervieweurs.

Il est souvent non seulement difficile de préparer la base de données pour le stockage des données, mais aussi d'acquérir le logiciel pour créer le tableau de bord. Le processus ne consiste pas simplement à choisir le logiciel, à l'acheter et à l'installer. Son déploiement dans un réseau à l'échelle d'une société ou d'un organisme est une tâche immense. Dans la plupart des cas, il faut obtenir l'approbation du personnel des services de sécurité et d'architecture d'entreprise, démarches qui peuvent prendre beaucoup de temps. Même si le logiciel est installé et disponible, il se peut qu'il ne soit pas configuré correctement pour cet usage particulier.

5. Conclusion

La création d'un système de contrôle de la qualité peut être une tâche formidable. Les étapes pour créer le système, c'est-à-dire déterminer les mesures qu'il convient de surveiller, choisir la méthode de surveillance et développer l'interface utilisateur pour se servir du système, sont simples. Cependant, l'environnement dans lequel le système est développé peut rendre ces simples étapes bien plus compliquées qu'il n'apparaissait au départ. Il est important de noter que la patience (ainsi que de solides compétences en négociation) est une ressource essentielle à la création d'un système de contrôle de la qualité. Bon nombre des étapes de la création du système requièrent l'intervention de beaucoup d'autres personnes et de systèmes sur lesquels on n'a pas de contrôle, de sorte que se montrer patient tout en pressant pour que la tâche s'accomplisse est une attitude importante à adopter dans ce genre d'entreprise, de même que dans de nombreux autres projets recoupant diverses fonctions.

Une fois que tous les obstacles sont surmontés et que le système est en place (pas nécessairement entièrement), de nombreux avantages peuvent être glanés. Le système peut ouvrir la voie à d'éventuelles améliorations continues du processus qui n'auraient pas été découvertes sans le suivi des mesures. Si un système de contrôle de la qualité est établi et saisit l'information nécessaire sur les mesures souhaitées, l'étape de la prise des mesures du processus est déjà accomplie. Un autre avantage tient à la possibilité d'optimiser globalement le processus complet de collecte des données à tous les emplacements de collecte par opposition à l'optimisation locale qui ne serait effectuée qu'à des emplacements de collecte individuels. La saisie de ces mesures donne aussi l'occasion de réaliser des économies, par exemple, en optimisant la gestion de la collecte des données et en réduisant le besoin de vérification. D'autres avantages comprennent la détermination des domaines demandant une formation ciblée, l'amélioration des procédures de recrutement, le développement des enquêtes et une application éventuelle des plans de collecte adaptatifs.

Bibliographie

Bates, N., Dalhammer, J., Phipps, P., Safir, A. et L. Tan. (2010), « Assessing Contact History Paradata Quality Across Several Federal Surveys », *Proceedings of the American Statistical Association 2010 Joint Statistical Meetings*, American Statistical Association, disponible à https://www.amstat.org/sections/srms/Proceedings/y2010/Files/306005_55654.pdf.

George, M. (2003), *Lean Six Sigma for Service: How to Use Lean Speed and Six Sigma Quality to Improve Services and Transactions*, New York: McGraw-Hill.

Laflamme, F., Mayden, M et A. Miller (2008), « Using Paradata to Actively Manage Data Collection Process », *Proceedings of the American Statistical Association 2008 Joint Statistical Meetings*, American Statistical Association, disponible à http://www.amstat.org/sections/srms/proceedings/y2008/Files/300608.pdf.

Laflamme, F. (2009), « Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada », *Proceedings of the 57th Session of the International Statistical Institute*, disponible à http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/0049.pdf.

Laflamme, F. et H. St-Jean (2011), « Proposed Indicators to Assess Interviewer Performance in CATI Surveys », Communication sur invitation lors du 2011 Joint Statistical Meetings, Miami, Floride.

Lepkowski, J.M., Axinn, W., Kirgis, N., Brady T. West, Shonda Kruger Ndiaye, Mosher, W. et R.M. Groves (2010), « Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection », NSFG Survey Methodology Paper Series, no 10-012, disponible à http://www.psc.isr.umich.edu/pubs/pdf/ng10-012.pdf.

Lyberg, L. (2009), « The Paradata Concept in Survey Research », présentation au NCRM Paradata Network, London, Royaume-Uni, 24 août 2009, disponible à http://www.natcen.ac.uk/ncrm-paradata-network/docs/Lyberg-paradata-concept.ppt.

O'Reilly, J. (2010), « Paradata and Blaise: a Review of Recent Applications and Research », *Proceedings of the 12th International Blaise Users Conference*, Riga, Lettonie, disponible à http://ibuc2009.blaiseusers.org/papers/7d.pdf.

Organisation internationale de normalisation (2005), ISO 9000:2005, *Systèmes de management de la qualité – Principes essentiels et vocabulaire*, Genève: Presse ISO.

Project Management Institute (2008), *A Guide to the Project Management Body of Knowledge*, 4e édition, Newton Square, PA: Project Management Institute, Inc.

SÉANCE 8A

UTILISATION DE MÉTHODES ET D'OUTILS NORMALISÉS POUR LE TRAITEMENT POST-COLLECTE

Normalisation du traitement des données après la collecte dans les enquêtes-entreprises à Statistique Canada

Serge Godbout[1]

Résumé

Statistique Canada a entrepris une refonte de ses enquêtes auprès des entreprises. Ceci représente une occasion unique de développer un nouveau cadre pour le traitement qui est bâti en fonction de la vérification sélective et de la gestion active de la collecte. Statistique Canada propose de mettre en œuvre un ensemble complet de fonctionnalités applicables sur des données partiellement recueillies pour produire des estimations en cours de collecte. Associé à ces estimations, un ensemble d'indicateurs de qualité choisis minutieusement sera également produit. À partir de ces indicateurs, les décisions relatives à la gestion active de la collecte seront prises, y compris la création d'une liste d'unités prioritaires pour le suivi de la non-réponse et de règles échouées de validation. Cette présentation donnera une idée des défis liés au développement des processus pour produire des estimations et des indicateurs de qualité fondés sur un échantillon partiellement recueilli et décrira la stratégie de gestion active de la collecte.

Mots clés : Normalisation ; traitement des données ; indicateurs de qualité ; gestion active de la collecte.

1. Introduction

En 2010, Statistique Canada a lancé l'initiative de l'Architecture opérationnelle du Bureau (AOB). Des pressions croissantes sur les finances de l'organisme ont mené à une revue de ses méthodes et systèmes, afin d'identifier des avenues pour réaliser des économies de coûts, rehausser l'assurance de la qualité et améliorer la capacité de réaction des nouveaux programmes statistiques. Une composante clé de cette initiative est le développement et l'utilisation obligatoire de services intégrés génériques et de systèmes généralisés pour l'échantillonnage, la collecte, le traitement, la publication et l'entreposage des informations statistiques des programmes d'enquêtes auprès des ménages et des entreprises. Afin de réaliser ces objectifs, Statistique Canada a entrepris une refonte majeure de ses programmes statistiques auprès des entreprises. Le Programme intégré sur la statistique des entreprises (PISE) a pour but de mettre en œuvre une plateforme commune pour un grand nombre de ses enquêtes auprès des entreprises. D'ici 2016, environ 120 enquêtes annuelles, infra-annuelles ou ponctuelles provenant de dix différents programmes seront intégrées dans ce nouveau cadre harmonisé.

Les objectifs du PISE sont de construire une seule plateforme harmonisée intégrant des enquêtes auprès des entreprises, réduire les coûts de développement et de maintenance, de simplifier les processus, réduire la courbe d'apprentissage du personnel et accroître la rapidité d'exécution des enquêtes, moderniser les processus, réduire le fardeau de réponse et réaliser des économies. Pour ce faire, la stratégie du PISE repose sur six piliers : une forte gouvernance, une utilisation accrue des données fiscales, une stratégie commune de validation, une gestion plus efficace de la collecte active, une collecte plurimodale avec la collecte électronique des données comme premier mode de collecte et l'utilisation du Registre des entreprises comme base de sondage unique pour toutes les enquêtes auprès des entreprises (Statistique Canada, 2010). La normalisation sera mise à profit afin de produire des statistiques d'une manière efficiente tout en répondant aux besoins précis des différentes enquêtes intégrées au PISE.

Le présent article décrit les composantes clés menant à la normalisation des méthodes et outils du PISE. À la section 2, les modèles fondamentaux de traitement seront présentés. La section 3 porte sur les moyens adoptés afin d'assurer une normalisation des méthodes et outils. La section 4 détaille les modèles de processus opérationnels et fait l'inventaire des méthodes et outils qui leur sont associés.

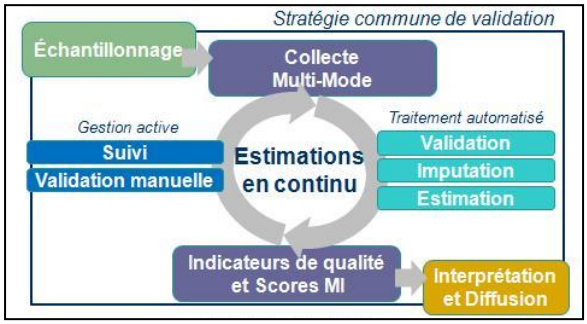
[1]Serge Godbout, Statistique Canada, 100, promenade pré Tunney's, Ottawa, (Ontario), K1A 0T6 (serge.godbout@statcan.gc.ca).

2. Modèles de traitement du PISE

Le modèle d'échantillonnage et de collecte du PISE repose sur un plan à deux phases afin de mieux cibler la population pour la production d'estimations financières, de marchandises et autres caractéristiques liées à l'industrie. Le modèle d'échantillonnage et de collecte suppose également une approche descendante centrée sur l'entreprise avec une utilisation maximale des données fiscales.

Le traitement post-collecte du PISE est fondé sur une stratégie commune de validation définie à l'aide d'un modèle de traitement itératif qui combine la collecte, le traitement et l'analyse des données, qui produit périodiquement les estimations et les indicateurs de qualité et qui fournit dynamiquement une rétroaction à la collecte et à l'analyse des données. La stratégie commune de validation a pour but d'harmoniser les méthodes et les outils de validation, d'étendre les opérations automatisées de validation, réduire les activités de suivi auprès des répondants et limiter les interventions manuelles aux unités influentes. Une composante clé de la stratégie commune de validation est une méthode de priorisation efficiente du suivi et de la validation manuelle. Ceci devrait permettre de récolter des gains en efficacité et en amélioration de la qualité en termes d'actualité et de précision (Saint-Pierre et Bricault, 2011).

Figure 1
Modèle de traitement du PISE



Au centre de cette stratégie se trouve un modèle de traitement itératif nommé « estimations en continu ». Telles que montrées à la figure 1, les données recueillies au moyen d'une collecte multimodale et provenant de différentes sources sont d'abord intégrées puis traitées de façon automatisée pour les règles de validation, ainsi que pour l'imputation et l'estimation. À la sortie, des indicateurs de qualité sont associés aux estimations. Si la qualité ciblée est atteinte, les estimations sont mises de côté pour être interprétées avant la diffusion et les ressources de la collecte sont réallouées. Sinon, ces indicateurs de qualité sont désagrégés au niveau des unités afin de produire des scores de mesure d'impact servant à prioriser les unités pour le suivi téléphonique et la validation manuelle. Lors de l'itération suivante, les données intégrées sont mises à jour à l'aide des données nouvellement recueillies et des corrections manuelles pour produire un nouvel ensemble d'estimations et d'indicateurs de qualité. Le cycle s'arrête lorsque toutes les cibles de qualité sont satisfaites ou lorsque la date prévue de la fin de la période de collecte est atteinte.

3. Normalisation des méthodes et outils

La normalisation des méthodes et des outils passe par la construction de modèles de processus opérationnels exhaustifs, l'utilisation de systèmes généralisés et la définition de métadonnées efficaces.

3.1 Rôle des modèles de processus opérationnels

Le point de départ de la construction des modèles de processus opérationnels est le modèle statistique général du processus opérationnel (MSGPO). Ce modèle général a été développé par le groupe de travail conjoint de l'Union européenne et de l'Organisation des Nations Unies, d'Eurostat et de l'Organisation de coopération et de développement économiques (METIS – métadonnées statistiques) dans le but de servir de base pour le développement d'une terminologie uniforme et de métadonnées pour des systèmes et des processus statistiques (Secrétariat de la Commission économique pour l'Europe des Nations Unies, 2009). Il s'applique à toute activité entreprise par les organisations de statistiques officielles qui résulte en la production de données. Le MSGPO est

composé de quatre niveaux décrivant de manière hiérarchique l'ensemble des processus et sous-processus opérationnels statistiques. Des processus globaux s'ajoutent dont deux (la gestion de la qualité et la gestion des métadonnées) sont étroitement liés au modèle.

À partir du MSGPO, les processus et les sous-processus sont par la suite décrits afin de constituer le niveau le plus détaillé du modèle opérationnel qui devient propre au PISE. L'exercice constitue une étape importante pour la normalisation des processus en établissant une liste exhaustive et ordonnée des activités statistiques à accomplir pour le PISE, en précisant les rôles de chaque partenaire et en définissant un langage commun. De plus, les services génériques concernés dans le projet sont clairement identifiés.

3.2 Rôle des services intégrés génériques

À Statistique Canada, des services intégrés génériques ont été mis en oeuvre et d'autres seront mis en valeur. Ces services génériques, y compris la méthodologie, le Registre des entreprises, le Centre de ressources en conception de questionnaires, le Centre de service de données, la collecte, la publication et le traitement post-collecte centralisé, ont pour but d'accroître l'efficacité opérationnelle et de normaliser leurs services respectifs. Le groupe de travail sur l'AOB a recommandé que l'utilisation de ces services soit obligatoire pour tous les programmes et projets.

3.3 Rôle des systèmes généralisés

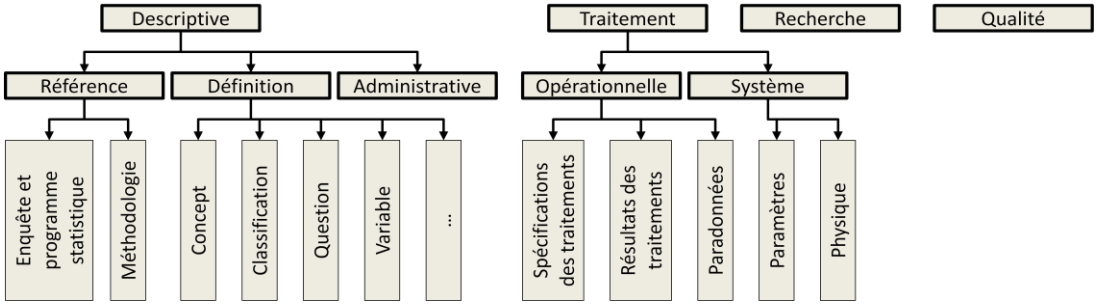
Depuis la fin des années 1980, Statistique Canada a développé un ensemble d'outils généralisés destinés à répondre aux besoins des enquêtes liés à l'échantillonnage, à la vérification, à l'imputation, à l'estimation, à la confidentialité et aux séries chronologiques. Avec le temps, ces outils ont démontré leur utilité dans la normalisation des méthodes, dans la réduction des coûts de développement et dans la facilitation de la mobilité des personnes entre les projets (Mohl, 2007). Les systèmes généralisés jouent un rôle très important dans la normalisation en fournissant un ensemble de méthodes statistiques connues, reconnues et éprouvées couvrant presque toutes les étapes de traitement. Ces méthodes sont ainsi mises en commun pour l'ensemble des enquêtes, réduisant le besoin de développement et de programmation. Construits de façon modulaire et permettant une paramétrisation flexible, ils s'adaptent aux besoins propres d'une enquête. Le développement, le soutien et la maintenance de ces systèmes sont assurés par des équipes de méthodologistes et de programmeurs, garantissant une mise à l'essai approfondie, une documentation complète et un soutien technique par des experts. Une fois en place, les systèmes généralisés réduisent grandement la programmation requise pour construire les systèmes spécifiques nécessaires à chaque enquête et en deviennent ainsi le pivot.

Plusieurs systèmes généralisés de Statistique Canada sont en cours de refonte. En particulier, de nouveaux systèmes généralisés d'échantillonnage (G-Sam) et d'estimation (G-Est) seront développés respectivement à partir du système généralisé d'échantillonnage (SGECH) et de la paire de systèmes généralisés d'estimation (SGE et StatMx), tous construits par Statistique Canada. Le système BANFF demeurera l'outil généralisé principal pour la vérification et l'imputation. D'autres outils généralisés tels que G-Confid et G-Coup ont également été développés par Statistique Canada (Deguire, Reedman et Wenzowski, 2011).

3.4 Rôle des métadonnées

Les métadonnées sont perçues de façon générale comme étant des données qui définissent et décrivent des données. Plus précisément, Gartner Research (Blechar et coll., 2010) définit les métadonnées comme de l'information qui décrit diverses facettes d'un élément d'information pour en améliorer son utilité à travers son cycle de vie. Les différents types de métadonnées incluses dans le PISE peuvent être organisés selon la classification donnée à la figure 2.

Figure 2
Classification des métadonnées du PISE



Les principes fondamentaux du PISE incluent le développement de systèmes gérés par des métadonnées, sur la base que les métadonnées devancent les diverses étapes du traitement et sont utilisées en entrée au lieu d'être documentées une fois la production effectuée. Dans la vision du PISE, l'entrepôt des métadonnées sera centralisé afin d'éliminer la duplication du travail, d'améliorer leur efficacité et réduire le risque d'erreurs dans l'ensemble des étapes d'une enquête (Statistique Canada, 2010). Les besoins en termes d'identification et de traçabilité des métadonnées ont été établis de façon prioritaire.

Dans la normalisation des méthodes et des systèmes, les métadonnées jouent divers rôles importants. Les métadonnées descriptives permettent d'harmoniser les définitions des données et des métadonnées et assurent une documentation en continu. Les métadonnées de traitement rehaussent l'efficacité, la flexibilité et la cohérence des systèmes en gérant par une paramétrisation efficace les complexités générées par les besoins spécifiques des enquêtes. Elles améliorent les processus en détaillant les opérations réalisées dans les différentes étapes du traitement et elles simplifient la connectivité avec les systèmes généralisés.

4. Description des modèles de processus opérationnels et inventaires des méthodes

Afin d'appuyer les différents processus de traitement du modèle du PISE, un grand nombre de méthodes statistiques doivent être évaluées et choisies dont plusieurs sont déjà disponibles dans l'éventail des différents systèmes généralisés existants de Statistique Canada. Dans certains cas, plus d'une méthode statistique peut être utilisée par les gestionnaires pour répondre aux besoins spécifiques des enquêtes. Mais dans d'autres occasions, les méthodes peuvent avoir des répercussions significatives sur la complexité du modèle, sur les opérations ou sur le développement des systèmes. Dans ces cas, les méthodes proposées doivent être précisément évaluées et comparées afin de faire un choix optimal d'un point de vue global.

Pour ce faire, six critères de sélection ont été établis, soient les impacts sur 1) les coûts de collecte, 2) sur la qualité des données, 3) sur la complexité, 4) sur les systèmes généralisés, 5) sur les opérations et 6) sur la capacité de réaction et la flexibilité. Dans le cas où plusieurs méthodes statistiques sont comparées, une pondération est assignée à chacun des six critères, permettant de relativiser un gain prévu sur la précision (par exemple meilleur coefficient de variation) par rapport à l'ensemble des effets sur les autres critères.

Dans les sections 4.1 à 4.3, les modèles d'échantillonnage et de collecte, de traitement post-collecte et d'indicateurs de qualité et de gestion active seront décrits. Leurs processus opérationnels ainsi que l'inventaire de leurs méthodes statistiques seront présentés.

4.1 Échantillonnage et collecte

Le modèle d'échantillonnage et de collecte du PISE repose sur un plan à deux phases. La première phase est un échantillon intégré de plus grande taille choisi à partir d'informations disponibles sur le Registre des entreprises. Elle sert à recueillir les informations de prise de contact, les variables de classification mises à jour et des informations de base sur les activités, les marchandises et autres caractéristiques liées à l'industrie. À noter que les unités dont l'information historique disponible est récente seront exclues de la collecte de première phase si les informations

sont jugées toujours valides. Dans le modèle du PISE, l'échantillonnage est centré sur l'entreprise selon une approche descendante, c'est-à-dire que chaque phase de collecte se fait à partir d'une unique base-liste dont l'unité d'échantillonnage est l'entreprise, ou son morceau qui est dans le champ de l'enquête, et sa contribution à chaque domaine d'intérêt est tenue en compte grâce à une stratégie multivariée. Dans un objectif de gestion du fardeau de réponse, le PISE vise à coordonner les échantillons intra-enquêtes et entre les enquêtes au moyen de la rotation.

Au niveau de la collecte, le modèle du PISE cherche une utilisation optimale des données fiscales dans le but de réduire le fardeau de réponse. Pour les données recueillies des unités enquêtées, la collecte électronique sera le mode principal. La stratégie de suivi pour la non-réponse inclut des actions par lot avec un faible coût marginal (rappels par télécopieur et courriel, des messages vocaux) et des suivis par interview téléphonique assistée par ordinateur dont le coût marginal est plus important.

Le tableau 3 décrit les méthodes et outils généralisés disponibles pour l'échantillonnage et la collecte pour les processus identifiés dans le modèle de processus opérationnel (figure 6 en annexe). À noter que les processus 4.1.2 à 4.1.5 ainsi que 4.1.7 à 4.1.10 sont absents du tableau, car ce sont essentiellement des étapes de gestion de métadonnées et d'approbation non liées à des méthodes statistiques. Les opérations et les outils liés à la gestion de la base de sondage et de collecte sont sous la responsabilité du service générique du Registre des entreprises (Statistique Canada, 2009). Pour ce qui est des composantes principales du plan d'échantillonnage que sont la stratification, la répartition et la sélection de l'échantillon, le système généralisé d'échantillonnage G-Sam (Statistique Canada, 2006) permettra l'utilisation de méthodes communément acceptées dans les enquêtes auprès des entreprises. Les opérations de collecte ne font pas explicitement partie du modèle de processus opérationnels du PISE puisqu'elles sont entièrement prises en charge par le service générique de collecte.

Tableau 3
Méthodes et outils généralisés pour l'échantillonnage et la collecte

Étapes du processus opérationnel	Méthodes statistiques disponibles et/ou recommandées	Outils généralisés
4.1.1- Créer et réviser la base de sondage	Paramétrisation	Registre des entreprises
4.1.6.1- Stratifier selon la taille	Racine cumulée de f (Dalenius-Hodges); Géométrique (Gunning-Horgan); Lavallée-Hidiroglou	G-Sam
4.1.6.2- Répartir l'échantillon	Répartition de puissance multivariée (y compris puissance univariée/Neyman)	G-Sam
4.1.6.3- Sélectionner l'échantillon	Bernoulli stratifié/Échantillonnage aléatoire simple (ÉAS) stratifié/Poisson; Nombres aléatoires permanents; Échantillonnage par collocation	G-Sam
4.1.6.4- Produire les rapports diagnostiques	Paramétrisation	Système PISE
4.1.11- Créer les entités de collecte	Paramétrisation	Registre des entreprises

4.2 Traitement post-collecte

Une fois les données recueillies et intégrées, la validation a pour but d'identifier les enregistrements et les variables qui doivent être imputées ou exclues du groupe des donneurs et/ou des modèles d'imputation. Selon la stratégie commune de validation du PISE, la majorité des règles automatisées sont déplacées de la collecte pour être intégrées dans le traitement post-collecte, afin de prioriser efficacement les unités ayant échoué des règles de validation et de réduire de façon importante le délai entre la réception des données et leur disponibilité pour le traitement.

En plus de la non-réponse partielle, la stratégie du PISE est d'imputer les unités pour la non-réponse totale lorsque des données auxiliaires sont disponibles. De plus, la stratégie de validation sélective proposée pour le PISE requiert d'imputer les variables clés pour les répondants afin de servir de valeurs prédites. La nature des données imputées variera selon la phase de collecte. À la première phase, l'imputation devrait se limiter aux variables de taille et de classification pour les unités non-répondantes ou pour les unités exclues de la collecte de première phase lorsque l'information est jugée toujours valide. À la deuxième phase, les données d'enquête et les données administratives seront sujettes à l'imputation.

À l'étape de l'estimation, les enquêtes du PISE visent à estimer des totaux, ratios, proportions et tendances pour des données financières, de marchandises et de caractéristiques de l'industrie recueillies à la deuxième phase. Les systèmes devront permettre la production des estimations par domaine et leurs variances associées pour les portions enquêtées et non enquêtées. De plus, les algorithmes d'échantillonnage de deuxième phase, en particulier ceux de stratification et de répartition, requièrent des estimations de totaux auxiliaires servant comme paramètres d'entrée.

Le tableau 4 décrit les méthodes et outils généralisés disponibles pour le traitement post-collecte pour les processus tels qu'identifiés dans le modèle de processus opérationnel (figure 6 en annexe). Les méthodes de validation et d'imputation seront appuyées par le système généralisé BANFF (Statistique Canada, 2011) alors que G-Est (Statistique Canada, 2005) sera le système généralisé appuyant l'estimation.

Tableau 4
Méthodes et outils généralisés pour le traitement post-collecte

Étapes du processus opérationnel	Méthodes statistiques disponibles et/ou recommandées	Outils généralisés
5.2.1 et 5.5.2- Détecter les données aberrantes	Hidiroglou-Berthelot; Écart Sigma	BANFF
5.2.2- Détecter des erreurs	Principes de Fellegi-Holt	BANFF
5.3.1- Imputer par item, (complet et détails)	Plus de 20 méthodes d'imputation disponibles, Y compris: imputation par donneur, valeur précédente, tendance, etc.	BANFF
5.3.2- Imputer les répondants	Méthodes similaires d'imputation	BANFF
5.3.3- Ajuster proportionnellement	Ajustement proportionnel de base ou pro-rating	BANFF
5.5.1- Repondérer pour la non-réponse	Ajustements par groupes de réponse homogène	G-Est
5.6.1- Caler les poids	Hajek; Régression généralisée avec double calage	G-Est
5.6.2- Calculer les estimations par domaine et les variances	Méthodes à être déterminées	G-Est

4.3 Indicateurs de qualité et gestion active

De façon générale, les indicateurs de qualité servent à évaluer la qualité du produit. Dans le PISE, ils jouent également un rôle important dans la redistribution des ressources liées aux opérations de suivi et de validation et pour l'arrêt de la collecte active. Le modèle du PISE prévoit l'utilisation d'indicateurs de qualité fondés sur la variance totale combinant la variance échantillonnale et la variance due à la non-réponse ou sur l'erreur quadratique moyenne dans le cas où le biais est mesuré. D'autres indicateurs de qualité fondés sur des taux de réponse ou de couverture, pondérés ou non, viennent les compléter afin de prévenir des erreurs en raison de la volatilité des estimations de mesure de la qualité en début de collecte. Les indicateurs de qualité sont par la suite désagrégés au niveau des unités, formant les scores de mesure d'impact (MI) dans le but de prédire leur impact respectif sur la qualité. Les scores MI servent aussi de mesure de taille pour la sélection des échantillons de suivi et pour leur priorisation pour la vérification manuelle. Puisque chaque unité a un score MI pour chaque variable clé et indicateur de qualité pris en compte, un score MI global (Hedlin, 2008) doit être dérivé afin de les combiner en une mesure de taille unique.

Dans le modèle du PISE, la gestion active est composée de deux branches : la gestion active de la collecte des données et la gestion active de l'analyse. La gestion active de la collecte tente de rentabiliser les ressources pour le suivi de la non-réponse et le suivi de règles échouées de validation (Godbout, Beaucage et Turmelle, 2011). La gestion active de l'analyse vise à prioriser les unités pour la révision manuelle dans un processus de validation sélective (Brundell, 2011). Le modèle du PISE propose de se fonder sur les scores MI pour sélectionner un sous-échantillon aléatoire d'unités influentes pour suivi de la non-réponse et pour prioriser les unités et les domaines pour suivi des règles de validation échouées et de révision manuelle.

Le tableau 5 décrit les méthodes et outils généralisés disponibles pour les indicateurs de qualité et la gestion active pour les processus identifiés dans le modèle de processus opérationnel (figure 6 en annexe). La majorité des

méthodes liées aux indicateurs de qualité et à la gestion active seront spécifiques au système du PISE sauf le sous-échantillonnage pour le suivi de la non-réponse qui reprendra certaines méthodes mises en oeuvre dans le système généralisé G-Sam pour l'échantillonnage initial.

Tableau 5
Méthodes et outils généralisés pour les indicateurs de qualité et la gestion active

Étapes du processus opérationnel	Méthodes statistiques disponibles et/ou recommandées	Outils généralisés
6.1.1.1- Calculer les indicateurs de qualité	Synthèse de mesures de la qualité calculées aux étapes précédentes	Système PISE
6.1.1.2- Calculer les scores MI	Désagrégation des indicateurs de qualité par linéarisation	Système PISE
6.1.2.1- Comparer les indicateurs de qualité et les cibles	Calcul de variables dérivées	Système PISE
6.1.2.2- Dériver les scores MI globaux	Calcul de variables dérivées	Système PISE
6.1.4.1- Stratifier et répartir le sous-échantillon de suivi	Répartition de puissance multivariée	G-Sam
6.1.4.2- Sélectionner le sous-échantillon de suivi	Bernoulli stratifié/ÉAS stratifié/Poisson; Nombres aléatoires permanents	G-Sam
6.1.3- Générer les rapports d'analyse	Agrégation de tableaux	Système PISE

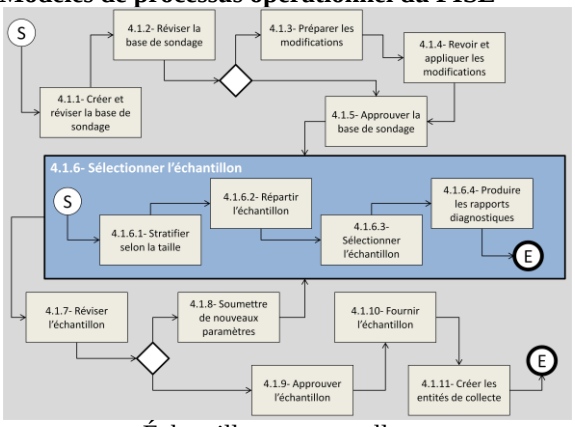
Conclusion

La normalisation est un élément critique du PISE étant donné la complexité du traitement, les spécificités provenant de chacune des enquêtes intégrées et la flexibilité requise pour les requêtes ponctuelles.

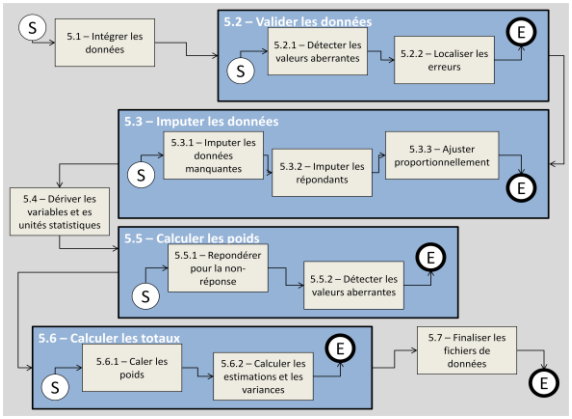
Les modèles de processus opérationnels établissent une liste complète des activités statistiques à accomplir. Les rôles de chaque partenaire sont clairement précisés et un langage commun est également défini. Les systèmes généralisés et les services intégrés génériques deviennent les points de convergence des processus. Ils assurent un développement, un soutien et une maintenance de haute qualité pour des composantes souvent jugées plus complexes ou spécialisées. Les métadonnées permettent de gérer efficacement la complexité et les spécificités. Elles harmonisent les définitions, assurent une documentation en continu et rehaussent l'efficacité, la flexibilité et la cohérence des données. De plus, elles font le pont entre les systèmes spécifiques du PISE et les systèmes généralisés. Les méthodes statistiques proposées aux différentes enquêtes sont soigneusement évaluées et choisies en mettant l'accent sur plusieurs dimensions et non pas seulement sur la précision des estimations.

Annexe – Modèles de processus opérationnels

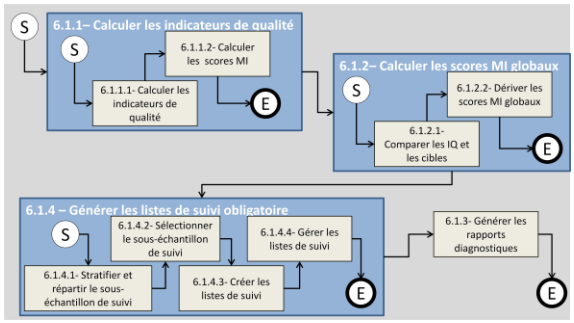
Figure 6
Modèles de processus opérationnel du PISE



Échantillonnage et collecte



Traitement post-collecte



Indicateurs de qualité et gestion active

Bibliographie

Blechar, M., Beyer, M.A., Thompson, J., Lapkin, A. et N. Gall (2010), « Gartner Clarifies the Definition of Metadata », 2H10-1H11, Gartner Research, 19 août 2010.

Brundell, P. (2011), « La vérification sélective des données et sa mise en œuvre à Statistics Sweden », *Recueil du Symposium international de 2011 sur les questions de méthodologie de Statistique Canada*, Ottawa (Canada).

Deguire, Y., Reedman, L. et M. Wenzowski (2011), « Systèmes généralisés : l'expérience de Statistique Canada », *Recueil du Symposium international de 2011 sur les questions de méthodologie de Statistique Canada*, Ottawa (Canada).

Godbout, S., Beaucage, Y. et C. Turmelle (2011), « Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program », Conférence européenne des statisticiens, Ljubljana (Slovénie).

Hedlin, D. (2008), « Local and Global Score Functions in Selective Editing », Conférence européenne des statisticiens, Session de travail sur la validation des données statistiques, Vienne (Autriche).

Mohl, C. (2007), « The Continuing Evolution of Generalized Systems at Statistics Canada for Business Survey Processing », International Conference on Establishment Surveys III, Montréal (Canada).

Saint-Pierre, É. et M. Bricault (2011), « The Common Editing Strategy and the Data Processing of Business Statistics Surveys », Conférence européenne des statisticiens, Ljubljana (Slovénie).

Secrétariat de la Commission économique pour l'Europe des Nations Unies. (2009), « Generic Statistical Business Process Model, version 4.0 », Session de travail conjoint UNECE/Eurostat/OECD sur les métadonnées statistiques (METIS).

Statistique Canada (2005), GES V4.3 – Guide de l'utilisateur, Ottawa (Canada).

Statistique Canada (2006), Système généralisé d'échantillonnage, version 2.3, Guide de l'utilisateur, Division des méthodes des enquêtes auprès des entreprises, Ottawa (Canada).

Statistique Canada (2009), Guide sommaire sur le Registre des entreprises, Division du registre des entreprises, Ottawa (Canada).

Statistique Canada (2010), « Integrated Business Statistics Program Blueprint », Division de la statistique des entreprises, Ottawa (Canada), document interne.

Statistique Canada (2011), « BANFF - description des fonctions du système BANFF pour la vérification et l'imputation, version 2.04 », Équipe de soutien de BANFF, Ottawa (Canada).

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Normalisation des processus

Frank Hofman, Astrea Camstra et Robbert Renssen[1,2,3]

Résumé

Au cours des cinq dernières années, Statistics Netherlands a entrepris un programme ambitieux, en vue de remanier le processus statistique. Les concepts généraux qui sous-tendent ce programme sont représentés par une architecture intégrée globale. Plus récemment, l'architecture a été complétée par une série de méthodes et d'outils standard qui devraient faciliter la conception du processus de production. Avant que les méthodes ou les outils standard puissent être appliqués sans difficulté à la production de statistiques, les processus statistiques proprement dits doivent être normalisés dans une certaine mesure. À cette fin, un modèle opérationnel conceptuel pour le traitement des données statistiques a été élaboré. Un concept important de ce modèle est l'étape de processus normalisée. Les étapes de processus normalisées correspondent aux applications des fonctions statistiques qui peuvent être mises en œuvre comme services opérationnels. Les fonctions statistiques sont habituellement fondées sur des méthodes statistiques standard. En déterminant ces étapes normalisées et en fournissant des lignes directrices concernant leur utilisation, le modèle vise à combler le fossé entre la vision de haut niveau adoptée dans l'architecture opérationnelle et la conception des processus statistiques en pratique. Le présent article aborde brièvement le modèle et ses concepts. Le modèle sera illustré en l'appliquant au domaine de la vérification des données.

Mots clés : Étape de processus normalisée ; composante fonctionnelle ; élément, fonction.

1. Introduction

Les instituts statistiques sont constamment soumis à des pressions en vue d'améliorer l'efficacité et de réduire le fardeau de réponse, particulièrement pour les entreprises. Par ailleurs, on leur demande de maintenir des normes de qualité élevées, d'améliorer la souplesse et de mettre davantage l'accent sur les besoins des utilisateurs qui évoluent rapidement et sur l'innovation des produits. Pour répondre à ces objectifs contradictoires, Statistics Netherlands a entrepris un programme de remaniement ambitieux (voir Braaksma 2009, pour un aperçu). Les concepts généraux qui sous-tendent ce programme sont intégrés dans une architecture opérationnelle exhaustive (par exemple, Huigen et coll. 2009).

L'un de ces concepts est de favoriser la réutilisation. La réutilisation des données permet la production plus efficace de statistiques, tandis que la réutilisation des méthodes, processus et outils contribue à une conception ou à un remaniement plus efficace des statistiques. Récemment, l'architecture a été complétée par une série de méthodes standard et un ensemble d'outils standard qui peuvent être utilisés pour rationaliser davantage le processus de production de base. La série de méthodes est un catalogue de méthodes statistiques approuvées, qui sont actuellement utilisées par Statistics Netherlands, et vise à informer et à aider les statisticiens à appliquer les méthodes appropriées. En dernier ressort, tous les processus statistiques devraient utiliser ou réutiliser uniquement les méthodes comprises dans la série. La nécessité d'outils standard découle des coûts constamment élevés d'entretien des technologies de l'information (TI) résultant de la grande diversité d'outils (souvent développés sur mesure). Une étude évaluant les outils actuels de la production de statistiques (Renssen, Wings et Paulussen, 2008) a donné lieu à une liste préliminaire de 18 outils privilégiés pour le domaine du traitement des données statistiques, les deux critères les plus importants d'inclusion dans la liste étant la possibilité pour l'outil de traiter des métadonnées, et sa capacité de faire une distinction entre la conception et la mise en œuvre.

[1]Frank Hofman, Astrea Camstra et Robbert Renssen, Statistics Netherlands, Henri Faasdreef 312, 2492 JP La Haye, Pays-Bas. Courriel : f.hofman@cbs.nl, a.camstra@cbs.nl, rh.renssen@cbs.nl.
[2]Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands.
[3]Lien avec la série des méthodes (en anglais) : www.cbs.nl/en-GB/menu/methoden/gevalideerde-methoden/default.htm.

Avant que les méthodes ou les outils standard puissent être appliqués à la production de statistiques, les processus statistiques doivent aussi être normalisés (en partie). Renssen et coll. (2009) présentent certains concepts initiaux pour la normalisation des processus statistiques. Dans Renssen (2010), ces concepts sont élargis pour former un modèle opérationnel conceptuel général pour le traitement des données. Un concept important de ce modèle est ce qu'on appelle l'étape de processus normalisée. Les étapes de processus normalisées correspondent aux applications de fonctions statistiques mises en œuvre au moyen d'une méthode statistique standard. Plusieurs méthodes différentes peuvent remplir les mêmes fonctions, par exemple, les méthodes hot deck et du plus proche voisin sont deux méthodes utilisées pour l'imputation. Par ailleurs, une méthode particulière peut être appliquée à différentes fonctions. Une méthode de régression, par exemple, peut être utilisée pour imputer les valeurs de données ou pour estimer des totaux de population. L'idée est de concevoir des processus en termes d'étapes de processus normalisées, qui peuvent servir de lien entre les méthodes et les outils. La création d'un répertoire d'étapes de processus normalisées fera augmenter la transparence et la souplesse du processus de conception, et facilitera la réutilisation de méthodes et de processus (et en dernier ressort d'outils).

La section 2 décrit les modèles théoriques qui ont été élaborées à cette étape de la recherche. Ces modèles ont été appliqués au domaine de la vérification des données, laquelle est brièvement expliquée à la section 3. Enfin, la section 4 présente certaines conclusions préliminaires et un aperçu des recherches futures.

Le présent article est fondé en partie sur deux documents antérieurs portant sur le même sujet (Camstra et Renssen, 2011 et Renssen et Camstra, 2011b).

2. Modèle conceptuel pour le traitement des données

Le concept d'étape de processus normalisée n'est pas indépendant et ne devient utile que dans le contexte de notre modèle de conception de processus statistiques. La section 2.1 présente brièvement le modèle en décrivant la première étape de la conception statistique : l'opérationnalisation des concepts. La deuxième étape consiste à déterminer les fonctions statistiques nécessaires (section 2.2) et, enfin, l'application de ces fonctions à des processus statistiques (section 2.3). La section 2.4 explique comment nous voulons normaliser la conception du processus en fournissant des éléments génériques. Enfin, un aperçu de tous les concepts et des rapports qui existent entre eux est fourni à la section 2.5.

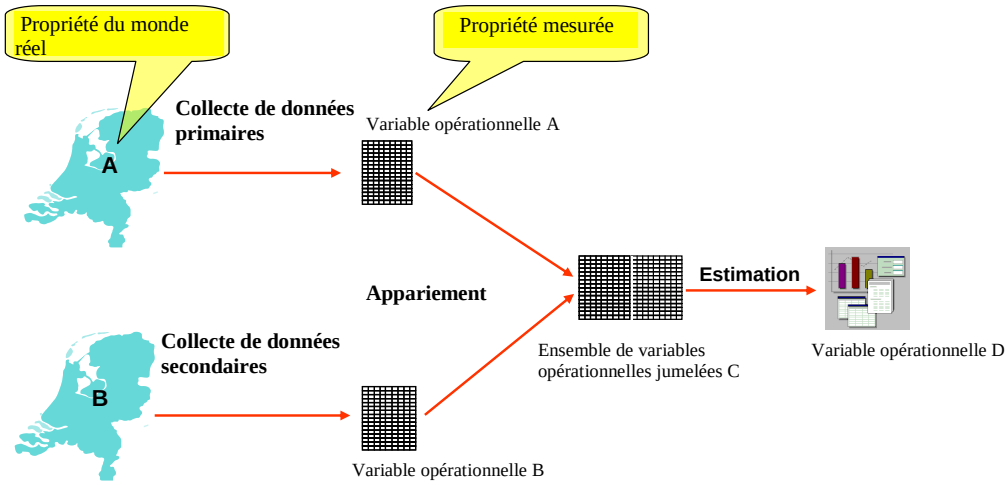
2.1 Opérationnalisation des concepts

La conception d'un processus statistique commence par la détermination des besoins de données statistiques et la transposition de ces besoins en produits. Cela nécessite la définition des propriétés du monde réel, y compris la population et la période visée de déclaration des renseignements. Par exemple, les chiffres sur le chômage ont trait à la population des résidents hollandais de 15 à 65 ans, le « chômage » étant défini comme le « fait de travailler moins de 12 heures par semaine et de chercher activement du travail (ou plus de travail) » et la période de référence correspondant au mois civil.

La propriété « chômage » est appelée variable d'attribut conceptuelle. Pour mesurer une variable d'attribut conceptuelle, celle-ci doit d'abord être transposée en variable d'attribut opérationnelle. Une définition opérationnelle décrit les mesures requises pour mesurer le concept, y compris la classification et les échelles de mesure. Par exemple, on peut opérationnaliser le chômage au moyen d'une ou de plusieurs questions dans une enquête.

Au moment de la production de statistiques, l'opérationnalisation d'un concept statistique est généralement un processus à plusieurs étapes. Dans la figure 2.1-1, on présente un exemple très simple. Pour publier des chiffres sur le nombre total de chômeurs selon le groupe d'âge (D), la propriété « chômage » (A) est mesurée dans un échantillon au moyen de l'Enquête sur la population active, tandis que les caractéristiques des personnes, comme l'âge (B), sont obtenues à partir du registre de la population. Ces variables opérationnelles mesurées sont combinées en une nouvelle mesure C, à partir de laquelle D est finalement estimée. En raison des erreurs de collecte de données, des erreurs d'échantillonnage et/ou des erreurs d'appariement, l'estimation qui en résulte peut différer du total conceptuel qui a dû être estimé.

Figure 2.1-1
Opérationnalisation d'une variable en plusieurs étapes



Il arrive fréquemment que plusieurs opérationnalisations (ou ensembles d'opérationnalisations) soient nécessaires pour obtenir le même produit statistique (D), même si les stratégies statistiques peuvent limiter les choix possibles. Ces stratégies découlent habituellement des politiques de gestion efficace des processus de production statistique et favorisent souvent une solution particulière. Selon la stratégie générale pour la collecte des données à Statistics Netherlands, tous les efforts doivent être faits pour réduire la déclaration statistique, ce qui signifie que la collecte de données primaires doit être effectuée uniquement lorsqu'il n'y a pas d'autre source de données (secondaires). Ainsi, dans la figure 2.1-1, on montre deux sources de données, plutôt que la collecte de toutes les données au moyen de l'enquête.

Fonctions statistiques

Les séries d'opérationnalisations de la figure 2.1-1 montrent déjà les grandes lignes de la conception du processus statistique. L'étape suivante de la conception consiste à élaborer ces étapes davantage. Cela signifie de déterminer l'ensemble de fonctions nécessaires à chaque étape, afin d'obtenir le produit final à partir d'un ou de plusieurs produits d'entrée. Dans la figure 2.1-1, la deuxième étape utilise une fonction d'appariement pour obtenir C à partir de A et B. À la troisième étape, on utilise une fonction d'estimation. L'appariement et l'estimation sont des exemples de fonctions statistiques. Ces fonctions statistiques sont quant à elles liées à des méthodes statistiques. La fonction d'estimation, par exemple, est souvent fondée sur une méthode de régression.

L'exécution d'une étape particulière peut nécessiter un certain nombre de mesures préparatoires. Pour apparier deux ensembles de données, il peut être nécessaire de calculer et/ou d'encoder un ensemble de variables pour obtenir une variable clé unique. On procédera à la modélisation en appliquant de façon successive une fonction de dérivation, une fonction de codage et une fonction d'appariement. Pour obtenir un résultat d'appariement réussi, plusieurs itérations de la fonction d'appariement seront peut-être nécessaires, au moyen de variables clés différentes à chaque itération.

Les mesures des propriétés du monde réel peuvent aussi comprendre des erreurs ou des données manquantes. Outre la mesure des propriétés proprement dites, il est par conséquent important de fournir des renseignements concernant la qualité de ces mesures. Par ailleurs, nous considérons les fonctions de qualité comme des fonctions statistiques particulières qui mesurent la qualité selon une méthode statistique ou des connaissances spécialisées.

Dans la section 3, nous examinerons plus étroitement les fonctions statistiques, plus particulièrement celles liées au contrôle des données.

2.3 Conception d'un processus statistique

Après avoir déterminé les fonctions requises, les données d'entrée, le produit et la méthode de chaque application des fonctions (les étapes du processus) doivent être précisés avec exactitude. De tout temps, nous avons décrit les processus statistiques en termes d'activités (étapes), qui étaient parfois regroupées en sous-processus. Les données d'entrée et les produits étaient principalement décrits au niveau des sous-processus et, à l'occasion seulement, les activités se reportaient à la série de méthodes pour leur fonctionnement interne. Même si cette façon de décrire les processus donne un aperçu d'un processus et sert de point de départ pour le développement d'outils des TI, il semblait assez difficile de comparer différentes conceptions de processus ou de déterminer les parties réutilisables possibles. Une étude de plusieurs processus statistiques remaniés récemment a fait ressortir plusieurs causes :

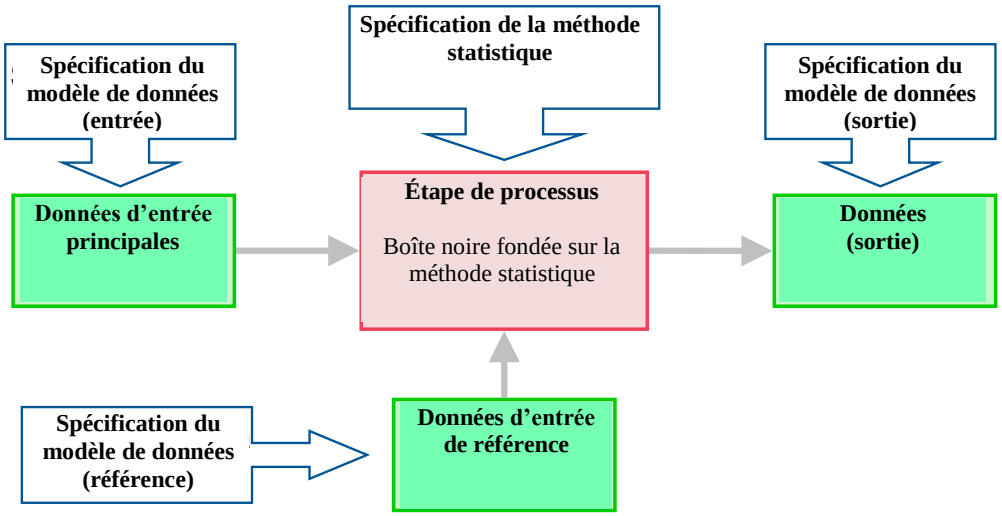
- Des étapes similaires à l'intérieur de plusieurs processus sont désignées et décrites de façon assez différente, ce qui rend difficile de déterminer leur similitude.
- Les applications de méthodes statistiques ne sont pas reconnues, particulièrement lorsqu'elles sont simples. Par exemple, lorsqu'on estime des totaux de population à partir des données d'un registre, ces données sont implicitement pondérées au moyen de poids qui sont égaux à 1.
- Les applications des méthodes statistiques peuvent être très complexes et peuvent nécessiter des activités préparatoires. Par exemple, lorsque l'on utilise les données *t*-1 dans un processus de vérification, celles-ci doivent être appariées auparavant. En outre, il existe des procédures pour le cas où les données t-1 ne seraient pas complètes.
- Les applications d'outils particuliers peuvent nécessiter des activités préparatoires (non statistiques), comme la transformation du format, le remaniement des colonnes de données ou une nouvelle désignation des variables.
- La réutilisation des données et les stratégies de mode mixte compliquent les activités d'un processus, parce que plusieurs sources de données doivent être combinées et que les processus doivent être reliés les uns aux autres.

Grâce à la normalisation des étapes du processus, nous tentons d'accélérer la conception et, en dernier ressort, le processus de mise en œuvre. Pour atteindre cet objectif, nous établirons un répertoire d'éléments génériques. En combinant et en configurant ces éléments, on peut facilement concevoir un processus particulier.

La figure 2.3-2 montre comment un tel élément est mis en œuvre comme étape de processus normalisée à l'intérieur d'un processus statistique particulier. Au cours de la production, on aura besoin des données principales et peut-être de certaines données de référence comme données d'entrée pour produire les données de sortie requises (qui sont montrées en vert). Pour concevoir cette étape de processus normalisée, nous devons sélectionner le bon élément et préciser les éléments en bleu : les modèles de données d'entrée et de sortie ainsi que la méthode utilisée. Il s'agit par exemple des règles de contrôle pour une fonction de contrôle ou des variables et des critères d'appariement pour une fonction d'appariement.

Nous avons distingué de façon explicite les principales données d'entrée et les données d'entrée de référence (ou auxiliaires). Les principales données d'entrée sont les données qui sont traitées dans les faits ou auxquelles de la valeur est ajoutée à cette étape, tandis que les données d'entrée de référence ne sont modifiées d'aucune façon. Au moment de l'imputation, par exemple, l'unité (enregistrement) qui sera imputée représente les principales données d'entrée, tandis que les données t-1 sont utilisées uniquement comme données d'entrée de référence. Si les données d'entrée de référence ne sont pas facilement disponibles, nous avons recours à des sous-processus supplémentaires pour recueillir et préparer ces données de référence. Nous avons noté des sous-processus supplémentaires qui sont assez complexes, ce qui rend l'ensemble du processus moins transparent et la conception, la mise en œuvre et/ou la production plus coûteuses. Il serait possible d'utiliser une méthode différente, nécessitant des données de référence facilement disponibles, mais donnant lieu à un produit similaire. Grâce à une distinction claire entre les données principales et le déroulement du processus, d'une part, et les données de référence et les sous-processus auxiliaires, d'autre part, nous obtenons un meilleur aperçu de l'ensemble du processus. Cela permet aussi d'effectuer une meilleure évaluation de la qualité du produit par rapport aux coûts.

Figure 2.3-2
Spécification d'une étape de processus normalisée



La distinction entre les données d'entrée principales et de référence augmente la souplesse des processus. Nous voyons que les principales données d'entrée et de sortie peuvent être identiques pour une fonction, peu importe la méthode utilisée. Lorsque nous remplaçons une méthode par une autre (pour la même fonction), nous devons uniquement modifier les données d'entrée de référence et la spécification de la méthode. Pour une imputation par la moyenne dans un groupe simple, par exemple, nous avons besoin uniquement d'une ou de plusieurs variables auxiliaires pour établir le groupe et toutes les unités valides (sans données manquantes) à l'intérieur de ce groupe. Contrairement aux imputations par le ratio et aux imputations par la régression, celles-ci nécessitent d'autres données de référence (et souvent davantage). C'est donc dire que nous n'avons pas à modifier le flux de données principales, mais uniquement cette étape du processus et peut-être des sous-processus auxiliaires pour les données d'entrée de référence.

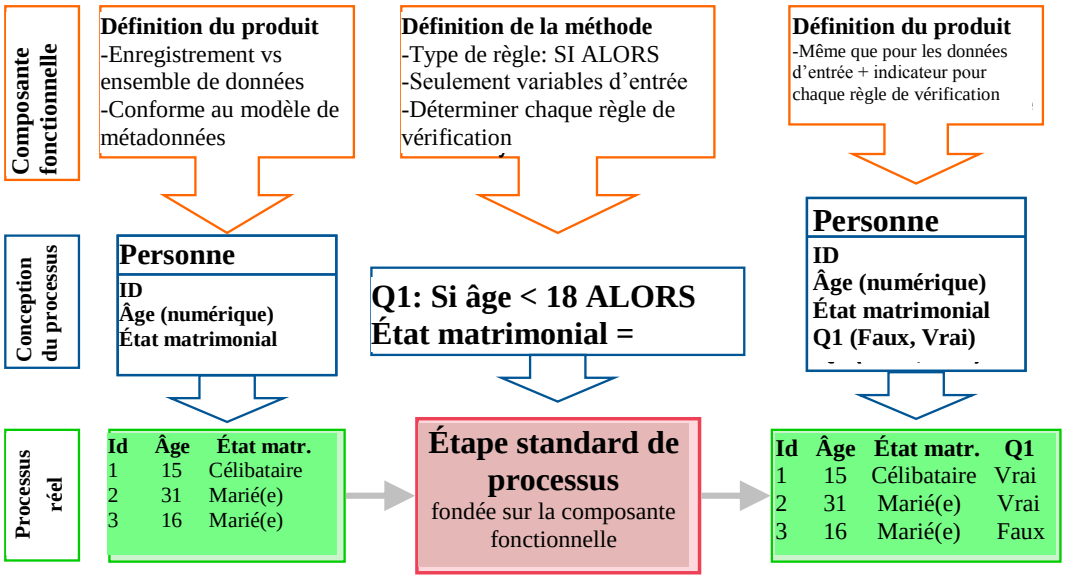
L'idée de concevoir un processus statistique au moyen d'un répertoire d'éléments génériques a été mise à l'essai dans deux cas. Le premier cas a pris la forme d'un genre de répétition générale, dans laquelle nous avons réorganisé, sur papier seulement, le processus existant des statistiques sur les entreprises à court terme remanié récemment (voir Renssen et Slootbeek, 2011). Le deuxième cas était un remaniement réel du processus statistique pour certaines statistiques environnementales, qui est actuellement mis en œuvre. Le premier cas a montré qu'il est possible d'appliquer le concept des étapes de processus normalisées. Le deuxième cas a confirmé les objectifs de plus grande transparence et les possibilités de réutilisation, même à l'intérieur d'un processus.

2.4 Composante fonctionnelle

Dans la section précédente, nous avons illustré comment nous souhaitons utiliser les étapes de processus normalisées pour concevoir des processus statistiques. Nous souhaitons établir un répertoire d'éléments qui peuvent être mis en œuvre comme étapes de processus normalisées dans un processus particulier. Ces éléments sont génériques, tandis que les étapes de processus normalisées qui sont fondées sur eux sont configurées en vue d'être utilisées dans un processus statistique particulier. Le nom officiel d'un élément est « composante fonctionnelle ». Une composante fonctionnelle décrit sa fonction (valeur ajoutée) et la méthode utilisée, ainsi que les exigences et les restrictions de ses spécifications.

La figure 2.4-3 illustre comment les composantes fonctionnelles peuvent être utilisées dans la conception d'un processus. Au bas, en vert, nous voyons le processus de production proprement dit. Dans ce cas, nous avons un ensemble de données simple constitué de personnes, y compris leur âge et leur état matrimonial. Nous souhaitons déterminer si les données de cet ensemble sont conformes à la règle de contrôle selon laquelle les « personnes de moins de 18 ans ne peuvent être mariées ». Après le traitement, nous nous attendons à ce que les deux premiers enregistrements soient correct, et le troisième, incorrect.

Figure 2.4-3
D'une composante fonctionnelle à une étape de processus normalisée



Dans la couche intermédiaire en bleu, nous concevons ce processus simple en récupérant une composante fonctionnelle pour la validation des données et en spécifiant les modèles de données d'entrée et de sortie, ainsi que la règle de contrôle.

La couche du haut en orange montre les exigences et les restrictions qui accompagnent la composante fonctionnelle et qui doivent être respectées pendant la conception. Par exemple, cette composante de validation des données peut traiter des unités individuelles (enregistrements), tandis qu'une composante d'agrégation nécessitera un ensemble complet d'unités. Il existe un autre type de restriction, à savoir le type de données des variables à imputer (numériques ou catégoriques). Pour la spécification de la méthode dans l'exemple, les règles se limitent au type « SI ALORS ». D'autres recherches sont nécessaires pour élaborer davantage ces exigences et restrictions.

2.5 Modèle d'information

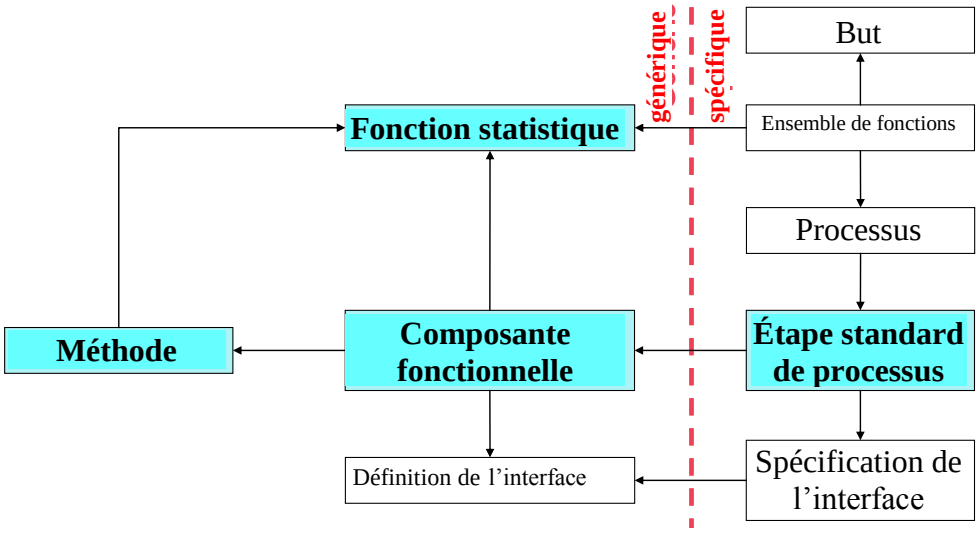
Le modèle d'information de la figure 2.5-4 fournit un aperçu des principaux concepts et de leurs relations présentés jusqu'à maintenant. Dans cette section, nous résumons brièvement les concepts principaux. Pour plus de renseignements, voir Gelsema et Hofman (2011).

Le côté gauche de la figure montre tous les concepts génériques, tandis que les concepts du côté droit sont propres à une enquête, par exemple, l'Enquête sur la population active. La composante fonctionnelle est le concept central. Il s'agit de l'élément générique qui peut être utilisé pour concevoir un processus particulier. Chaque composante fonctionnelle met en œuvre une fonction statistique grâce à l'application d'une méthode (statistique). Parmi les exemples de fonctions statistiques figurent l'imputation, la dérivation et la validation. Par ailleurs, pour une fonction particulière, nous pouvons avoir plusieurs composantes fonctionnelles correspondant aux différentes méthodes qui peuvent être utilisées, comme le plus proche voisin ou la méthode de régression pour la fonction d'imputation.

Du côté droit de la figure, nous voyons les concepts utilisés au moment de la conception d'un processus statistique particulier. Au niveau le plus élevé, chaque processus comporte un ou plusieurs objectifs, la sortie à produire (ou les besoins d'information à combler). En connaissant les entrées et les sorties d'un processus statistique, nous pouvons tracer une esquisse des fonctions statistiques dont nous avons besoin. Ce n'est qu'alors que nous entreprenons la conception proprement dite du processus (statistique), le choix des composantes fonctionnelles (et des méthodes que nous utiliserons), ainsi que la configuration de ces composantes pour leur utilisation particulière, c'est-à-dire leur transformation en étapes normalisées de processus.

La spécification d'interface décrit la configuration de chaque étape normalisée de processus, tout comme les modèles de données d'entrée et de sortie et la spécification (plus poussée) de la méthode. La définition d'interface générique de chaque composante fonctionnelle décrit les exigences et les restrictions pour la spécification.

Figure 2.5-4
Modèle d'information



3. Modèle appliqué à la vérification des données

Dans la présente section, nous appliquerons les principaux concepts du modèle décrits dans la section précédente au domaine de traitement de la vérification des données. La section 3.1 aborde la stratégie de vérification des données utilisée à Statistics Netherlands, tandis que la section 3.2 donne un exemple d'un processus très simple de vérification des données.

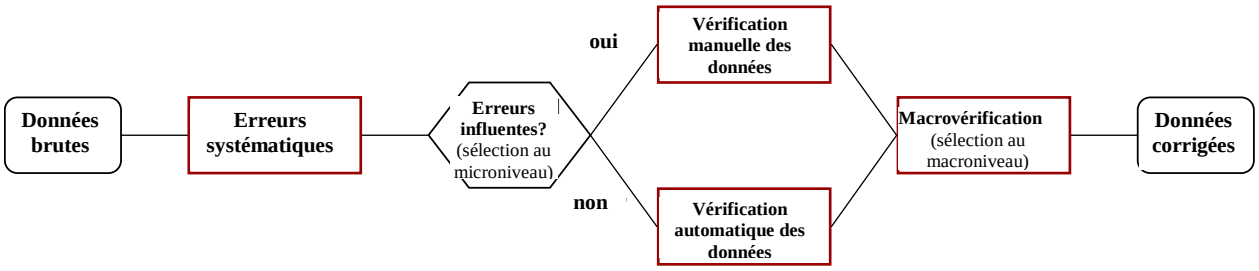
3.1 Stratégie de vérification des données statistiques

Le sujet « Validation des données et correction » de la série des méthodes (Hoogland et coll., 2010) porte sur les techniques de vérification des données les plus fréquemment utilisées à Statistics Netherlands. Les auteurs décrivent une stratégie globale pour le processus de validation et de correction des données, comme le montre la figure 3.1-5. La façon particulière dont cette stratégie est appliquée aux différents processus statistiques peut varier et ce ne sont pas toutes les étapes qui doivent être exécutées.

À la première étape du processus de validation et de correction des données, les erreurs systématiques (évidentes) sont détectées et corrigées. Parmi les exemples d'erreur systématique figure une « erreur de millier », c'est-à-dire une valeur qui est erronée selon un facteur de 1 000. Si les erreurs systématiques sont connues, elles peuvent facilement être corrigées au moyen de méthodes déductives. L'une des étapes suivantes consiste à vérifier et à corriger manuellement les données. Compte tenu des instructions générales de vérification des données, le spécialiste détermine les données qui sont erronées et comment elles devraient être corrigées. Les corrections sont habituellement apportées sur la base des connaissances d'expert, que viennent parfois compléter des données de référence. Étant donné que la vérification et la correction manuelles des données sont coûteuses et longues, elles se limitent souvent aux erreurs influentes qui ne peuvent être corrigées automatiquement de façon fiable. Pour déterminer les erreurs qui devraient être traitées manuellement et celles qui devraient l'être de façon automatique, un score est attribué à chaque unité indiquant les répercussions attendues sur les chiffres publiés si cette unité doit être corrigée manuellement (on appelle aussi cela le contrôle sélectif). Les scores élevés comportent une priorité élevée d'examen interactif. Les erreurs qui restent sont moins importantes et peuvent par la suite être vérifiées automatiquement. Un ensemble de valeurs de données pour une unité sont vérifiées par rapport à un ensemble de règles de contrôle déterminées au préalable, et la ou les valeurs erronées sont automatiquement localisées. À

Statistics Netherlands, les méthodes de localisation des erreurs sont fréquemment fondées sur le paradigme de Felligi-Holt. À la dernière étape, les chiffres provisoires destinés à la publication sont estimés et comparés avec les données historiques ou des sources de données externes. Si les chiffres estimés ne sont pas plausibles, les microdonnées sous-jacentes sont analysées à nouveau et corrigées au besoin. Le processus de macro-détection et de micro-correction est appelé macro-vérification. La macro-vérification peut être interprétée comme une forme de vérification sélective, grâce à laquelle la sélection des erreurs influentes est effectuée par l'entremise des estimations de population.

Figure 3.1-5
Stratégie de vérification des données



3.2 Fonctions statistiques comprises dans la vérification des données

Dans les différentes techniques comprises dans la stratégie globale de vérification décrite dans la section précédente, un certain nombre d'opérations ou de fonctions de vérification des données de base peuvent être distinguées. Nous définissons les quatre fonctions de vérification suivantes :

- Fonction de validation des données : vérification des variables en vue de déterminer les erreurs et les incohérences.
- Fonction de localisation des erreurs : détermination, dans le cas des incohérences, de la variable qui est « erronée » et qui doit être corrigée.
- Fonction de score : détermination des observations influentes, c'est-à-dire les observations qui ont des répercussions substantielles sur les chiffres publiés.
- Fonction de correction : correction des erreurs ou des incohérences (localisées).

Les quatre fonctions sont suffisantes pour concevoir un processus qui met en œuvre la stratégie de vérification des données dont il est question à la section précédente, même si une étape de la stratégie nécessite généralement plusieurs fonctions. Par exemple, le traitement des erreurs systématiques nécessite des fonctions de validation, de localisation et de correction, tandis que la détermination des erreurs influentes nécessite des fonctions de validation et de score.

À noter que même si ces quatre fonctions couvrent l'ensemble de la stratégie, différentes méthodes peuvent être utilisées pour chaque fonction, ce qui fait que plus de quatre composantes fonctionnelles seront nécessaires. Toutefois, l'élaboration complète de cette stratégie en composantes fonctionnelles dépasse la portée du présent article.

4. Conclusions et recherche à venir

Jusqu'à maintenant, les recherches qui ont été menées sur la normalisation des processus statistiques ont montré que l'idée de concevoir des processus statistiques à partir d'un répertoire d'éléments génériques est faisable. L'objectif de plus grande transparence dans la conception du processus semble être confirmé par deux études de cas. À cette étape de la recherche, les objectifs de souplesse et de réutilisation ne peuvent être prouvés, même si on a procédé à une certaine forme de réutilisation à l'intérieur d'un processus statistique.

Après avoir mis l'accent sur les fonctions comprises dans la vérification des données, nous en élargirons la portée. Nous avons déjà déterminé environ 15 fonctions qui sont les plus fréquemment utilisées dans les processus statistiques (voir Renssen et Camstra, 2011b). Celles-ci devront être élaborées davantage pour pouvoir être décrites comme composantes fonctionnelles.

L'objectif de la présente recherche est principalement descriptif et nous ne sommes pas directement touchés par la mise en œuvre proprement dite des étapes de processus normalisées et leur lien avec la boîte à outils standard. Cette question sera abordée à une étape ultérieure de la collaboration avec les architectes des TI. L'objectif ultime consiste à élaborer un répertoire de composantes fonctionnelles (sur papier), ainsi que de composantes des TI, qui serviront de base à la fois pour la conception et la mise en œuvre des processus de production statistique.

Bibliographie

Braaksma. B. (2009), « Redesigning a Statistical Institute: The Dutch case », *Proceedings of MSP2009, workshop on Modernisation of Statistics Production 2009*.

Camstra, A. et R. Renssen, (2011), « Standard process steps based on standard methods as part of the business architecture », *Proceedings of World Statistics Congress* (*ISI2011*), Dublin, Irlande.

Gelsema, T. et F. Hofman, (2011), « Conceptual Information Model for Standard Process Steps », rapport non publié (rédigé en néerlandais), La Haye, Pays-Bas, Statistics Netherlands.

Hoogland, J., Van der Loo, M., Pannekoek, J., et S. Scholtus (2010), « Methodology series, theme Data validation and correction », rapport non publié, La Haye, Pays-Bas, Statistics Netherlands.

Huigen, R., Bredero, R., Dekker, W. et R. Renssen (2009), « Statistics Netherlands Architecture; Business and Information model », rapport non publié, La Haye, Pays-Bas, Statistics Netherlands.

Renssen, R., Morren, M., Camstra, A. et T. Gelsema (2009), « Standard processes », rapport non publié, La Haye, Pays-Bas, Statistics Netherlands.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Codage des réponses dans les enquêtes – Efforts d'assurance de la qualité et outils des TI à Statistics Sweden

Jörgen Svensson[1]

Résumé

En plus de prendre beaucoup de temps, le codage des réponses dans les enquêtes est coûteux et est sujet à plusieurs erreurs. La présente communication porte sur les efforts de Statistics Sweden en vue de s'assurer de la qualité du processus de codage. Des travaux ont été effectués dans de nombreux domaines.

La vérification du codage sera mise en œuvre durant l'exercice 2011-2012 pour toutes les enquêtes pertinentes. Dans le cas du codage (manuel) assisté par ordinateur, le contrôle de la qualité comprendra la vérification indépendante du codage, effectuée sur au moins 5 % des enregistrements codés dans une enquête. Si le codage original et le codage de vérification diffèrent, un processus d'arbitrage (de rapprochement) sera suivi pour décider du code « correct ». Si un codeur commet fréquemment les mêmes erreurs, son travail sera vérifié et une formation appropriée lui sera donnée. Dans le cas du codage automatique, un contrôle analogue de la qualité sera effectué au moins tous les trois ans, dans une enquête appropriée. Si le taux d'erreur est inacceptable, des révisions doivent être apportées au dictionnaire.

En outre, des directives ont été établies pour les listes de codes (lorsqu'il n'existe pas de classifications types). Toutes les classifications et listes de codes sont sauvegardées dans un entrepôt commun. Les divers employés qui participent au processus de codage au sens large doivent coopérer afin d'assurer la qualité du processus. À cette fin, les divers rôles ont été définis clairement. Comme le codage des réponses aux questions ouvertes est souvent de nature subjective, une formation appropriée est essentielle. À Statistics Sweden, les codeurs reçoivent une formation en ce qui concerne la façon d'utiliser les classifications, les fichiers de référence, etc., c'est à dire qu'on les sensibilise aux règles concernant ce qui doit être inclus ou exclu dans un code particulier. La formation est prodiguée au moyen d'une application Internet raccordée à une base de données assortie d'exercices, qui sont offerts avec trois niveaux de difficulté.

Un outil moderne des TI pour le codage assisté par ordinateur a été développé en étroite collaboration avec les codeurs. Le soutien pour les décisions de codage est fourni au moyen d'une interface conviviale. Des fonctionnalités sont également disponibles pour traiter les droits d'accès des codeurs originaux, des codeurs de vérification et des « arbitres ».

Le système intranet d'information sur les processus fournit des renseignements sur le processus de codage, les mesures d'assurance de la qualité requises et l'outil des TI. Le suivi de la mise en application sera effectué par l'entremise des directeurs des programmes spécialisés et de collecte des données.

Mots clés : Codage ; codage assisté par ordinateur ; erreurs de codage ; listes de codes ; normalisation ; codage de vérification ; rapprochement.

1. Introduction

Dans ce contexte, le *codage* est défini de la façon suivante : utiliser des codes pour classer des objets d'enquête dans différentes catégories, selon une classification établie ou une autre liste de codes définie au préalable. Le présent document porte sur les efforts déployés par Statistics Sweden ces dernières années, en vue de s'assurer de la qualité du processus de codage. Plusieurs raisons ont justifié que l'on mette l'accent sur ce processus. L'une d'elles est que le processus de codage des réponses aux enquêtes est sujet à l'erreur. C'est ce qu'ont démontré des évaluations du codage des professions effectuées en 2007 à Statistics Sweden. Des contrôles réguliers de la qualité sont donc nécessaires. Une autre raison est que le codage prend du temps et est donc coûteux. Une rationalisation, grâce au soutien moderne des TI, est requise. En 2008, le directeur général a décidé que Statistics Sweden devait tenter d'obtenir une certification selon la norme internationale ISO 20252 pour les études de marché, les études sociales et d'opinion (voir Organisation internationale de normalisation (2006)). La norme exige que le processus de codage fasse l'objet de mesures d'assurance et de contrôle de la qualité. Les exigences sont assez précises du point de vue du pourcentage d'enregistrements qui doivent faire l'objet d'une vérification. À Statistics Sweden, on a examiné en long

[1]Jörgen Svensson, Statistics Sweden, Département des processus, Örebro, SE-701 89, Suède, courriel : jorgen.svensson@scb.se.

et en large la façon de répondre à ces exigences. L'objectif global des travaux généraux décrits dans le présent document est d'améliorer de façon continue le processus de codage.

2. Contrôle de la qualité du codage

Statistics Sweden a établi une démarche uniforme pour le contrôle de la qualité du processus de codage, qui est décrit dans le système d'information sur les processus dans intranet. Les principales activités consistent à procéder à un *codage de vérification indépendant*, après le codage original et – si le code original et le code de vérification diffèrent – à décider du code « correct », grâce à un processus de *rapprochement* (arbitrage). Le gestionnaire d'enquête, de concert avec les responsables du processus de codage, doit alors stipuler ce que l'on considère comme des *taux d'erreur inacceptables*, au sujet desquels des mesures doivent être prises pour obtenir une meilleure qualité. Cette démarche a été adoptée en 2011 pour la majorité des enquêtes pertinentes.

Le *codage automatique au moyen d'un dictionnaire* doit faire l'objet d'un contrôle de la qualité, grâce à un codage de vérification manuel indépendant assisté par ordinateur et à un rapprochement pour au moins 5 % des enregistrements codés de façon automatique, dans une enquête au moyen de questionnaires sur papier sélectionnés à cette fin. Ce contrôle de la qualité doit être effectué au moins une fois tous les trois ans, à partir de 2011. Si le taux d'erreur est inacceptable, le dictionnaire doit être révisé. (La somme de codage automatique est relativement faible à Statistics Sweden.)

Le *codage* pour l'Enquête sur la population active, *au moment des interviews téléphoniques*, à partir d'une liste de professions avec le système par ITAO, doit faire l'objet d'un contrôle de la qualité, grâce à un codage de vérification manuel indépendant assisté par ordinateur et ensuite à un rapprochement, pendant une période de trois mois (cet exercice est limité à la première fois que les enregistrements apparaissent dans cette enquête longitudinale). À partir de 2011, ce contrôle de la qualité doit être effectué au moins une fois tous les trois ans. Si le taux d'erreur est inacceptable, le gestionnaire d'enquête, de concert avec le responsable du processus, doit décider des mesures à prendre pour améliorer les résultats.

Le *codage manuel assisté par ordinateur* au moyen d'un fichier de référence exhaustif doit faire l'objet d'un contrôle de la qualité, grâce à un codage de vérification manuel indépendant assisté par ordinateur et à un rapprochement pour au moins 5 % des enregistrements codés dans chaque enquête pertinente. Ce contrôle de la qualité représente une forme d'échantillonnage d'acceptation et doit être exécuté de façon continue, à partir de 2011. Le pourcentage sera souvent supérieur à 5 %, étant donné qu'il doit être possible de mesurer les résultats du codage pour chaque codeur. Si un codeur commet fréquemment les mêmes erreurs, son travail doit être vérifié et une formation appropriée doit lui être donnée. Des efforts seront peut être nécessaires pour améliorer le niveau de compétence, tant des personnes que du groupe. Mise à part la formation, les fichiers de référence et les instructions peuvent être améliorés, afin de réduire les taux d'erreur dans le codage. À noter que la limite de 5 % ne comporte pas de motivation statistique, mais représente une exigence pour l'accréditation selon la norme ISO 20252.

Le *codage externe* est une option courante à Statistics Sweden. De deux choses l'une : les fournisseurs des données procèdent au codage eux mêmes et nous envoient les enregistrements codés, ou les fournisseurs de données codent les enregistrements selon leurs propres nomenclatures et nous les envoient, avec des clés de traduction qui servent à établir les codes de classification. Dans le cas du codage externe, il n'est pas possible d'effectuer un contrôle de la qualité selon la démarche mentionnée précédemment. Toutefois, des évaluations et une assurance de la qualité sont recommandées pour les enquêtes concernées.

L'*échantillonnage pour le codage de vérification* peut être conçu de diverses façons. Une option simple consiste à procéder à un échantillonnage aléatoire simple. Une autre option est l'échantillonnage aléatoire simple stratifié, et avec des strates correspondant aux codeurs. Une troisième option est l'échantillonnage systématique stratifié, qu'il est pertinent d'utiliser lorsque le codage de vérification doit être mené en parallèle avec le codage ordinaire à l'intérieur de la période de production de l'enquête.

Le contrôle de la qualité comporte un *objectif* double. D'une part, le processus de codage doit être amélioré, afin que les erreurs soient moins fréquentes au cycle d'enquête suivant. Par ailleurs, les données d'enquête doivent (si le temps le permet) être rajustées directement, grâce à des corrections des enregistrements erronés au moyen des codes

finaux, possiblement après rapprochement, des enregistrements de vérification échantillonnés, ou peut être grâce à des rajustements du niveau de population à partir de ces codes corrigés. Pour commencer, on visera le premier objectif dans la plupart des cas.

Les lacunes dans les résultats du codage peuvent être présentées de deux façons. Tout d'abord, les taux d'erreur pour *chaque codeur* sont calculés directement, à partir de données non pondérées sur les codes finaux tirées des enregistrements de vérification échantillonnés. Le résultat sert à assurer le suivi des codeurs et des groupes de codeurs qui affichent des erreurs fréquentes. En deuxième lieu, les *erreurs brutes* et les *erreurs nettes* sont calculées, afin de montrer les effets sur les microdonnées (importantes pour les analyses des associations statistiques et des flux entre les catégories) et les macrodonnées (statistiques descriptives), respectivement. Ces erreurs peuvent être produites de façon non pondérée pour les enregistrements de vérification échantillonnés et pondérée pour la population. Il reste un problème pratique à résoudre, à savoir la définition d'un taux d'erreur inacceptable. Les caractéristiques propres à un produit doivent alors être prises en compte. Des niveaux différents, par exemple, dans la classification des professions ISCO, c'est à dire un nombre différent de chiffres dans les codes, imposeraient des exigences différentes en ce qui a trait au taux d'erreur. Une approche de contrôle de processus devrait être appropriée. Le niveau d'erreur doit alors être contrôlé grâce à un échantillonnage d'acceptation. Un graphique de contrôle peut être utile; il montre quand le taux d'erreur dépasse la variation aléatoire normale ou dévie vers des niveaux trop élevés.

3. Prisma – un nouvel outil des TI de codage assisté par ordinateur

Un *outil moderne des TI* pour le codage assisté par ordinateur, appelé Prisma, a été développé à Statistics Sweden au cours de la période de novembre 2010 à avril 2011. Ce développement était nécessaire parce que les applications des TI utilisées pour l'Enquête sur la population active et plusieurs autres enquêtes sont pratiquement périmées et dépendent trop de quelques personnes. Il était aussi nécessaire d'élaborer des fonctions pour le codage de vérification et le rapprochement, selon la nouvelle démarche uniforme décrite précédemment. Les fonctions nécessaires ne se retrouvaient dans aucun outil commercial des TI, ni dans aucun autre utilisé dans les bureaux statistiques.

Prisma est programmé en C#.NET et appuiera le codage de toutes les classifications différentes et cela pour (presque) toutes les enquêtes. Les classifications sont faciles à mettre à jour et à charger à nouveau. L'outil fournit une interface conviviale pour le codeur, ce qui est important pour l'environnement de travail. Prisma rationalisera le travail des codeurs en appuyant les décisions de codage. Lorsqu'un nouvel enregistrement est ouvert, une recherche automatique est effectuée dans les fichiers de référence, *etc*. Si un résultat est trouvé, le codeur n'a qu'à cliquer sur la catégorie appropriée. Puis, tous les codes pour les différentes classifications d'un domaine, comme les professions, sont établis automatiquement. Il arrive parfois qu'aucun résultat ou que beaucoup de résultats soient obtenus et que le codeur doive effectuer des recherches pour recueillir plus de renseignements concernant la personne et la profession. Les enregistrements qui causent des difficultés peuvent être placés sur une liste d'attente. Il est possible d'importer des images balayées de questionnaires. Des fonctions pour le traitement des droits d'accès des codeurs originaux, des codeurs de vérification et des « responsables de la vérification » sont aussi disponibles. Au départ, quelques techniques d'échantillonnage sont utilisées et des données de processus sont produites. La configuration des différentes classifications et enquêtes est assurée grâce à un module dans Prisma.

Le projet de développement a été mené en étroite collaboration avec le personnel de codage qui, à différentes étapes préliminaires, a mis à l'essai le logiciel. Leurs souhaits et demandes ont été intégrés dans une large mesure dans la spécification des exigences pour Prisma. Des séminaires ont été organisés pour présenter l'outil des TI et le processus de codage uniforme. En mai 2011, le directeur général a approuvé Prisma comme outil uniforme pour le codage assisté par ordinateur.

Le plan était et est toujours de *mettre en œuvre* Prisma dans toutes les enquêtes pertinentes. Parmi ces quelques 10 à 15 enquêtes, l'Enquête sur la population active est la plus importante et aussi la première pour laquelle Prisma a été mise en œuvre (en avril mai 2011). Les classifications utilisées dans cette enquête sont ISCO pour les professions (en deux versions), la classification socioéconomique suédoise, la classification type des industries de Suède, le secteur et le comté.

4. Formation et autres mesures d'assurance de la qualité

Comme le codage des réponses aux questions ouvertes est souvent de nature subjective, une formation appropriée est essentielle. À Statistics Sweden, les codeurs reçoivent une formation en ce qui concerne la façon d'utiliser les classifications, les fichiers de référence, *etc*., c'est à dire qu'ils sont sensibilisés aux règles concernant ce qui doit être inclus ou exclu dans un code particulier. La formation est prodiguée au moyen d'une application Internet raccordée à une base de données assortie d'exercices, qui sont offerts avec trois niveaux de difficulté.

Les différents collègues, au sens large, qui participent au processus de codage doivent coopérer afin d'assurer la qualité du processus. À cette fin, grâce à une décision du directeur général, les divers rôles ont été clairement définis. Parmi ces derniers figurent : le rôle du codeur, de l'intervieweur, du gestionnaire d'enquête, du responsable de la classification, du responsable du processus (du processus et de l'analyse) et celui du personnel des TI responsable de Prisma. Le propriétaire de la classification, par exemple, devra réviser le dictionnaire de classification si le taux d'erreur est inacceptable selon le contrôle de qualité.

Des *directives abrégées ont été établies pour l'élaboration des listes de codes* (lorsqu'il n'existe pas de classifications types). Elles mettent notamment l'accent sur la façon d'élaborer des catégories et des instructions, ainsi que de traiter les « autres » catégories (globales). Le gestionnaire d'enquête est, selon les instructions, chargé de la production, de la documentation et de la révision possible de la liste de codes. Toutes les classifications et toutes les listes de codes doivent être conservées dans un répertoire commun.

Le personnel responsable du codage est principalement *centralisé* dans un des deux services de collecte des données. La centralisation complète est presque terminée. La justification d'une approche centralisée est de regrouper des codeurs compétents qui travaillent de façon similaire au moyen d'un outil commun des TI. Les dialectes des codeurs de différentes enquêtes devraient être évités dans la mesure du possible, afin d'assurer l'uniformité et la comparabilité. Par ailleurs, l'« écart entre les codeurs » devrait être réduit au minimum, grâce à l'utilisation de nombreux codeurs pour chaque enquête, plutôt que d'un groupe de codage divisé entre quelques codeurs par enquête.

L'*information* sur le processus de codage et les mesures d'assurance de la qualité requises est comprise dans un système intranet d'information sur les processus. Le suivi de la mise en application sera effectué par l'entremise des directeurs des programmes spécialisés et des services de collecte des données.

5. Conclusion et travaux futurs

Afin d'améliorer le processus de codage des réponses aux enquêtes à Statistics Sweden, des travaux exhaustifs ont été effectués au cours des dernières années. La principale tâche consiste maintenant à mettre en œuvre la méthode et la démarche adoptées, ainsi que Prisma, le nouvel outil des TI. Des analyses des taux d'erreur par codeur et des erreurs brutes et nettes dans les statistiques seront effectuées, et des mesures devront être prises pour améliorer le processus de codage. La sensibilisation et les compétences en matière d'assurance de la qualité du processus doivent être intensifiées chez toutes les parties concernées. La démarche adoptée doit être évaluée.

En 2012, un nouveau projet de développement de Prisma sera probablement entrepris. La spécification complète des exigences n'a pu être respectée dans le projet récent. L'utilisation de Prisma en 2011 a donné lieu à de nouvelles demandes de fonctions de la part des codeurs. Le codage automatique au moyen d'un dictionnaire pourrait être inclus dans la prochaine version de Prisma. Il faut aussi établir de meilleurs liens entre Prisma et d'autres outils normalisés des TI à Statistics Sweden. On prévoit mettre en œuvre le contrôle des interviews téléphoniques en 2012, ce qui pourrait mener à une démarche partiellement différente pour le contrôle de la qualité du codage des interviews téléphoniques.

Bibliographie

International Organization for Standardization (2006), Market, opinion and social research – Vocabulary and service requirements (ISO 20252:2006, IDT).

SÉANCE 8B

QUESTIONNAIRES ET EFFETS DU MODE DE COLLECTE

Dispositions relatives à la qualité des données d'enquête dans la solution de questionnaire électronique de Statistique Canada : Rétrospective et perspectives

Yamina Abiza[1]

Résumé

Le mode de collecte de données en ligne est relativement nouveau, et les possibilités, avantages et aspects méthodologiques complets liés à son utilisation, seul ou en parallèle avec d'autres modes (c'est-à-dire différents modes pour différents répondants), sont toujours explorés. Néanmoins, en ce qui a trait à la qualité des données d'enquête, ce mode de collecte présente un potentiel énorme permettant de dépasser les résultats d'autres modes bien établis, ou à tout le moins d'obtenir l'équivalent. Ce potentiel ne devrait qu'augmenter à l'avenir, du fait des progrès technologiques et de la recherche dans le domaine des méthodes d'enquête.

Au cours des trois ou quatre dernières années, les responsables de la collecte à Statistique Canada ont conçu et mis en œuvre une solution de collecte en ligne pleinement fonctionnelle, qui a servi à recueillir les données auprès des répondants des enquêtes de Statistique Canada. Le mode de collecte en ligne a été utilisé comme principal outil de collecte, en combinaison avec d'autres modes (par exemple, enquête sur papier et enquête téléphonique assistée par ordinateur).

Dans cette présentation, nous décrivons l'architecture de la solution de questionnaire électronique de Statistique Canada et ses caractéristiques, qui contribuent aux différents aspects de la qualité des données d'enquête. Nous décrivons aussi comment nous pouvons améliorer davantage notre solution en intégrant les résultats de la recherche de nouvelles méthodes d'enquête dans les pratiques exemplaires concernant l'utilisation des enquêtes en ligne fondées sur la collecte des données multimodale.

[1]Yamina Abiza, Statistique Canada.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Conception d'un questionnaire pour examiner le Programme de sports des Forces canadiennes

Krystal K. Hachey[1]

Résumé

La forme physique est l'une des exigences essentielles pour les membres actifs des Forces canadiennes (FC). Compte tenu de cela, des programmes sont offerts au personnel des FC et lui permettent d'atteindre et de maintenir les normes établies par les FC (Hillier, 2009). Un de ces programmes est le programme de sports des FC qui, en dépit de son importance, n'a pas encore été évalué. Cette communication fait partie d'une étude plus vaste qui examinera la participation des employés des FC au programme de sports des FC et leur satisfaction à cet égard. Les participants et les non participants seront échantillonnés. Le projet comporte cinq objectifs globaux : 1) examiner le type de personnes qui participe au programme de sports des FC ; 2) déterminer les raisons pour lesquelles les personnes participent au programme de sports des FC ; 3) examiner les avantages globaux du programme de sports des FC ; 4) déterminer si le programme de sports des FC respecte ses principaux objectifs; et 5) déterminer la satisfaction globale des participants à l'égard du programme de sports des FC. Cette communication vise à présenter l'approche méthodologique utilisée pour l'atteinte des cinq objectifs, y compris l'élaboration du questionnaire. La présentation portera sur l'élaboration du questionnaire, l'échantillonnage, ainsi que les questions méthodologiques qui se poseront.

1. Introduction

La condition physique fait partie intégrante de la vie des militaires. Des recherches antérieures ont montré que la participation à des sports sert de base pour l'amélioration des caractéristiques personnelles (par exemple, leadership et cohésion des équipes; Alimo Metcalfe et Alban Metcalfe, 2001), outre ses bienfaits pour la santé physique (par exemple, souplesse) et psychologique (Fentem, 1994). Même si des bienfaits sont liés à l'amélioration continue de la condition physique, la participation à des sports de groupe comporte aussi d'autres avantages uniques (Pate, Trost, Levin et Dowda, 2000). Les recherches ont montré que la participation à des sports de groupe organisés favorise l'esprit sportif, la compétitivité, la performance (Pate et coll., 2000), ainsi que les liens sociaux (Long, 2004; Sherry, 2010). Cela est particulièrement important pour tisser des liens avec d'autres personnes travaillant dans le même domaine ou s'acquittant de la même tâche urgente (par exemple, évacuation en mer) (Directives et ordonnances administratives de la Défense [DOAD] 5023 2, 2010).

Le Programme de sports fait partie intégrante des Forces canadiennes (FC), étant donné qu'il s'agit d'une façon d'assurer la formation et le perfectionnement des membres des FC (Ordonnances administratives des Forces canadiennes [OAFC] 50 3, 2010). Le personnel des FC a la possibilité de participer à des sports à la base (c'est-à-dire intramuros) et aux niveaux régional, national et international. Le programme vise principalement à développer la cohésion de l'unité et le travail en équipe, ainsi que les caractéristiques personnelles, comme l'estime de soi, l'abnégation et le leadership, et à favoriser la condition physique (OAFC 50 3, 2010). Dans l'ensemble, le Programme de sports des FC fournit aux membres la possibilité de participer à des sports de compétition tout en leur donnant la chance de développer des caractéristiques importantes qui font partie intégrante des FC.

Comme il existe certaines exigences physiques que le personnel des FC doit respecter, il est essentiel de pouvoir offrir des programmes d'activité physique. Le présent projet vise donc à examiner l'importance du Programme de sports des FC pour le recrutement, le maintien de l'effectif, les niveaux de condition physique et les bienfaits physiques et psychologiques, ainsi qu'à évaluer la satisfaction à l'égard du programme.

[1]Krystal K. Hachey, Ministère de la Défense nationale et les Forces canadiennes, Canada.

2. Méthodologie

2.1 Questions de recherche

Les principales questions de recherche étaient les suivantes :

1. Le Programme de sports des FC peut-il être utilisé comme outil de recrutement?
2. Existe t il un lien entre la participation au Programme de sports des FC et le maintien de l'effectif des FC?
3. La participation au Programme de sports des FC permet elle à l'effectif des FC de maintenir son niveau de condition physique?
4. La participation au Programme de sports des FC permet elle à l'effectif des FC d'acquérir les caractéristiques personnelles recherchées par les FC (leadership, cohésion d'équipe, *etc*.)?
5. Existe t il un lien entre la participation au Programme de sports des FC et les bienfaits pour la santé chez l'effectif des FC?
6. Les participants sont ils satisfaits du Programme de sports des FC?

2.2 Participants et enquête

L'échantillon aléatoire stratifié comprenait à la fois des personnes qui avaient participé au Programme de sports des FC et d'autres qui n'avaient pas participé, afin de fournir une base pour la comparaison des mesures de la santé et des influences sur le recrutement et le maintien de l'effectif. Les données ont été recueillies au moyen d'une enquête électronique, dans le cadre de laquelle les participants ont reçu un courriel comprenant une fiche d'information, un formulaire de consentement, ainsi qu'un lien à l'enquête. L'enquête était disponible en ligne pour les participants, de mai 2011 jusqu'à septembre 2011, et a obtenu un taux de réponse final de 34 %.

2.3 Élaboration de l'enquête

Pour pouvoir établir les principales variables devant être prises en compte, un examen exhaustif des ouvrages publiés sur la participation aux sports et à l'activité physique a été mené. On a utilisé dans la mesure du possible les questions d'enquête déjà employées au ministère de la Défense nationale (MDN) (par exemple, questionnaire sur la santé des recrues [QSR]) qui s'appliquaient à l'étude du Programme de sports des FC. Cela était important pour disposer d'une base de comparaison avec d'autres enquêtes. Comme aucune étude ne comportait d'examen de la participation au Programme de sports des FC et de la satisfaction à l'égard de ce programme, plusieurs questions ont été élaborées pour examiner ces domaines. Enfin, une fois une banque de questions d'enquête établie, celles ci ont été réparties en catégories particulières en fonction des questions de recherche.

Les sections qui suivent passent en revue les principales composantes du questionnaire et décrivent les questions qui ont été choisies pour répondre aux objectifs particuliers du projet.

2.4 Variables démographiques

Les variables démographiques comprenaient des variables courantes utilisées dans les questionnaires administrés aux échantillons des FC, y compris l'âge, le sexe, la langue officielle, le rang, l'uniforme distinctif de l'élément et le centre de soutien de la base/l'escadre.

2.5 Questions de recherche 1 et 2 : Le Programme de sports des FC peut il être utilisé comme outil de recrutement?et Existe t il un lien avec l'érosion?.

Afin de déterminer si la participation à des sports pourrait être utilisée comme outil de recrutement/maintien de l'effectif et si le Programme de sports des FC avait des répercussions sur l'érosion, six questions de l'Enquête sur le maintien de l'effectif des FC de 2010 (Holden, 2011 ; Howe, 2006) ont été utilisées. Les réponses ont été mesurées selon une échelle de Likert à cinq points allant de « définitivement pas à définitivement » (par exemple, « J'ai l'intention de demeurer dans les FC jusqu'à la fin de mon mandat actuel »). Par ailleurs, d'autres questions ont été élaborées par le chercheur principal, afin d'examiner le recrutement (par exemple, si les répondants étaient d'avis que le Programme de sports des FC pouvait être utilisé comme outil de recrutement), la façon dont ils avaient

entendu parler du programme (par exemple, dans le journal de la base, dans des médias civils, par des amis), les raisons de se joindre aux FC (par exemple, partir de la maison, rémunération et avantages des FC), les raisons de se joindre au Programme de sports des FC (par exemple, développer/améliorer un esprit de camaraderie), et les obstacles à la participation (par exemple, trop occupé).

2.6 Question de recherche 3 : La participation au Programme de sports des FC permet elle à l'effectif des FC de maintenir son niveau de condition physique?

Afin de déterminer si l'amélioration/le maintien des niveaux de condition physique du personnel des FC pourrait être assuré grâce à la participation au Programme de sports des FC, plusieurs questions du QSR, un outil de surveillance de base de la santé des recrues des FC, de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) (Statistique Canada, 2008) et du Sondage sur la santé et le style de vie des FC (SSSV ; Centre de recherche Decima, 2002, Prévention de la santé des forces, 2003) ont été utilisées. Les questions comprenaient les résultats des tests de condition physique, la taille et le poids pour le calcul de l'indice de masse corporelle (IMC) (ESCC, 2001), l'état de santé/santé mentale auto-déclaré (ESCC, 2001), les blessures physiques (y compris les blessures répétitives) (ESCC, 2001) et les stratégies de maintien de la condition physique. Ces domaines ont été inclus parce qu'ils portaient sur la santé physique en général, y compris les blessures pouvant nuire à l'amélioration ou au maintien des niveaux de condition physique du personnel des FC.

2.7 Question de recherche 4 : La participation au Programme de sports des FC permet elle à l'effectif des FC de développer les caractéristiques personnelles recherchées par les Forces canadiennes (par exemple, leadership)?

Afin de déterminer si les caractéristiques personnelles, comme le leadership, sont favorisées par le Programme de sports des FC, deux questions ont été élaborées : 1) « Croyez vous que votre participation au Programme de sports des FC a un effet sur la façon dont vous interagissez avec les membres de votre unité? » (réponse par oui ou par non) ; et 2) « Participez vous au Programme de sports des FC pour améliorer vos capacités de leadership? » (réponse par oui ou par non). Les deux questions comportaient un espace pour fournir des commentaires additionnels.

2.8 Question de recherche 5 : Existe t il un lien entre la participation au Programme de sports des FC et les bienfaits pour la santé chez l'effectif des FC, dans les domaines comme la dépression, l'anxiété, le tabagisme et la consommation d'alcool?

Les questions du QSR et de l'ESCC ont servi à évaluer les comportements en matière de santé personnelle et de santé mentale. La santé mentale, y compris la dépression et l'anxiété, ont été mesurées au moyen de l'échelle de détresse psychologique de Kessler comprenant dix questions (K10 ; Kessler et coll., 2002). Les répondants ont coté les réponses aux questions sur une échelle à 5 points de Likert, allant de jamais à toujours (par exemple, « Vous sentiez vous nerveux? »). Les questions concernant la consommation d'alcool et le tabagisme ont été tirées de l'ESCC (ESCC, 2001). Parmi les points abordés figurait la situation d'usage du tabac (par exemple, n'a jamais fumé, ancien fumeur ou fumeur), ainsi que la fréquence de consommation de boissons alcoolisées.

2.9 Question de recherche 6 : Les participants sont ils satisfaits du Programme de sports des FC?

Afin d'évaluer la satisfaction du personnel des FC à l'égard du Programme de sports des FC, des questions ont été élaborées sur la base des recherches antérieures (Sigrist et coll., 2005; Chin, White, Howel, Harland et Drinkwater, 2006). On a demandé aux répondants d'évaluer leur satisfaction globale à l'égard des différents niveaux du Programme de sports des FC (par exemple, sport sur la base, sport régional, sport national, sport international, sport extrême/sport à l'extérieur des FC et autres) sur une échelle de Likert à cinq points, allant de complètement satisfait à complètement insatisfait.

3. Conclusion

L'objectif du projet consistait à fournir un aperçu général du Programme de sports des FC, y compris son influence sur le maintien de l'effectif, le recrutement, les niveaux de condition physique, les caractéristiques personnelles et la santé mentale. Même si l'enquête a porté sur tous les objectifs énoncés, elle comportait certaines limites. L'une des principales était qu'il s'agissait d'une enquête transversale, qui limite la capacité d'établir des rapports de cause à effet. De plus, la composante qualitative de l'enquête a été limitée en raison de contraintes liées à la longueur de l'enquête. Malgré ces limites, les résultats de l'étude actuelle fourniront des renseignements sur le Programme de sports des FC, du point de vue de son lien avec l'érosion/le maintien de l'effectif, la condition physique et les caractéristiques personnelles qui correspondent à celles recherchées par les FC. Étant donné qu'aucune autre étude n'a porté sur la participation au Programme de sports des FC ou sur les objectifs de ce programme, la présente étude servira de base pour les études futures sur les bienfaits du programme.

Bibliographie

Alimo-Metcalfe, B. et R.J. Alban-Metcalfe (2001), « The development of a new transformational leadership questionnaire », *Journal of Occupational and Organizational Psychology*, 74, 1 à 27.

Canadian Forces Administrative Orders (CFAO) 50-3 (2010), « CFAO 50-3 – Sports », manuscript non-publié.

Decima Research Inc. (2002), « *CF Health and Lifestyle Information Survey 2000 Regular Force Report* » (Report prepared for Canadian Forces Department of National Defence), Ottawa, Canada: Decima Research Inc.

Defence Administrative Orders and Directives (DAOD) 5023-2 (2008), « Physical fitness program », retrieved online August 13th from: http://admfincs.mil.ca/admfincs/subjects/daod/5023/2_e.asp.

Directorate of Force Health Protection (2003), « *Canadian Forces Health and Lifestyle Information Survey 2004 Regular Force Report* (A-MD-015-FHP/AF-001) », Ottawa, Canada: Department of National Defence, Directorate of Force Health Protection. Retrieved August, 2010 from: http://www.dnd.ca/health-sante/pub/hlis-sssv/pdf/ AMD015FHPAF001-20030901-eng.pdf

Fentem, P.H. (1994), « ABC of sports medicine: Benefits of exercise in health and disease », *British Medical Journal*, 308, 1291-1295.

Howe, D. (2006), « Building and sustaining a retention culture in the Canadian Forces », DGMPRA TR 2006-006, Ottawa, Canada: Director General Military Personnel Research and Analysis.

Hyams, K.C., Barrett, D.H., Duque, D., Engel, Jr., C.C., Friedl, K., Gray, G., Hogan, B., Kaforski, G., Murphy, F., North, R., Riddle, J., Ryan, M.A.K., Trump, D.H. et J. Wells (2002), « The Recruit Assessment Program: A program to collect comprehensive baseline health data from U.S. military personnel », *Military Medicine*, 167 (1), 44-47.

Kessler, *et al*. (2002), « Short screening scales to monitor population prevalences and trends in non-specific psychological distress », *Psychological Medicine*, 32, 959-976.

Long, J. (2004), « The social benefits of sport: Measurement and evaluation », retrieved online July 16th, 2010 from: http://www.cpl.biz/isrm/infonotesite/recreation/documents/REmay04_16_17.pdf.

Pate, R.R., Trost, S.G., Levin, S. et M. Dowda (2000), « Sports participation and health related behaviours among US youth », *Archives of Paediatric and Adolescent Medicine*, 154, 904-911.

Sherry, E. (2010), « (Re) engaging marginalized groups through sport: The homeless world cup », *International Review for the Sociology of Sport*, 45 (1), 59-71.

Sigrist, L.D., Anderson, J.E. et G.W. Auld (2005), « Senior military officer's educational concerns, motivators and barriers for healthful eating ad regular exercise », *Military Medicine*, 170 (10), 841-845.

Spitzer, R.L., Kroenke, K., Williams, J.B. et The Patient Health Questionnaire Primary Care Study Group (1999), « Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study: Primary Care Evaluation of Mental Disorders: Patient Health Questionnaire », *JAMA*, 282, 1737-1744.

Statistique Canada (2001), « Canadian Community Health Survey (CCHS) questionnaire for cycle 1.1 », retrieved on 12 December, 2007, from: http://www.statcan.ca/english/sdds/instrument/3226_Q1_V1_E.pdf

Statistique Canada (2008), « Canadian Community Health Survey (CCHS) 2008 questionnaire », retrieved online June 24, 2010 from http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2.

Thompson, M.M. et L.S. Smith (2002), « Peace Support Operations Predeployment Survey: Scale reliability analysis », DRDC CORA TR 2002-190, Toronto, Canada: Defence Research and Development Canada – CORA.

Ware, Jr. J.E. et C.D. Sherbourne (1992), « The MOS 36-item short-form health survey (SF-36), I. Conceptual framework and item selection » *Medical Care*, 30 (6), 473 à 483.

Contenu harmonisé : Le nouveau paradigme d'élaboration des enquêtes à Statistique Canada

Richard Nadwodny et Pamela Best[1]

Résumé

Comme de nombreux autres organismes statistiques dans le monde, Statistique Canada réalise une vaste gamme d'enquêtes gouvernementales et spéciales, conçues à des moments différents pour répondre à des besoins distincts. Au fil du temps, ce contexte a entraîné un manque de cohésion dans les concepts, les définitions, les classifications et les questions d'enquête proprement dites, qui cause un manque d'harmonie entre les sources de données d'enquête – ce qui rend particulièrement difficile la comparaison de données sur un même thème ou une même variable provenant d'enquêtes différentes.

Au cours des trois dernières décennies, Statistique Canada a reconnu qu'il fallait améliorer la rentabilité et la souplesse du programme d'enquêtes auprès des ménages. En 2005, l'organisme a adopté la Nouvelle stratégie pour les enquêtes auprès des ménages, un livre blanc établissant la vision et les priorités, afin de lancer le processus de transformation nécessaire pour mieux s'adapter aux besoins des clients, être plus concurrentiel financièrement et pouvoir fournir des données d'enquête plus rapidement et pour une gamme plus vaste de produits, grâce à une plus grande intégration et harmonisation du contenu et des processus d'enquête.

Des groupes de travail ont été établis pour examiner les diverses méthodes d'enquête, comme les plateformes de collecte, le nivellement de la charge de travail, les interviews continues, le remaniement de certaines enquêtes de prestige, les outils de traitement communs et l'uniformisation des modules de questions les plus couramment utilisés dans les enquêtes auprès des ménages.

Le présent document décrit comment deux composantes importantes de la Nouvelle stratégie pour les enquêtes auprès des ménages, l'uniformisation des modules de questions et le développement d'outils de traitement communs, ont contribué à modifier le paradigme d'élaboration des enquêtes à Statistique Canada.

Mots clés : Harmonisation des données ; uniformisation des questions ; outils de traitement communs.

1. Introduction

1.1 Description

Selon un vieil adage, les bonnes idées ne disparaissent jamais; il faut seulement plus de temps à certaines pour porter ses fruits. Même si l'on a utilisé des titres différents depuis les années 1970 pour décrire le sujet (coordination des données, harmonisation, uniformisation, intégration des données/statistiques), l'intention est toujours demeurée la même, à savoir : réduire le chevauchement des fonctions de traitement et d'entreposage, réduire les coûts d'entretien des systèmes, réduire le fardeau pour les répondants et profiter des économies découlant de l'utilisation de modules uniformisés de concepts, de questions, de processus et de classifications[2]. L'harmonisation des données est le terme actuellement utilisé à Statistique Canada pour décrire ce concept.

En 2005, Statistique Canada a adopté la *Nouvelle stratégie pour les enquêtes auprès des ménages* (NSEM), un livre blanc établissant la vision et les priorités, afin de lancer le processus de transformation nécessaire pour mieux s'adapter aux besoins des clients, être plus concurrentiel financièrement et pouvoir fournir des données d'enquête

[1]Richard Nadwodny, gestionnaire principal de projet, Projet d'harmonisation des données, Division des enquêtes spéciales, Statistique Canada, Ottawa (Ontario) Canada, KIA 0T6, richard.nadwodny@statcan.gc.ca. Pamela Best, directrice adjointe, Division des enquêtes spéciales, Statistique Canada, Ottawa (Ontario) Canada, KIA 0T6 pamela.best@statcan.gc.ca.
[2]Priest, Gordon. 1996. *The Issue of Harmonization of Data from Diverse Sources*. Communication sur invitation à l'atelier d'EUROSTAT sur l'harmonisation, Londres, Angleterre, novembre 1996.

plus rapidement et pour une gamme plus vaste de produits, grâce à une plus grande intégration et harmonisation du contenu et des processus d'enquête.

Au cours des décennies qui ont précédé la NSEM, les enquêtes sociales menées par Statistique Canada ont été de plus en plus ambitieuses, visant simultanément à élargir et à approfondir leur contenu ainsi qu'à être plus spécifique au niveau géographique. Cela a eu pour conséquence que l'organisme a repoussé les limites des capacités des répondants de fournir des données suffisamment précises, ce qui a rendu nombre de nos enquêtes coûteuses et lourdes, difficiles à administrer et lentes à produire des résultats.

La NSEM nous a fourni l'occasion de mieux nous positionner pour mener des enquêtes simples à petite échelle, dans des délais courts, en tenant compte de l'évolution de l'environnement de collecte imposée par les clients, qui souhaitaient obtenir leurs données plus rapidement et à un moindre coût que par le passé. L'architecture du plan opérationnel de la NSEM a été adoptée par les cadres supérieurs, et des groupes de travail ont été rapidement établis pour analyser et recommander des changements au programme d'élaboration des enquêtes auprès des ménages de Statistique Canada.

Le présent document porte sur la façon dont deux composantes importantes de la NSEM (**harmonisation des données** et **outils communs**) ont été élaborées, leurs résultats et leur mise en œuvre.

1.2 Harmonisation des données

Au Symposium sur les questions de méthodologie, en 1995, Gordon Priest faisait état d'un problème, à savoir que les méthodes, systèmes, concepts, définitions, classifications, produits et services étaient élaborés de façon indépendante, ce qui donnait lieu à des problèmes d'efficacité, de chevauchement et de manque d'harmonie et à certaines frustrations de la part des clients[3]. Il proposait l'utilisation de l'intégration des données/statistiques comme une façon de réduire ces problèmes.

Aujourd'hui, nous utilisons le terme « harmonisation des données », qui se rapporte au processus d'élaboration de modules de questions uniformisés pour les variables d'enquêtes auprès des ménages qui se recoupent. Ces modules comprennent des concepts, définitions, classifications et libellés uniformes pour de multiples modes de collecte. On s'attendait à ce que l'utilisation de modules de questions uniformisés favorise l'amélioration de l'efficacité et de l'actualité, grâce à la réutilisation des spécifications d'IAO (interview assistée par ordinateur), de la mise à l'essai, du traitement, de la documentation et de la diffusion.

Au fil du temps, comme les questions étaient posées de différentes façons, on a assisté à un manque de cohérence dans les concepts, les définitions et les classifications, ce qui a donné lieu à un manque d'harmonie entre les sources d'enquête et a rendu particulièrement difficile la comparaison de données de différentes sources d'enquête pour le même thème ou une même variable.

Le manque actuel d'harmonie dans l'analyse des données pourrait s'atténuer si les clients savaient que les différentes enquêtes de Statistique Canada utilisent les mêmes questions pour un thème particulier, et que les questions ont été regroupées et traitées de la même façon, afin que les résultats puissent être analysés et comparés avec confiance.

Un exemple de ce manque d'harmonie ressort de l'analyse de la question sur l'occupation du logement, qui a été posée de différentes façons dans les enquêtes de Statistique Canada menées en 2006. Dans le cadre de l'Enquête nationale auprès des ménages de 2011, on a utilisé la version uniformisée de la question, qui est fournie en référence.

[3]Priest, Gordon. 1995. « Data Integration: The View from the Back of the Bus ». *Recueil du Symposium 1995 de Statistique Canada, Des données aux renseignements – Méthodes et systèmes*, novembre 1995, Ottawa, n° 11-522-XPF au catalogue de Statistique Canada.

RECENSEMENT DE 2006	ENQUÊTE SOCIALE GÉNÉRALE (ESG)	ENQUÊTE SUR LES DÉPENSES DES MÉNAGES (EDM)	EPA, ESCC, EDTR	ENQUÊTE NATIONALE AUPRÈS DES MÉNAGES (ENM) DE 2011
Êtes-vous (ou un membre du ménage est-il) : propriétaire de ce logement ou en train de le payer?	**Le propriétaire de ce logement est-il un membre de votre ménage?** 1. Oui 2. Non NSP, R **Payez-vous un loyer pour vivre dans ce logement?** 1. Oui 2. Non NSP, R	**Au 31 décembre 2007, votre logement était-il :** 1. Possédé **sans hypothèque** par votre ménage? 2. Possédé **avec une hypothèque** par votre ménage? 3. Loué **par votre ménage**? 4. Occupé gratuitement **par votre ménage** (c.-à-d. qu'aucun membre ne possède le logement et qu'aucun loyer n'est exigé)?	**Ce logement appartient-il à un membre de ce ménage?** 1. Oui 2. Non NSP, R	**Êtes-vous (ou un membre du ménage est-il) :** 1. propriétaire de ce logement ou en train de le payer? 2. locataire (même si aucun loyer en argent n'est versé)?

EPA : Enquête sur la population active, **ESCC :** Enquête sur la santé dans les collectivités canadiennes, **EDTR :** Enquête sur la dynamique du travail et du revenu

Par conséquent, l'effet des différences dans la façon dont les questions sur le mode d'occupation du logement (logements possédés et logements loués) ont été posées a eu des répercussions sur la comparabilité des données entre les différentes sources d'enquête. Même si ces différences pourraient être attribuées à des différences dans la taille d'échantillon, la méthode de collecte, les instructions aux intervieweurs, etc., qui sont les raisons habituelles des problèmes de comparabilité entre les enquêtes, l'élimination des incohérences dans la façon dont les questions étaient libellées, même si elles portaient sur les mêmes données, a constitué un point de départ évident de l'amélioration de la qualité des données.

Ménages propriétaire	Nombre	Pourcentage
EDM	8 215 000	66 1 %
EDTR	8 603 731	68,7 %
Recensement	8 381 125	68,5 %
Ménages locataires	**Nombre**	**Pourcentage**
EDM	4 218 331	33,9 %
EDTR	3 922 799	31,3 %
Recensement	3 861 155	31,5 %

Le contenu harmonisé comportait des modules de questions uniformisés déjà mis à l'essai, accompagnés de spécifications de traitement et d'un code BLAISE programmés au préalable, approuvés et vérifiés, permettant aux enquêtes d'être plus souples au chapitre des approches de collecte des données et en mesure de répondre aux besoins des clients plus rapidement, en réduisant le temps et le coût de mise en œuvre d'une enquête. Les blocs de questions uniformisés se sont aussi révélés avantageux pour les intervieweurs par interview téléphonique assistée par ordinateur (ITAO)/interview sur place assistée par ordinateur (IPAO), qui se sont habitués au libellé et à l'ordonnancement des questions uniformisés, ce qui a accéléré le temps de réponse par question et réduit le temps de formation.

L'ENM a aussi fourni l'occasion de collaborer plus étroitement avec les responsables du recensement, afin de mettre en commun des expériences sur une vaste gamme de domaines, y compris le contenu, le codage et les systèmes de classification, les bases de sondage, l'expertise technologique, les outils de gestion, les relations avec les médias et les répondants, et les études sur l'effet du mode (le recensement comporte un plan de sondage multimodal – courrier, intervieweur et Internet), qui pourraient servir de modèles pour une collaboration entre le bureau central et les bureaux régionaux au sujet des enquêtes auprès des ménages en général[4].

1.3 Processus de développement d'un contenu harmonisé

La Division des enquêtes spéciales a été mandatée pour gérer le projet d'harmonisation, qui a commencé par un examen interne des pratiques actuelles de l'organisme, et plus particulièrement par la détermination et la documentation des enquêtes auprès des ménages, y compris le recensement, qui comportaient des questions ou des modules de questions sur les thèmes les plus courants, comme l'état matrimonial, la santé, la scolarité, les langues, etc. Des similitudes et/ou différences dans la façon dont ces questions étaient posées, regroupées et documentées, ont été notées. Le projet a aussi permis d'examiner les normes internationales élaborées par les Nations Unies et d'autres organismes internationaux. Une fois l'analyse des pratiques actuelles de l'organisme et des recommandations internationales documentée, des groupes de travail d'experts spécialisés et d'intervenants ont été réunis pour passer en revue la documentation et entreprendre le processus de recommandation de modules de questions uniformisés pour 18 des thèmes multisectoriels les plus courants utilisés dans les enquêtes auprès des ménages. Les membres du groupe de travail étaient constitués d'analystes de la Division des normes responsables d'élaborer des concepts, définitions et classifications uniformes pour les questions uniformisées recommandées. Les métadonnées des Normes constituaient une composante très importante de ce projet, parce que sans elles, les données des enquêtes ou des recensements ne peuvent être diffusées.

Une fois terminée l'analyse par les groupes de travail spécialisés, il a été déterminé que plusieurs variantes de la même question nécessitaient un essai auprès des répondants. De concert avec le Centre de ressources en conception de questionnaires (CRCQ), différentes présentations et variantes de questions ont fait l'objet d'un essai qualitatif partout au Canada, auprès de groupes de discussion et sur une base individuelle. Après que le groupe de travail ait passé en revue les résultats des essais et ait pris sa décision concernant les questions devant être incluses dans le programme de contenu harmonisé, les recommandations ont été soumises aux cadres supérieurs, à divers niveaux, jusqu'à ce que l'adoption des normes soit approuvée.

2. Création d'outils communs

Les modules de questions uniformisés et les métadonnées connexes ne permettent pas à eux seuls de réaliser les économies requises pour que la *Nouvelle stratégie pour les enquêtes auprès des ménages* soit un succès. Les économies et les gains d'efficacité véritables sont le résultat du développement d'un nouveau système de traitement utilisé de façon universelle par toutes les divisions de l'organisme qui mènent des enquêtes auprès des ménages.

Dans ce contexte, le projet d'outils communs a été créé pour harmoniser les processus opérationnels des enquêtes auprès des ménages, afin d'élaborer les outils communs qui permettront aux secteurs d'enquête de créer, traiter et diffuser efficacement les données des enquêtes sociales. L'Environnement des métadonnées des enquêtes sociales (EMES) a été élaboré sur la base de quatre outils communs : l'outil de développement du questionnaire (ODQ), l'outil pour les spécifications et le traitement (OST), l'outil pour le dictionnaire de données (ODD) et l'outil des variables dérivées (OVD).

Outre les améliorations de l'actualité et de la qualité des données, l'utilisation de questions ayant fait leurs preuves, favorise la cohérence entre les enquêtes. Il se peut qu'un répertoire de métadonnées connexes soit accessible, grâce à une composante de l'interface, en vue d'accéder aux blocs précodés.

[4]Statistique Canada. 2005. *New Household Survey Strategy: Summary Report.* Document interne, Ottawa (Ontario), 5 octobre 2005.

L'EMES a été conçue en fonction des exigences suivantes des utilisateurs :

1. créer des spécifications de questionnaire dans un environnement structuré, peu importe le mode de collecte;
2. fournir l'accès à un répertoire commun de blocs de spécifications de questionnaire pour tous les modes de collecte (IPAO, ITAO, questionnaires électroniques et questionnaires papier);
3. favoriser l'utilisation d'un contenu harmonisé;
4. favoriser l'utilisation des dernières normes d'IAO;
5. créer divers types de questionnaires permettant d'améliorer l'actualité et d'uniformiser le processus de diffusion de métadonnées, pour différents usages, relativement à tous les aspects du cycle de vie de l'enquête, de la pré-collecte à la diffusion.

L'outil commun le plus étroitement lié au projet de contenu harmonisé est l'outil de développement du questionnaire (ODQ), qui utilise des modules de questions uniformisés et permet aux employés spécialisés de définir et de diffuser les questionnaires rapidement au moyen d'une approche uniforme.

De façon plus particulière, la fonction de l'ODQ permet aux développeurs d'enquêtes :

- de regrouper, de gérer et d'uniformiser le travail lié au développement d'un questionnaire d'enquête;
- de permettre aux utilisateurs d'avoir accès à un répertoire de spécifications de questionnaire et d'élaborer un tel répertoire;
- de créer des spécifications de questionnaire dans un environnement structuré, peu importe le mode de collecte;
- de donner accès à un répertoire commun de blocs de spécifications de questionnaire, pour tous les modes de collecte (IPAO, ITAO, collecte électronique et collecte sur papier),
- de créer divers types de questionnaires permettant d'améliorer l'actualité et d'uniformiser le processus de diffusion de métadonnées, pour différents usages, relativement à tous les aspects du cycle de vie de l'enquête, de la pré-collecte à la diffusion;
- de favoriser l'utilisation d'un contenu harmonisé;
- de favoriser l'utilisation des dernières normes d'interview assistée par ordinateur (IAO),;
- de suivre les progrès du développement des enquêtes.

L'ODQ est actuellement équipé pour créer des spécifications Blaise pour la collecte par ITAO/IPAO, mais à l'avenir, il pourra créer des spécifications de questionnaire électronique, ainsi que de questionnaire papier[5].

3. Travaux futurs d'élaboration d'un contenu harmonisé et d'outils communs

3.1 Modules de questions uniformisés pour les questionnaires électroniques

L'exercice initial comprenait l'élaboration de questions uniformisées pour les questionnaires papier et les enquêtes par ITAO/IPAO. Statistique Canada a déjà exprimé le souhait de procéder à d'autres économies, grâce à l'élaboration des enquêtes futures dans l'environnement Internet, sous forme de questionnaires électroniques. En 2006, 18 % des questionnaires du recensement ont été remplis sur Internet. Cette proportion est passée à 54 % pour le Recensement/l'Enquête nationale auprès des ménages de 2011, ce qui constitue une bonne indication que, grâce à une stratégie de promotion sur Internet, les questionnaires électroniques deviennent le mode d'enquête préféré des Canadiens. D'autres enquêtes comme le Sondage auprès des fonctionnaires fédéraux et le Sondage sur la dotation sont déjà sur Internet, l'Enquête sur la population active faisant actuellement l'objet d'un essai pilote d'un questionnaire électronique mensuel, et des plans étant en voie d'élaboration pour une version Internet de l'Enquête sociale générale.

Les enquêtes Internet représentent une version hybride des enquêtes sur papier et des enquêtes par ITAO/IPAO. Par conséquent, il est impérieux que les travaux à venir touchant le contenu harmonisé comprennent l'élaboration de

[5]Statistique Canada. 2011. *Common Tools Project for Social Surveys: Bringing it all together*. Document interne, Ottawa (Ontario), juin 2011.

modules de questions uniformisés sur Internet. L'établissement d'un équilibre entre la maximisation de la portée du mode, c'est-à-dire fournir des conseils aux intervieweurs pendant les enquêtes avec intervieweur, l'utilisation d'une aide interactive dans un questionnaire électronique et la comparaison avec le questionnaire papier constituera l'un des principaux défis de l'élaboration de ce contenu.

3.3 Examen sur cinq ans et mise à jour du contenu harmonisé

Statistique Canada a pris la décision de passer en revue les questions uniformisées pour assurer le maintien de leur pertinence, cet examen devant être effectué sur une base quinquennale, en parallèle avec le programme du recensement.

Ce processus d'examen pourrait permettre d'ajouter, de modifier et de supprimer des questions uniformisées de la base de données. Les modifications ou ajouts nécessiteraient la même méthode de recherche que celle utilisée pour la première ronde d'élaboration du contenu harmonisé, c'est-à-dire l'établissement de groupes de travail spécialisés, des comparaisons internationales, des essais qualitatifs et des comités hiérarchiques de prise de décisions, etc.

Bibliographie

Colledge M. (1999), « Statistical Integration through Metadata Management », Direction des Statistiques, Organisation de coopération et de développement économiques, *International Statistical Review*, Paris, France.

Moore, T., Bailie, L. et G. Gilmour (2009), « Établissement d'une analyse de rentabilisation de la collecte des données du recensement par Internet », *Recueil du Symposium 2011 de Statistique Canada, Collecte des données : défis, réalisations et nouvelles orientations*, Ottawa, n° 11-522-X au catalogue de Statistique Canada.

Priest, G. (1995), « Data Integration: The View from the Back of the Bus », *Recueil du Symposium 1995 de Statistique Canada, Des données aux renseignements – Méthodes et systèmes*, novembre 1995, Ottawa, n° 11-522-XPF au catalogue de Statistique Canada.

Priest, G. (1996), *The Issue of Harmonization of Data from Diverse Sources*. Communication sur invitation à l'atelier d'EUROSTAT sur l'harmonisation, Londres, Angleterre, novembre 1996.

Priest, G. (1998), « Report on the Progress on the Harmonization of Social Statistics - Working Paper 3 », Conference of European Statisticians, Statistical Commission and Economic Commission for Europe, Genève, Suisse, 18 au 20 février 1998.

Statistique Canada (2004), *Policy on Standards (Revised)*, document interne, Ottawa (Ontario), 14 juillet 2004.

Statistique Canada (2005), *New Household Survey Strategy: Summary Report*, document interne, Ottawa (Ontario), 5 octobre 2005.

Statistique Canada (2009), *Questionnaire Development Tool, Business Requirements Version 1.1*, document interne, outils communs de la statistique sociale, de la santé et du travail, Ottawa (Ontario), 14 septembre 2009.

Statistique Canada (2011), *Common Tools Project for Social Surveys: Bringing it all together*, document interne, Ottawa (Ontario), juin 2011.

Commission de statistique – Conseil économique et social (1999), *Draft standards of the United Economic and Social Information System for data structure and metadata in international data exchange and dissemination*, n° 98-35662 (E) au catalogue des Nations Unies, Nations Unies, New York, mars 1999.

Bureau de statistique des Communautés européennes (EUROSTAT) et la Commission économique des Nations Unies pour l'Europe (2006), *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing*, New York et Genève.

Effets du mode d'étalonnage sur l'enquête hollandaise sur la criminalité

Bart Buelens et Jan van den Brakel[1]

Résumé

Dans l'enquête hollandaise sur la criminalité et la victimisation, on utilise divers modes de collecte des données de façon séquentielle : lorsque l'on n'obtient aucune réponse au moyen d'un mode, on utilise un mode différent pour communiquer à nouveau avec les répondants. On a observé des instabilités d'une année à l'autre des estimations GREG pour cette enquête. L'analyse a permis de supposer qu'elles étaient attribuables au moins en partie aux changements dans la composition du mode de réponse, ce qui indique la présence d'effets de mode. Parallèlement, les effets de sélection font en sorte que les sous populations contactées au moyen des divers modes diffèrent. Les effets de mode et les effets de sélection se confondent et sont impossibles à isoler. On présente une approche pratique pour réduire le problème des instabilités temporelles. On présume que le modèle qui sous tend l'estimateur GREG corrige entièrement la sélectivité de la réponse. Selon cette hypothèse, les différences dans la composition du mode de réponse d'une année à l'autre ne sont pas un signe de sélectivité non supprimée. Il peut exister d'autres causes, qui peuvent être responsables des effets de mode. Ces effets ne peuvent être supprimés, mais ils peuvent être nivelés, grâce à l'étalonnage des modes de réponse en fonction de repères fixes. Un tel étalonnage se fait grâce au prolongement simple du modèle de régression GREG. Du fait que la collecte de données au moyen de modes mixtes est devenue la norme de facto pour les enquêtes sociales à Statistics Netherlands, une attention particulière est accordée aux questions relatives à l'adoption de l'étalonnage des modes comme norme dans le processus statistique générique des enquêtes sociales.

[1]Bart Buelens et Jan van den Brakel, Statistics Netherlands, Pays-Bas.

SÉANCE 9A

NORMALISATION DANS LE CADRE D'ÉTUDES COMPARATIVES INTERNATIONALES : AVANTAGES ET DÉFIS

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Conception, normalisation et suivi des opérations d'enquête dans les études internationales à grande échelle en éducation

Ralph Carstens[1]

Résumé

La recherche à grande échelle en éducation est complexe, et plus particulièrement dans le cadre des études internationales faisant intervenir plusieurs langues et cultures. Le besoin de données comparables d'un pays à l'autre rend nécessaires la conception d'opérations d'enquête solides et normalisées, la fourniture de manuels et de lignes directrices complets, l'organisation de séances de formation efficaces, l'instauration d'un processus de contrôle de la qualité aux étapes clés du processus d'enquête, la définition de protocoles de données communs, ainsi que des analyses de validité et de fiabilité. Toutefois, compte tenu des structures organisationnelles, de l'expertise et des particularités nationales dans le domaine de la recherche par sondage, une approche offrant une certaine souplesse doit être adoptée pour mettre en œuvre les plans internationaux d'échantillonnage et de collecte des données et maintenir le niveau global de qualité et de comparabilité. L'une des considérations connexes consiste à déterminer s'il convient de centraliser certaines tâches, comme la traduction ou sa vérification, au niveau international ou d'en répartir la responsabilité au niveau national. En s'inspirant des stratégies et expériences découlant de l'Étude internationale sur la maîtrise de l'ordinateur et de l'information (EIMOI) de l'IEA (Association internationale pour l'évaluation du rendement scolaire) et du Programme pour l'évaluation internationale des compétences des adultes (PIAAC) de l'OCDE, l'article portera sur les objectifs et les limites de la normalisation et de la centralisation.

Mots clés : Normalisation ; opérations d'enquête ; normes techniques ; évaluations à grande échelle ; international.

1. Conception et normalisation des opérations d'enquête

1.1 Objectifs de qualité

Pour chaque nouvelle enquête ou évaluation transnationale, il faut tenir compte d'un nombre important de buts, d'attentes, de besoins, d'intrants, ainsi que de budgets et d'échéanciers souvent variés qui, mis ensemble, influent sur les cadres conceptuels, la conception et les opérations d'une enquête et, en bout de ligne, sur sa qualité. L'objectif global qui est poursuivi à l'étape de la conception d'une enquête et, par la suite, lorsque des révisions deviennent nécessaires au moment de sa mise en œuvre, est de maximiser la qualité générale de l'enquête ou ce que l'on appelle parfois la qualité *totale*. Certaines des dimensions de qualité communes (Brackstone, 1999, Biemer et Lyberg, 2003 et Biemer, 2010) peuvent être appelées dimensions des « utilisateurs », du fait qu'elles sont principalement établies par les intervenants et les commanditaires. Elles se rapportent à la pertinence, à l'abondance ou au caractère complet des données recueillies, à la rapidité avec laquelle les données, les analyses et les rapports deviennent disponibles, ainsi qu'à l'accessibilité et à la facilité d'utilisation d'un produit de données à grande diffusion pour favoriser les analyses secondaires. En outre, la façon dont la mise en œuvre opérationnelle des enquêtes transnationales est assurée au niveau international et dans les pays participants – qui sont conjointement considérés comme les « producteurs » des données – représente une composante essentielle de la chaîne de qualité d'une enquête, étant donné qu'elle peut avoir une incidence directe sur les aspects mentionnés précédemment et d'autres dimensions de la qualité, en particulier l'exactitude des estimations tirées des données recueillies au niveau national et la comparabilité de ces estimations entre les pays, les langues, les domaines démographiques et les cycles d'enquête. Le degré d'attention que l'on porte à ces qualités, collectivement, définit la mesure dans laquelle on peut utilement se servir des données pour produire des estimations, des comparaisons et des inférences.

Les dimensions de la qualité peuvent aussi être exprimées au moyen d'un modèle à trois niveaux de qualité des *produits*, des *processus* et de *l'organisation* (à ce sujet, voir Lyberg et Stukel, 2010). La qualité d'un produit de données est établie par les intervenants respectifs (dans le cas d'évaluations à grande échelle, il s'agit des pays

[1]Ralph Carstens, Centre de traitement des données et de recherche de l'IEA, 37, Mexikoring, 22297 Hambourg, Allemagne (ralph.carstens@iea-dpc.de).

participants et de l'organisation qui commandite l'enquête) et précisée du point de vue de son potentiel analytique, y compris les variables, et de la précision requise des estimations. Naturellement, la qualité d'un produit dépend de la qualité des processus qui interviennent dans sa production. Concrètement, le choix des méthodes, la définition des procédures opérationnelles normalisées et leur suivi représentent des aspects essentiels de l'assurance que les processus peuvent générer un produit de données précis et comparable. Enfin, les organisations et équipes qui mettent en œuvre une enquête à l'échelle locale, ainsi qu'au niveau international, sont essentielles à la réussite et à la qualité globale d'une enquête.

1.2 Approche générale

Au cœur de la plupart des études à grande échelle en éducation se trouve une enquête transnationale par sondage sur les résultats obtenus dans un ou plusieurs domaines de contenu, qui est enrichie par la collecte de renseignements contextuels à divers niveaux (par exemple, système, école, enseignant, classes et parents), qu'on estime liés aux variations observées dans les résultats obtenus, par exemple, l'efficacité de certains types d'école, en reliant les intrants (antécédents) aux extrants. Dans ce cas, une étude réalisée à l'échelle de plusieurs pays peut produire des constatations impossibles à obtenir à partir de l'étude d'un seul pays et permettre de cerner les facteurs malléables que l'on pourrait manipuler, en vue d'améliorer les attitudes, les résultats ou l'efficience dans le système d'éducation.

Au plan opérationnel, la mise en œuvre d'une étude dans, par exemple, plus de 60 pays[2] représente une entreprise extrêmement complexe et exigeante du point de vue politique, financier et opérationnel (voir aussi Tamassia, 2005). À l'évidence, on prendra tous les moyens pour normaliser les opérations d'enquête dans tous les pays participants, en vue de réduire au minimum ou d'éliminer la variabilité des processus spéciaux, qui pourrait conduire et, de fait, conduit souvent à des erreurs qui sont difficiles, voire impossibles, à corriger. En matière de normalisation, l'approche générale qui est retenue dans nombre des enquêtes sur l'éducation, sinon la totalité, consiste donc à procéder à une harmonisation des intrants ou une harmonisation *ex ante*, c'est-à-dire une stratégie dans laquelle tous les pays utilisent les mêmes définitions et des processus uniformes, par opposition aux extrants ou à une harmonisation *a posteriori*, qui repose sur la recherche de dénominateurs communs concernant les données produites, ainsi que sur l'évaluation de la qualité et de la comparabilité de ces données, après collecte au moyen de méthodologies diverses. Par exemple, dans les études de l'IEA, les pays participants commencent leur travail de traduction à partir d'une, et parfois de deux versions de l'instrument source, et toutes les adaptations en fonction des questions et concepts sont documentées, revues et approuvées avant la mise en œuvre, ce qui fait que les adaptations structurelles peuvent raisonnablement être reliées à un schéma international et que les concepts sont traduits de façon appropriée. Pour un plan de sondage donné, il n'existe souvent pas de normes universellement valides ou acceptées pour certaines parties du processus d'enquête. Les coordonnateurs d'évaluations à grande échelle tentent plutôt d'utiliser ou d'adapter les meilleures méthodes et pratiques d'enquêtes courantes (voir Harkness et coll., 2010, De Leeuw, 2008, Statistique Canada, 2010, et Survey Research Center, 2010) et, le cas échéant, leurs versions historiques, c'est-à-dire les approches et méthodes employées dans les cycles antérieurs. Cependant, le degré de complexité des méthodes d'enquête actuelles (par exemple, l'échantillonnage ou les interviews) peut être remarquable et, habituellement, on sélectionne des méthodes pouvant procurer la qualité désirée et dont la mise en œuvre peut se faire de façon réaliste et fiable par toutes les équipes nationales.

L'avantage clé d'une normalisation initiale des méthodes et des opérations réside dans la prévisibilité des produits du processus, habituellement à l'intérieur d'un échéancier très serré, si bien qu'il n'y a tout simplement pas place à la variation en ce qui concerne, par exemple, les formats des fichiers de données fournis par les pays, puisque tout le temps et les budgets disponibles doivent être consacrés au travail de fond plutôt qu'au rapprochement ou qu'à l'harmonisation des particularités. Si l'on peut habituellement déterminer des approches optimales à l'échelle globale dans des domaines comme la sélection de l'échantillon ou la saisie et la vérification des données, la normalisation n'a pas à être stricte et prescriptive dans d'autres domaines. Cela pourrait même nuire aux objectifs de qualité. On pense par exemple à la participation d'organismes locaux compétents pour appuyer et promouvoir la recherche ou les particularités de l'enquête, dans les cas où les stratégies de prise de contact avec les écoles ou les répondants peuvent varier à l'intérieur de certaines limites, lorsqu'une solution optimale locale meilleure et défendable existe, qui ne nuit

[2]Par exemple, voir la liste des pays participants à la TEIEMS 2011 de l'IEA à l'adresse : http://timss.bc.edu/timss2011/countries.html.

pas à la solution globale. Cela découle souvent d'une philosophie dans laquelle on ne se soucie pas de la méthode suivie pour autant qu'elle fonctionne bien, et une normalisation légère ou d'orientation dans ce contexte signifie que les pays doivent mettre au point et documenter des plans concrets pour ces aspects.

Les coordonnateurs d'enquête s'efforcent de tenir compte des sources d'erreur, comme les erreurs de couverture et d'échantillonnage et celles dues à la non réponse, de la validité interculturelle des concepts, de l'absence d'invariance des mesures, des erreurs de collecte des données et des erreurs de traitement. De plus, ils tentent d'anticiper les nouvelles sources d'erreur, dans le cas de méthodes remaniées ou nouvellement utilisées (plus récemment avec l'avènement de la collecte de données assistée par ordinateur). Dans chaque domaine, l'objectif est de réduire au minimum la variation naturelle (aléatoire) et d'éliminer les variations systématiques de processus (et donc le biais ou la variance excessive; voir Biemer, 2010) qui surviennent à la suite d'un défaut accidentel – quoique parfois par négligence, voire de façon intentionnelle – de suivre une procédure spécifiée. De toute évidence, on pourrait consacrer une quantité presque infinie de temps et de ressources à la détermination et à l'étude des sources d'erreur. Cependant, dans la réalité, il faut faire des compromis entre les coûts et les erreurs, c'est-à-dire qu'il faut atteindre un juste équilibre entre la détermination et la réduction des erreurs et le coût que cela représente en temps et en argent. Dans cette optique, les planificateurs accordent une certaine priorité aux sources d'erreur et de biais qui comportent le plus de conséquences (comme la non réponse ou la collecte de données erronées) ou les plus notoires (comme la variation naturelle des notations humaines), sans pour autant laisser complètement de côté les autres aspects et en faisant reposer ces priorités sur l'expérience tirée des cycles précédents. Les coordonnateurs mettent ensuite au point des plans de suivi du contrôle de la qualité, qui ont pour objet la collecte d'information sur l'échantillon, du point de vue du degré d'uniformité de la mise en œuvre des processus (nous reviendrons sur cet aspect un peu plus loin).

1.3 Répartition et partage des responsabilités et de la charge de travail

Dans toute tentative d'élaboration de normes pour une étude transnationale, il faut tenir compte du fait que l'étude se déroule essentiellement à deux niveaux, national et international, et que cette réalité et la nature coopérative de l'entreprise doivent se refléter dans tout ensemble d'opérations. En théorie, toutes les tâches devraient être réparties au niveau international, en vue de maximiser les aspects communs et la normalisation, mais cette approche n'est pas opportune du point de vue de la qualité. Il n'est d'ailleurs pas non plus réaliste ni indiqué de décentraliser complètement ces tâches. Ce sont plutôt certains facteurs et contraintes qui permettent habituellement de déterminer la meilleure façon de partager les responsabilités et la charge de travail.

L'un des facteurs contraignants est la complexité de la tâche d'enquête par rapport à la capacité organisationnelle au niveau national. On observe de grandes variations d'un pays à l'autre en ce qui a trait aux traditions de réalisation d'enquêtes, à la capacité professionnelle, à l'expérience et aux ressources financières. Même si nombre d'organisations nationales de recherche ont acquis, au fil des ans, une impressionnante capacité de planifier et de mener des évaluations internationales (et nationales) – que l'on pense à la TEIEMS de l'IEA, amorcée en 1995, ou à l'enquête PISA de l'Organisation de coopération et de développement économiques (OCDE), qui a vu le jour en 2000 – on ne saurait de façon réaliste s'attendre à ce que tous les pays mettent en œuvre des opérations d'enquête d'un degré de complexité aussi élevé que celui des organismes d'enquête spécialisés. L'échantillonnage et la pondération sont un exemple d'opération complexe et donc principalement centralisée. La conception et la sélection d'échantillons probabilistes se font habituellement à l'échelle internationale, en se fondant sur les besoins analytiques nationaux (par exemple, la volonté de produire des données par domaine ou région), ce qui oblige les pays à compiler une base d'échantillonnage récente et complète, en fonction des spécifications définies au niveau international ou, au besoin, à procéder à une analyse du biais dû à la non réponse. Dans le cas de l'échantillonnage et de la pondération, il est habituellement à la fois plus efficace (que l'on pense, par exemple, à la documentation) et plus économique d'accomplir cette tâche au niveau international que de vérifier le travail effectué par les pays. Dans les évaluations réalisées au niveau des écoles, tous les échantillons sont habituellement constitués de façon centralisée. Dans d'autres cadres d'enquête cependant, des pays peuvent décider de sélectionner eux mêmes l'échantillon. Dans le cas de l'évaluation du PIAAC de l'OCDE, cela s'applique à environ la moitié des pays, et cette tâche est surtout effectuée par les bureaux de statistique nationale respectifs de ces pays. Les principales raisons de cela ont notamment trait à la confidentialité de l'information dans la base de sondage et à la complexité de certains plans de sondage stratifiés à plusieurs degrés.

Un autre facteur clé de la détermination du partage des responsabilités est celui de la connaissance du domaine et de l'accès au système (ou écosystème) local d'éducation. Quant à la traduction des instruments d'enquête, elle nécessite des connaissances et une expertise locales qui ne sont raisonnablement pas disponibles au niveau international, ce qui signifie qu'un processus de traduction entièrement centralisé n'est pas recommandable. Ainsi, les tâches de traduction sont partagées de telle manière que des équipes nationales sont responsables de veiller au caractère adéquat des traductions et adaptations (par exemple, en ce qui a trait aux définitions locales des niveaux de scolarité normalisés), afin de maximiser les correspondances avec les concepts et définitions stipulés par les cadres et les versions des instruments sources. La responsabilité de fournir des lignes directrices en matière de traduction, de revoir et d'approuver les adaptations, de faire une vérification linguistique des traductions et de procéder à une vérification optique de la présentation des questionnaires va alors au niveau international. Pour ce qui est de la collecte des données en tant que telle, les équipes nationales sont manifestement mieux à même de juger de la meilleure façon d'accéder aux unités échantillonnées et de prendre contact avec elles, d'obtenir l'appui des organismes locaux concernés, d'organiser la communication et de planifier l'administration des tests dans le respect d'un ensemble de lignes directrices et de limites applicables à tous les pays.

Enfin, la charge de travail et les coûts sont des facteurs contraignants, et les tâches doivent être réparties aux périodes de pointe de l'enquête, afin de limiter les dépenses au niveau international. Par exemple, la traduction d'un gros volume de documents (comme les tests, les questionnaires et les manuels), dans un délai habituellement court, doit être partagée non seulement pour des raisons de qualité, mais aussi parce qu'elle sollicite des efforts importants de la part de toutes les personnes qui y prennent part. Dans ce cas, le partage des responsabilités entre le niveau national et international fait aussi en sorte que la charge de travail est gérable d'un côté comme de l'autre. En revanche, la préparation et l'exécution du travail proprement sur le terrain et des tâches de traitement post collecte, comme la saisie des données, le codage des réponses selon la profession ou la notation des réponses fournies par les élèves, sont habituellement du seul ressort des équipes nationales, sous étroite supervision au niveau international.

2. Suivi de l'assurance et du contrôle de la qualité

Comme on l'a fait valoir ci dessus, la conception d'une enquête ou d'une évaluation doit être guidée par des objectifs de qualité. Au chapitre des opérations, l'assurance de la qualité (AQ) signifie que les organismes qui réalisent une enquête ont la capacité et l'expérience nécessaires pour le faire et que les processus définis et normalisés peuvent engendrer un produit de données qui répond aux besoins et attentes des intervenants. Dans le langage courant, l'AQ signifie que l'on « fait la bonne chose ». Étant donné que la mise en œuvre peut annuler une bonne conception, l'AQ est essentiellement axée sur les intrants et pourrait être considérée comme « fictive ». Le rôle du contrôle de la qualité (CQ) est de vérifier la « réalité », de produire des documents vérifiables et de mesurer la conformité, c'est-à-dire si les processus fonctionnent et si les produits sont bons dans les faits. Dans le langage de tous les jours, le CQ signifie que l'on « fait les choses de la bonne façon » (ou qu'on corrige la façon de faire, s'il y a lieu).

2.1 Approches et activités en matière d'assurance de la qualité

Comme on l'a décrit précédemment, l'assurance de la qualité vise essentiellement à faire en sorte que les exigences ou objectifs en matière de qualité du produit soient respectés. Concrètement, l'AQ représente les activités planifiées et systématiques qui sont menées à cette fin. Dans les enquêtes sur l'éducation, il s'agit de la documentation détaillée et exhaustive des procédures des opérations d'enquête que chaque pays doit compiler. Outre les aspects conceptuels qui procurent aux équipes nationales le cadre qui justifie une étude, un certain nombre de ressources et d'activités sont couramment utilisées pour transmettre l'information nécessaire au sujet de la mise en œuvre de l'étude. Les principales sont les suivantes :

- aperçus décrivant les rôles, responsabilités et compétences requises des coordonnateurs de recherche nationaux et d'autres employés clés ;
- manuels, lignes directrices et listes de vérification connexes clairs et simples pour toutes les opérations du chemin critique (par exemple manuels sur l'échantillonnage, lignes directrices sur la traduction et l'adaptation, manuels sur l'administration des tests et guides de l'intervieweur, manuels sur la gestion des données et guides de notation) ;
- logiciels, outils et services centralisés, lorsque le degré de complexité et le besoin d'uniformité le justifient ;
- normes techniques résumant les paramètres clés des produits et des processus (abordés un peu plus loin) ;

- réunions régulières en personne pour rendre compte des progrès et discuter des obstacles rencontrés, des leçons apprises, des aspects à améliorer et des tâches à venir (« communauté de pratique ») ;
- formation relative aux tâches clés, comme la collecte de données sur le terrain (stratégie de « formation du formateur »), ainsi que la saisie, le traitement ou l'analyse des données.

Ensemble, ces approches fournissent aux coordonnateurs nationaux une information exhaustive, ciblée et suffisamment détaillée pour assurer la mise en œuvre de l'enquête à l'échelle locale, et permettent de regrouper des équipes nationales possédant l'expertise et la capacité requises, de trouver du soutien à l'interne ou des sous traitants pour des tâches précises (comme la traduction, la collecte des données ou le codage par profession), de planifier des stratégies de communication avec les répondants, de mettre en œuvre la collecte des données et de préparer les fichiers de données requis. En résumé, les coordonnateurs reçoivent tous les renseignements, les documents et l'aide nécessaires pour réaliser l'étude sans avoir besoin d'élaborer leurs propres documents et outils pour des aspects particuliers. Par exemple, l'IEA fournit à chaque pays participant un exemplaire de son logiciel d'échantillonnage dans les écoles, afin de faciliter le dénombrement des classes, des élèves et des enseignants (selon le plan de sondage), de signaler les exclusions et d'entrer les données auxiliaires pertinentes d'échantillonnage, de tirer des échantillons selon le plan de sondage international, d'attribuer des codes d'identification hiérarchiques, de produire des listes de suivi et des étiquettes de questionnaire et de consigner la participation des répondants dont rendent compte les administrateurs des tests et les instruments reçus.

Pour certains aspects – dont les stratégies de prise de contact mentionnées précédemment – les coordonnateurs nationaux doivent élaborer un plan qui s'applique à leur contexte local en adaptant et en élargissant parfois les recommandations génériques. Il devient alors nécessaire de recueillir de façon systématique des renseignements sur la façon dont les coordonnateurs nationaux prévoient mettre en œuvre les lignes directrices, par exemple, en ce qui concerne le nombre d'intervieweurs (et de superviseurs), compte tenu du nombre visé de cas achevés et d'une période fixe de collecte de données. À cet égard, le consortium du PIAAC de l'OCDE s'est servi d'un « rapport sur la planification et la conception d'une enquête nationale » pour recueillir systématiquement de l'information sur les aspects clés de mise en œuvre, bien avant le travail sur le terrain, ainsi que pour déterminer si les pays sont susceptibles de mettre en œuvre l'étude de la bonne façon ou s'ils ont besoin d'un soutien supplémentaire.

Cela dit, les organisations qui coordonnent les évaluations internationales à grande échelle (c'est du moins le cas de l'IEA) n'« évaluent » pas la capacité d'un pays à mettre en œuvre une étude avant de leur permettre de le faire. Elles sont plutôt réceptives initialement à la volonté d'un pays de participer. Le test ultime du fonctionnement général des approches d'AQ et de la conformité locale aux opérations normalisées prend la forme, dans la plupart sinon la totalité des études transnationales, d'un essai sur le terrain, c'est-à-dire d'une répétition générale qui sert plusieurs buts, soit i) tester, valider et préciser les instruments d'enquête et/ou leurs traductions, ii) mettre à l'essai et réviser les opérations d'enquête et les approches d'AQ en général et iii) fournir aux coordonnateurs nationaux et internationaux de l'information sur la capacité de chaque pays de réaliser le travail. Au besoin, les coordonnateurs internationaux analysent les problèmes en profondeur et fournissent un soutien supplémentaire aux pays.

2.2 Approches et programmes de contrôle de la qualité

Le contrôle de la qualité consiste à produire et à recueillir systématiquement des données et des éléments probants indiquant que les opérations d'enquête ont été mises en œuvre dans le respect des normes et des plans et que les données recueillies conviennent à l'utilisation qu'on veut en faire. Les efforts de contrôle de la qualité doivent englober tout le processus d'enquête, peu importe si la responsabilité en était confiée aux coordonnateurs internationaux ou nationaux. Idéalement, les efforts de CQ ne devraient pas se heurter à des « boîtes noires », c'est-à-dire des processus non vérifiables, et devraient être déployés non seulement par des personnes proches de celles qui ont mis en œuvre le processus, mais aussi par des contrôleurs indépendants. Habituellement, les programmes de CQ sont centrés sur les processus qui sont du ressort des pays (par exemple, la traduction, le travail sur le terrain ou la saisie des données). Néanmoins, le travail effectué par les coordonnateurs internationaux est également documenté en détail, de façon à pouvoir être revu et contre vérifié par des coordonnateurs ou conseillers techniques nationaux.

Le travail de CQ vise à quantifier la somme d'erreurs et, lorsque cela est pertinent et possible, utilise même des méthodes de contrôle des processus statistiques. Cela est relativement aisé pour ce qui est, par exemple, du contrôle de la production de la collecte, des taux de réponse prévus, de la fiabilité de la notation ou de l'exactitude des données saisies en double. Dans la majorité des cas, cependant, il est impossible de couvrir chacune des actions, et

les programmes de CQ sont habituellement limités à des échantillons du travail. Par exemple, dans ses études, l'IEA met ordinairement en œuvre un programme ambitieux de contrôle international de la qualité, afin de documenter les activités de collecte de données, et désigne un superviseur du contrôle de la qualité (SCQ) dans chacun des pays participants. Après avoir reçu une formation exhaustive, cette personne supervise notamment l'administration des tests dans environ 10 % des écoles (soit 15 des 150 écoles selon le plan de sondage canonique). Dans l'évaluation de la TEIEMS de 2007, 248 SCQ et leurs adjoints ont assuré le contrôle de 1 371 séances d'administration de tests au total (Olson, Martin et Mullis, 2008). Ces programmes permettent de produire des preuves importantes du degré d'uniformité de la collecte des données et sont habituellement complétés par des programmes nationaux similaires, qui s'inspirent du modèle international. Il arrive parfois, cependant, qu'il soit impossible de recueillir l'information directement et de façon indépendante. Dans le cas de l'évaluation du PIAAC de l'OCDE, le déploiement de superviseurs internationaux indépendants pour vérifier le travail des intervieweurs pendant et/ou après la collecte s'est avéré impossible en raison de la confidentialité des données de contact. En pareil cas, le contrôle de la qualité est surtout assuré par les organisations qui sont aussi responsables de mener les interviews.

Pour toutes les activités d'enquête, il est expressément demandé aux coordonnateurs nationaux de rendre systématiquement compte de leur conformité et de leurs expériences (ou de celles des répondants) à l'égard d'opérations prédéterminées dans les rapports du travail sur le terrain, des questionnaires sur les activités d'enquête, des listes de vérification et/ou des appels téléphoniques, autant d'activités apparemment différentes qui servent cependant un but commun. Là encore, le travail sur le terrain et les éléments probants du contrôle de la qualité qui sont recueillis pendant son déroulement servent à vérifier si un pays est susceptible de s'acquitter avec succès de la collecte principale des données ou s'il convient d'apporter des améliorations substantielles, de fournir du soutien ou même de réévaluer la participation du pays. À l'issue de la collecte principale des données, tous les éléments probants disponibles sur le contrôle de la qualité (c'est-à-dire les taux de réponse obtenus, la fiabilité de la notation et les rapports produits par les superviseurs du contrôle de la qualité) sont compilés et utilisés pour évaluer les échantillons nationaux et, par la suite, formuler une recommandation sur l'inclusion sans condition, l'inclusion conditionnelle ou la non inclusion des données dans l'analyse, l'échelonnage et les rapports internationaux.

2.3 Normes techniques

Toutes les études transnationales à grande échelle sur l'éducation définissent des normes techniques, c'est-à-dire un ensemble d'exigences que doit respecter un produit ou un processus. Par exemple, les normes techniques définissent les principaux aspects des responsabilités, de la gestion, des plans de sondage, du processus de traduction, des procédures de collecte de données, du contenu et des formats des fichiers de données, du codage/de la notation ou des taux minimums de réponse au niveau national. Ainsi, les normes techniques peuvent être considérées comme un aperçu des aspects clés de l'assurance et du contrôle de la qualité et englobent les attentes communiquées et convenues d'une enquête en matière de qualité.

Les modes de documentation et de diffusion des normes techniques sont très variés. Dans les études de l'IEA, où les intervenants nationaux et les coordonnateurs nationaux de la recherche sont les mêmes, des guides des opérations, comme un manuel sur l'échantillonnage ou sur la gestion des données, suffisent pour communiquer les normes, les attentes en matière de qualité, les processus obligatoires et les points de contrôle, ainsi que des instructions détaillées. Cette approche a pour avantage qu'elle permet d'éviter que la documentation soit redondante. Dans les enquêtes et évaluations de l'OCDE, les normes techniques sont habituellement élaborées dans le cadre d'un document spécialisé, qui renferme des normes pour l'ensemble des domaines, normes qui sont davantage détaillées dans les manuels des opérations. Cette façon de faire est probablement liée à la façon dont les projets de l'OCDE sont gérés, en l'occurrence i) par un comité formé de membres des pays participants, qui représentent le principal organe décisionnel du projet, et qui ne se préoccupent pas tant de la documentation opérationnelle que de spécifications en matière de qualité de haut niveau du point de vue de l'utilisateur, et ii) par un groupe de gestionnaires de projets nationaux qui réalisent le travail au niveau local et ont donc besoin d'information sur les normes, ainsi que d'instructions détaillées. Un autre avantage d'un document de normes techniques, ainsi que de documents sur l'échéancier et les jalons, est que l'on peut s'en servir comme intrants importants dans le processus national d'appel d'offres auprès des organismes d'enquête. Il reste que le niveau de détail des normes techniques peut varier de façon considérable. Le consortium du PIAAC de l'OCDE a produit un ensemble de normes techniques (non encore publié), qui compte presque 200 pages, surtout en raison de la complexité méthodologique de l'étude et de son lien avec le modèle de normes des précédents travaux de l'EIAA et de l'EIACA. S'il est vrai que cela peut réduire la probabilité que ce document soit effectivement lu, les quelque 20 pages de l'évaluation du PISA 2009 (OCDE 2011) semblent

relativement maigres en comparaison. La principale différence entre les documents vient de ce que celui du PIAAC renferme, outre les normes en tant que telles, une grande quantité de renseignements d'appui dans les sections de chaque domaine consacrées à la justification, aux lignes directrices, aux recommandations et à l'assurance/au contrôle de la qualité. Pour l'Enquête internationale sur les enseignants, l'enseignement et l'apprentissage (TALIS) de l'OCDE, on a opté pour un certain compromis, le document sur les normes (pas encore publié) dans ce cas comportant une quarantaine de pages et laissant, comme celui du PISA, aux manuels opérationnels les détails techniques.

À l'évidence, cependant, les principales difficultés liées à la production de normes techniques n'ont pas trait au niveau de détail. Il importe plutôt d'élaborer un ensemble exhaustif de normes et d'indiquer les aspects pour lesquels elles seront appliquées à la lettre (par exemple, en ce qui concerne les exigences de taux de réponse ou la vérification admissible des données), ainsi que ceux pour lesquels une certaine marge de manœuvre pourrait être accordée et les pays pourraient demander de s'écarter des normes s'ils préféraient retenir une approche différente ou modifiée pour certains aspects des travaux, sous réserve de l'approbation par les coordonnateurs internationaux (par exemple, le recours au balayage et à la reconnaissance optique des caractères plutôt qu'à la saisie manuelle des données). Une autre considération importante a trait aux normes déjà en place, qui sont utilisées par un nombre croissant d'organismes d'enquête et par un grand nombre, sinon la totalité, des bureaux de statistique nationale. Dans ces cas, beaucoup de temps est habituellement consacré à l'examen des dérogations proposées, pour faire en sorte que les normes coïncident avec les pratiques locales et, parfois même, au rejet des demandes de dérogation.

3. Exemples : L'EIMOI de l'IEA et le PIAAC de l'OCDE

Le tableau 3.1 ci après présente les principaux aspects de la conception de deux évaluations internationales courantes à grande échelle, soit l'Étude internationale sur la maîtrise de l'ordinateur et de l'information (EIMOI) de l'IEA et le Programme pour l'évaluation internationale des compétences des adultes (PIAAC) de l'OCDE, afin d'illustrer la justification précédemment décrite de la normalisation des opérations pour un plan de sondage donné. Il ne s'agit pas vraiment d'une comparaison valide, les deux initiatives comportant des buts très différents, mais un degré élevé de similitude, les deux servant à produire des données internationales faisant autorité aux fins de la comparaison et de l'étalonnage. De plus, les deux études sont intéressantes, car elles ont invariablement recours à l'administration de tests assistée par ordinateur, ainsi qu'à des instruments contextuels comportant un lien naturel avec les domaines et concepts étudiés. Même si l'EIMOI étend le modèle éprouvé des études de l'IEA à l'administration assistée par ordinateur, le PIAAC représente sans doute l'enquête transnationale sur l'éducation et les compétences des adultes la plus ambitieuse, complexe et coûteuse entreprise à ce jour, étant donné qu'elle combine une enquête auprès des ménages, des évaluations en éducation et des méthodes d'administration assistée par ordinateur.

Comme on peut le constater à partir du tableau, les deux études diffèrent sensiblement quant aux organisations chargées de la coordination internationale et de la mise en œuvre au niveau national, à la façon dont les échantillons sont sélectionnés et aux responsables de la sélection, aux procédures de collecte des données (administration de tests en classe et interviews sur place), aux programmes correspondants de contrôle de la qualité et à la disponibilité directe (observateurs/superviseurs internationaux) ou indirecte (vérification au téléphone par des superviseurs) des éléments probants du CQ et, enfin, sur le plan des principaux défis auxquels chacune fait face. Cependant, les deux projets partagent aussi certaines approches et pratiques exemplaires, même si les approches, outils et systèmes particuliers varient. Cela s'applique aux modes faisant appel à divers instruments, au fait que les deux études fournissent aux pays les logiciels dont ils ont besoin pour assurer la collecte des données, aux processus de traduction et d'adaptation, ainsi qu'aux processus post collecte, comme la saisie des données consignées sur papier et la notation et le codage des réponses aux questions ouvertes pour lesquels les règles et les buts sont semblables.

Tableau 3-1 Aspects clés du plan de sondage de l'EIMOI de l'IEA et du PIAAC de l'OCDE

	Étude internationale sur la maîtrise de l'ordinateur et de l'information (EIMOI) de l'IEA	**Programme pour l'évaluation internationale des compétences des adultes (PIAAC) de l'OCDE**
Domaines	Maîtrise de l'ordinateur et de l'information	Littératie, numératie et capacités de résolution de problèmes
Cibles	Élèves de 8e année et leurs enseignants	Adultes de 16 à 65 ans
Coordination	Gestion conjointe par trois partenaires : ACER (gestion principale), Administration centrale de l'EA et Centre de traitement des données et de recherche de l'IEA	Dirigé par ETS et sept partenaires : Westat, IEA, cApStAn, ROA, DIPF, CRP et GESIS
Participants	Environ 20 pays	25 pays pour le cycle 1, environ huit pour le cycle 2
Échéancier	2010-2014, essais sur le terrain en 2012, collecte principale en 2013	2008-2013, essais sur le terrain en 2010, collecte principale de la fin de 2011 au début de 2012
Centres nationaux	Surtout les universités, ministères et instituts de recherche associés en éducation	Surtout des bureaux de statistique nationale soutenus par des organismes d'enquête privés et quelques instituts de recherche en éducation
Échantillons	Échantillonnage à deux degrés avec PPT de 150 écoles (20 élèves et 15 enseignants par école) ; sélection centralisée	Grande variation, allant des échantillons sur registre aux échantillons à plusieurs degrés avec filtre des ménages; sélection/pondération centralisée pour environ la moitié, l'autre moitié étant prise en charge par le pays
Instruments	Administration de tests assistée par ordinateur et de questionnaires électroniques aux élèves; questionnaires aux enseignants et aux écoles (sur papier ou en ligne)	Test adaptatif assisté par ordinateur, papier sur demande, questionnaire de contexte donné par Interview sur place assistée par ordinateur (IPAO)
Traduction	Adaptation/traduction, revue par un expert, révision, vérifications de la mise en page (niveau national) Vérification, vérifications de la mise en page, documentation (niveau international)	Adaptation/traduction (en double recommandée), rapprochement, vérification de la mise en page (niveau national) Vérification, vérification de la mise en page, documentation (niveau international)
Systèmes	Traduction, échantillonnage dans les écoles, évaluation des élèves (USB), collecte et saisie des données en ligne (centralisés)	Traduction, IPAO et test assisté par ordinateur (portable) et systèmes de saisie (centralisés) Système de gestion de l'étude/des cas (responsabilité locale)
Collecte des données	Séances chronométrées et surveillées, tous les questionnaires auto-administrés (en ligne, par défaut, papier sur demande)	Questionnaire de contexte par IPAO, portion cognitive assistée par ordinateur non chronométrée ou sur papier (contrôlée par l'intervieweur)
Tâches post-collecte	Codage de la profession des parents, notation des réponses, saisie des données des documents sur papier (responsabilité locale et études de fiabilité internationales)	Codage (profession, industrie, langue, pays, région), notation des réponses sur papier, saisie des données sur papier (responsabilité locale et études de fiabilité internationales)
Programme de contrôle de la qualité	Vérificateurs nationaux de la qualité; observateurs internationaux de l'administration des tests dans 10 % des écoles (mesure directe)	Appels de contrôle de la qualité, vérification de 10 % du travail de chaque intervieweur (100 % en cas de doute) par les centres nationaux, rapports au consortium (mesure indirecte)
Principales difficultés	Participation des écoles, infrastructure informatique dans les écoles, conditions uniformes d'administration du test, compte tenu des systèmes informatiques et des compétences des administrateurs du test	Conditions d'administration du test et d'interview uniformes, système complexe d'IPAO et d'administration assistée par ordinateur (également mise en place et correction), contact initial, taux de réponse, échéancier global

4. « Normaliser les normes »

On a souligné précédemment qu'il était essentiel de normaliser les définitions et les opérations entre les pays qui prennent part à une même enquête. On peut en dire autant des études à plusieurs cycles (par exemple, les programmes indicateurs de tendances, comme la TEIEMS/le PIRLS et le PISA, ainsi que des études réalisées par une même organisation (comme l'IEA ou l'OCDE). Les évaluations récentes à grande échelle ont en commun un nombre relativement important de caractéristiques, d'approches et de procédures au niveau opérationnel, mais leurs populations, concepts et plans de sondage peuvent varier dans une large mesure. Il s'agit évidemment d'un développement positif, mais tout ne fonctionne pas partout.

Au niveau national, la certification ISO 20252 (Organisation internationale de normalisation, 2006) commence à se généraliser et se rapporte à la normalisation des opérations liées à la planification et à la tenue des enquêtes dans les organismes d'enquête. On peut aussi observer une normalisation des pratiques d'enquête parmi les organismes de statistique nationale (par exemple, Statistique Canada, 2009) ou d'autres organismes gouvernementaux (par exemple, National Center for Education Statistics, 1991), même si la nature de ces documents varie, allant des recommandations ou lignes directrices à des instructions assez rigides laissant relativement peu de place aux dérogations ou pas du tout. Comme Susan Linacre l'a fait valoir dans son discours-programme, « s'il vaut la peine de normaliser une pratique, il vaut aussi la peine de la mettre en application ». On a appliqué cela au contexte australien et à la difficulté de normaliser, par exemple, les opérations de vérification et d'imputation entre les divers secteurs opérationnels de l'ABS. En plus des gains probables de qualité et d'uniformité qu'elle permettrait – pour autant que la solution optimale globale corresponde aussi à la solution optimale locale pour chaque enquête – la normalisation entre les enquêtes pourrait aussi faire économiser des ressources à moyen ou à long terme, à mesure que des systèmes généralisés seront créés, que des ressources seront mises en commun et que les nouveaux projets (avec un peu de chance) nécessiteront très peu d'adaptation. Souvent, toutefois, la normalisation engendre d'abord des coûts supplémentaires.

Dans le cas des enquêtes transnationales, l'atteinte d'un équilibre entre la solution optimale globale et la solution optimale locale a fait l'objet de travaux très utiles conçus et publiés par le Survey Research Center (2010) de l'University of Michigan. Le document « Cross-cultural Survey Guidelines » (CCSG) s'applique à la recherche sociale en général, mais a principalement trait aux enquêtes transversales auprès des ménages et des particuliers. Ces lignes directrices recensent les pratiques exemplaires, comme le travail d'équipe en traduction, et en font la promotion; elles comprennent de nombreux renvois aux études et travaux publiés, des exemples et un glossaire bien fait. Elles sont formulées en termes généraux et ne revêtent pas de caractère prescriptif, pas plus qu'elles n'imposent de pratiques particulières, ce qui signifie que pour chaque enquête comportant des objectifs et un plan de sondage qui lui sont propres, il faut définir des approches, procédures et opérations particulières. En ce sens, les CCSG ne sont pas des normes techniques « strictes » à proprement parler, mais constituent plutôt un cadre de méta normes, qui doit servir à l'élaboration de normes par les coordonnateurs internationaux.

La normalisation et l'harmonisation des études transnationales similaires ou des cycles d'une même étude au fil du temps continuent toutefois de présenter un défi. Dans le cas des actuels programmes de l'OCDE (PISA, TALIS, PIAAC, AHELO), les projets sont gérés par différents services du Secrétariat de l'OCDE et sont habituellement mis en œuvre par différents sous traitants internationaux, habituellement un consortium d'organismes d'enquête œuvrant au niveau international. Dans ses projets, du moins ceux entrepris au niveau des écoles, l'OCDE s'efforce de parvenir à une certaine uniformité conceptuelle, à la réutilisation des questions ou à l'harmonisation avec les approches mixtes de l'Organisation des Nations Unies pour l'éducation, la science et la culture, de l'OCDE et de l'Union européenne en matière de collecte de données. Néanmoins, les normes techniques et les opérations sont définies de façon indépendante pour chaque projet et cycle, quoiqu'il existe un degré assez élevé de similitude lorsque l'on retient les services d'un même sous traitant pour des cycles consécutifs d'une enquête et lorsque les sous traitants sont de plus en plus nombreux à adopter les pratiques exemplaires courantes en matière d'enquêtes. Il arrive aussi que des pays sollicitent un alignement opérationnel entre les divers projets de l'OCDE auxquels ils participent (« Pouvons nous faire cela comme dans le PISA? »).

Dans les évaluations transnationales de l'IEA, nombre d'aspects du processus d'enquête sont assez normalisés et suivent un modèle commun (par exemple, pour ce qui est de l'échantillonnage, de la traduction, de la saisie des données et de leur traitement), tandis que d'autres ne reprennent pas ou ne peuvent pas utiliser le plan de sondage particulier de l'étude. Tous les projets de l'IEA suivent les lignes directrices énoncées dans les normes techniques de

l'organisation (Martin, Rust et Adams, 1999, et aussi Gregory et Martin, 2001), qui sont adaptées au contexte paramètre particulier dans lequel l'IEA évolue, c'est-à-dire qu'elles sont clairement axées sur les évaluations en éducation. Comme c'est le cas pour les CCSG, ces lignes directrices ne comportent pas de prescriptions particulières, comme un taux de réponse fixe minimum, pour toutes les études. Ces « méta normes » servent plutôt de cadre imposant aux coordonnateurs internationaux de se pencher sur chaque partie du processus d'enquête – de la mise sur pied d'un centre international d'études à la diffusion des données et des documents techniques – en vue de définir des normes propres à l'étude. Cela pourrait aboutir à des plans de sondage très différents, les études parrainées par l'IEA (actuellement, la TEIEMS, le PIRLS, la TEDS, l'ICCS et l'EIMOI) étant aussi menées par différentes organisations et personnes. Cela dit, on observe un degré élevé de ressemblance, les projets s'inspirant les uns des autres et partageant des approches et pratiques éprouvées. Qui plus est, nombre des tâches liées aux enquêtes sont non seulement normalisées, mais aussi centralisées d'une étude à l'autre, et cela s'applique, à divers degrés, à l'échantillonnage, à la traduction/vérification, aux opérations, au contrôle de la qualité de la collecte des données, à la gestion des données, aux analyses (secondaires) et à la diffusion des données. En outre, un seul groupe technique supervise les plans, les progrès et la qualité de toutes les études. Collectivement, ces structures engendrent une grande communauté entre les études, pour nombre des normes, processus et activités de contrôle de la qualité, mais pas la totalité.

Bibliographie

Biemer, P.P. (2010), « Total survey error: Design, implementation, and evaluation », *Public Opinion Quarterly*, vol. 74, no 5, p. 817 à 848.

Biemer, P.P. et L.E. Lyberg (2003), *Introduction to Survey Quality*, Hoboken, NJ: John Wiley & Sons.

Brackstone, G. (1999), « La gestion de la qualité des données dans un bureau de statistique », *Techniques d'enquête*, vol. 25, no 2, p. 159 à 171.

De Leeuw, E.D., Joop J. Hox et D.A. Dillman (2008), *International Handbook of Survey Methodology* (éditeurs), New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Gregory, K.D. et M.O. Martin (2001), *Technical Standards for IEA Studies: An Annotated Bibliography*, Amsterdam: IEA.

Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell. B.-E. et T.W. Smith (2010), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (éditeurs), Hoboken, NJ: John Wiley & Sons.

Organisation internationale de normalisation, (2006), ISO 20252, *Études sociale, d'opinion et de marché – termes, définitions et exigencies de service*, Genèva, Suisse: ISO.

Lyberg, L.E. et D.M. Stukel (2010), « Quality assurance and quality control in cross-national comparative studies », dans Harkness, J.A. et coll. (éditeurs). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Hoboken, NJ: John Wiley & Sons, p. 227 à 249.

Martin, M.O., Rust, K. et R.J. Adams (1999), *Technical Standards for IEA Studies*, Amsterdam: IEA.

National Center for Education Statistics (1991), *SEDCAR Standards for Education Data Collection and Reporting*, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

OCDE. (2011), *PISA 2009, Normes techniques*, annexe 7 du PISA 2009 – Rapport technique (version préliminaire), Paris: OCDE. Consulté le 15 janvier 2012 à http://www.oecd.org/document/19/0,3746,en_2649_35845621_48577747_1_1_1_1,00.html.

Olson, J.F., Martin, M.O. et I.V.S. Mullis (2008), *TIMSS 2007 – Technical Report*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Statistique Canada (2009), *Lignes directrices concernant la qualité*, 5e édition, no 12-539-X au catalogue de Statistique Canada, Ottawa: Statistique Canada.

Statistique Canada (2010), *Méthodes et pratiques d'enquête*, no 12-578-X au catalogue de Statistique Canada, Ottawa: Statistique Canada.

Survey Research Center (2010), *Guidelines for Best Practice in Cross-Cultural Surveys*, Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Consulté le 6 janvier 2012 à http://www.ccsg.isr.umich.edu/

Tamassia, C. (2005), « Implementing Surveys In An International Context: An Overview », Article présenté à la réunion du groupe d'experts en évaluation de la qualité dans les services d'éducation du Programme on Educational Building (PEB) de l'OCDE, Lisbonne, Portugal.

Résumé des réponses d'item dans l'évaluation à grande échelle

Eugenio Gonzalez et Matthias Von Davier[1]

Résumé

Les évaluations à grande échelle et d'autres programmes d'enquête comportent généralement l'administration de multiples questions aux participants aux enquêtes. Ces questions peuvent être de nature différente ou visent à mesurer un concept commun sous des angles complémentaires ou différents. Lorsque plusieurs questions sont administrées pour mesurer un concept commun, il est généralement utile de résumer les réponses en une variable ou un indice unique. La combinaison des réponses à ces items en un nombre unique porte le nom d'échelonnage. Cette combinaison des questions peut se faire de plusieurs façons, allant de la simple somme de données ponctuelles sur les items à l'utilisation de méthodes plus complexes, telles que la théorie de la réponse d'item (TRI) et les tirages multiples à partir d'une distribution a posteriori prévue des résultats possibles. La communication présentera les avantages et les inconvénients des diverses méthodes d'échelonnage dans le contexte de l'évaluation à grande échelle dans le domaine de l'éducation. Les méthodes particulières présentées et qui feront l'objet d'une discussion comprennent les scores sommatifs, les scores moyens, les scores en pourcentage, les scores factoriels, les scores TRI et les valeurs plausibles.

[1]Eugenio Gonzalez et Matthias Von Davier, Educational Testing Service, États-Unis.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Standardisation des plans de sondage et assurance de qualité dans les enquêtes comparatives

Marc Joncas et Sylvie LaRoche[1]

Résumé

L'intérêt grandissant pour les études comparatives à caractère international dans le domaine de l'éducation pose de nombreux défis de standardisation des méthodes statistiques, en particulier celle des plans de sondage. Dans ce document, nous passons en revue une à une les différentes étapes définissant un plan de sondage normalisé pour des études comparatives de ce type : la définition des populations cibles et d'enquête ; la construction des bases de sondage ; le choix du mode de sélection des échantillons ; la détermination des tailles d'échantillons et précision des estimations ; et, l'évaluation de la mise en œuvre. Nous terminerons en discutant des leçons apprises au cours des nombreuses années de participation à différentes enquêtes internationales.

Mots clés : Standardisation ; plan de sondage ; études comparatives internationales ; éducation.

1. Introduction

Les études comparatives à caractère international dans le domaine de l'éducation existent depuis de nombreuses années. Ces enquêtes visent à mesurer l'efficacité des systèmes éducatifs dans leur ensemble et s'adressent généralement aux élèves et/ou aux enseignants. Plusieurs de ces enquêtes évaluent les compétences acquises par les élèves et permettent ainsi de comparer les performances des systèmes d'éducation en place dans les pays participants. Parmi les plus connues, mentionnons le Programme international pour le suivi des acquis des élèves (PISA[2]), le Programme international de recherche en lecture scolaire (PIRLS[3]) et les Tendances de l'enquête internationale sur la mathématique et les sciences (TIMSS[4]). Depuis 2008, l'Organisation de coopération et de développement économiques (OCDE) a aussi mise sur pied l'enquête internationale sur les enseignants, l'enseignement et l'apprentissage (TALIS) qui s'intéresse de son côté aux conditions de travail des enseignants et à leur environnement pédagogique.

Comme la plupart de ses études établissent un classement des systèmes d'éducation des pays participants basé sur la performance des élèves, leurs résultats sont hautement médiatisés et souvent sensible d'un point de vue politique. De par leur nature comparative, les études internationales demeurent donc sujettes à la controverse. Ainsi, les personnes responsables de ces études font face à de nombreux défis afin de s'assurer que tous les aspects des enquêtes soient contrôlés et vérifiés de façon à ce que les résultats en découlant soient comparables et crédibles.

L'échantillonnage est certainement une composante importante à cet égard. La méthodologie d'échantillonnage d'enquêtes antérieures a souvent été l'objet de critiques. Nous pouvons certainement nous poser les questions suivantes : Est-ce l'échantillon d'un pays est représentatif de la population cible? Est-ce que les niveaux d'exclusions sont comparables d'un pays à l'autre? Les échantillons ont-ils été sélectionnés de façon appropriée? Est-ce que les

[1]Marc Joncas, Statistique Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (marc.joncas@statcan.gc.ca) ; Sylvie LaRoche, Statistique Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (sylvie.laroche@statcan.gc.ca).

[2]*Programme for International Student Assessment* menée par l'Organisation de coopération et de développement économiques (OCDE).

[3]*Comparison of the Progress in International Literacy Study*. Cette étude est réalisée par le TIMSS&PIRLS International Study Center et financée par Association internationale pour l'évaluation du rendement scolaire (IEA - *International Association for the Evaluation of Educational Achievement*).

[4] *Trends in International Mathematics and Science Study*. Cette étude est également réalisée par ISC et financée et financée par IEA.

taux de participation sont acceptables? Ces questions et bien d'autres ont donné naissance à la nécessité de se doter d'une série de contrôles et de standards propres aux études comparatives dans le domaine de l'éducation, lesquels font l'objet d'une revue dans les sections subséquentes. Notez que cette revue se limitera aux aspects touchant particulièrement le plan de sondage. De manière plus précise, nous décrivons dans un premier temps les conditions définissant un contexte idéal menant à un plan de sondage adéquat, pour naturellement aborder les difficultés inhérentes à l'implantation d'un tel plan sur le terrain. Cette section est suivie d'une courte discussion sur les raisons qui nous amènent à devoir standardiser dans un contexte international. S'ensuivent une série d'exemples de contrôles et de standards parmi les plus couramment utilisées dans ce genre d'enquêtes avant de finalement conclure en mentionnant les leçons retenues au cours des nombreuses années d'expérience à travailler sur les enquêtes comparatives internationales en éducation.

2. Contextes

2.1 Contexte idéal

Dans un contexte idéal, un plan de sondage devrait mener à des résultats sans biais, exactes et comparables internationalement. Pour ce faire, trois conditions apparaissent incontournables lors de la conception d'enquêtes internationales.

D'abord, la population d'enquête doit correspondre à la population cible. Cette condition est primordiale étant donné que les inférences se rapportent à la population d'enquête. Une fois la collecte terminée, il n'est pas toujours possible des faire des ajustements lors de l'estimation afin de corriger les défauts de couverture. Il est souvent plus facile de faire comprendre l'importance de couvrir entièrement la population d'intérêt lorsqu'il s'agit d'un recensement. Que vaudrait un recensement avec une couverture de seulement 80 %?

Dans le cas d'une enquête-échantillon par contre, on a souvent tendance à penser que cette condition n'est pas aussi capitale alors qu'elle l'est tout autant. Dans un contexte international, ceci est encore plus important puisque que c'est le premier aspect pouvant miner la crédibilité de l'enquête. Des taux de couverture différents d'un pays à l'autre ouvrent irrémédiablement la porte à des discussions remettant en cause la validité des comparaisons et peuvent ainsi limiter les possibilités d'analyse. Dans le contexte des enquêtes touchant l'éducation, cet élément est particulièrement important puisqu'il peut laisser l'impression que les résultats seront biaisés si la population non couverte d'un pays incluent les étudiants les moins performants.

Ensuite, un plan d'échantillonnage valide est la deuxième condition essentielle. Il constitue notre lien avec la population d'enquête. Toutes les unités de la population d'enquête se doivent d'avoir une chance (probabilité d'inclusion non nulle) d'être choisies sous peine de voir le niveau d'exclusions s'accroître (et ainsi retomber avec une couverture non complète). Cette probabilité d'inclusion doit être connue et calculable afin d'exclure un risque de biais. Il est important aussi d'assurer une mise en œuvre adéquate de ce plan d'échantillonnage (car il ne sert à rien de faire de beaux plans d'échantillonnage si ceux-ci ne sont pas exécutés correctement sur le terrain). À défaut de disposer d'un plan d'échantillonnage valide, il devient hasardeux d'inférer les résultats de l'enquête à la population d'enquête.

Finalement, comme troisième condition, les erreurs d'échantillonnage devraient être les plus petites possibles. Les résultats d'enquête seraient donc le plus près possibles des valeurs qui auraient été obtenues d'un recensement.

Dans un contexte idéal où toutes ces conditions seraient remplies, avec une mise en œuvre parfaite du plan d'échantillonnage, des procédures de collecte, et des taux de réponse de 100%, le besoin d'établir des standards tels que définis à la section 4.1 n'existerait pas.

2.2 Contexte international (sur le terrain)

Dans un projet d'enquête regroupant plusieurs pays, il est rare de retrouver un contexte idéal dans lequel les conditions discutées à la section précédente sont pleinement respectées. À peu près tout distingue un pays d'un autre, et le monde de l'éducation n'y fait pas exception. Que ce soit en rapport aux systèmes d'éducation comme tels, aux

particularités géographiques ou culturelles, à la disponibilité et à l'accès aux données administratives pertinentes, à la capacité de mise-en-œuvre d'enquêtes, au fardeau de réponse, ou encore à la culture d'enquêtes[5] pour n'en nommer que quelques-uns, il est impossible de trouver deux pays parfaitement comparables. Invariablement, les taux de couverture, de réponse, la qualité de la mise-en-œuvre, bref tout ce qui touche au plan de sondage sera affecté à des degrés différents d'un pays à l'autre. Ce constat n'implique pas que toutes tentatives de comparaisons entre les pays soient vouées à l'échec, mais que pour ce faire, il est nécessaire d'établir des normes et des standards minimaux en deçà desquels les comparaisons deviennent douteuses.

3. Pourquoi standardiser

La mise en œuvre de standards et de contrôles dans ce genre d'études permet de valider la comparaison des résultats et renforce la crédibilité de ceux-ci yeux des utilisateurs. Le but recherché est de pouvoir attribuer une différence significative (statistiquement parlant) observée dans les résultats à une réelle différence dans les populations comparées, et non au fruit d'une combinaison d'erreurs non contrôlées (taux de réponse inadéquat, couverture mal assurée, piètre qualité des mises en œuvre des procédures sur le terrain, erreurs de mesures, *etc*.). La motivation des utilisateurs à procéder à des analyses des résultats est fortement influencée par la qualité des données et la pertinence des comparaisons de résultats.

L'établissement de standards facilite également grandement la mise en œuvre du plan de sondage. Dans le domaine particulier des enquêtes internationales touchant l'éducation, standardisation rime souvent avec unification des procédures. À titre d'exemple, on notera que le plan d'échantillonnage, la taille d'échantillon, le mode de collecte, ou encore le libellé des questionnaires d'enquêtes prestigieuses telles que PISA, ou TIMSS est à peu de choses près le même pour tous les participants (des exemples de standards suivront dans les prochaines sections). Cette unification permet un meilleur équilibre des charges de travail de chacun des pays participants à l'enquête et permet aussi la mise sur pied de mesures minimales efficaces et uniformisées de contrôle de la qualité des données.

En résumé, l'établissement de standards et de contrôles est profitable à tous les intervenants. Aux commanditaires qui seront rassurés du bien-fondé de leur investissement en obtenant des résultats valables et comparables, aux gestionnaires d'enquêtes qui peuvent se porter garants de la qualité des procédures et de la validité des résultats, et enfin aux pays participants qui obtiennent des résultats d'une qualité assurée et ce pour une charge de travail somme toute équivalente peu importe les conditions, contraintes et environnement qui leurs sont propres.

Nous ferons maintenant un court étalage des standards établis et des contrôles habituellement effectués dans les enquêtes comparatives internationales en l'éducation pour garantir la comparabilité des résultats.

4. Exemples de standards et contrôles

4.1 Introduction

La différence que nous faisons entre standards et contrôles est la suivante : les standards sont des normes établies pour assurer la qualité des données. Ces normes sont décrites et documentées. La non-conformité aux standards se traduit habituellement par des conséquences pour les pays fautifs, allant de simples remarques dans les tableaux de publication à la mise en annexe[6] de tous leurs résultats. Le caractère des contrôles est quant à lui moins formel et se traduit souvent par des critères plus souples. On parle plutôt de procédures de contrôle de qualité. Ces procédures de contrôle ont été définies et mises en application avec l'expérience acquise au cours des années. Cependant, certains contrôles qui existaient au début des enquêtes comparatives internationales sont maintenant devenus des standards établis.

[5]On entend par culture d'enquête l'ouverture qui existe face aux enquêtes dans les différents pays. Pour certains, la participation aux enquêtes peut être obligatoire alors que pour d'autres les enquêtes sont volontaires et parfois perçues comme étant une atteinte à la vie privée.
[6]Les résultats des pays fautifs sont exclus des tableaux principaux et se retrouvent en annexe à la fin des rapports.

4.2 Populations cible et d'enquête

Les premiers exemples de standards se rapportent à la population cible et la population d'enquête. Ainsi, chaque pays participant est tenu de fournir une description de leur population nationale. Ils doivent au minimum, décrire le système éducationnel (par exemple, l'âge minimal d'entrée à l'école, une description de la structure scolaire, le niveau ISCED[7]). La population d'enquête doit couvrir au moins 95 % de la population cible (ce standard est probablement le plus reconnu de la communauté en éducation). Toute divergence entre la population d'enquête et la population cible doit être documentée (type et ampleur des exclusions). Finalement, une couverture de la population cible pour un pays donné de moins de 95 % entrainera automatiquement une annotation dans les publications internationales à cet effet. Cette norme de 95 % peut sembler élevée mais il est important de se souvenir que dans les publications internationales, les tableaux de résultats comportent une ligne par pays. L'utilisateur suppose naturellement que les résultats sont représentatifs du pays dans son ensemble et c'est donc sur cette base même que les comparaisons entre les pays sont faites. Tel que mentionné précédemment à propos des enquêtes en éducation, il y a souvent une forte présomption de corrélation entre la couverture et la performance mesurée. On peut difficilement réduire ces exigences et encore prétendre que les comparaisons sont crédibles.

Certains contrôles sont fréquemment utilisés pour assurer que la définition de la population cible nationale corresponde à la définition de la population cible internationale. Des vérifications additionnelles de l'information fournie par le pays (tel que l'âge, le nombre d'années de scolarisation, la fréquentation scolaire[8], *etc*.) sont effectuées en consultant des sources externes de données. Dans le contexte d'enquêtes répétées, nous vérifions que la définition des populations est comparable d'un cycle à l'autre pour permettre des estimations de tendances valides. Si la population cible change d'un cycle à l'autre, il faut identifier la portion de la population qui est commune aux deux cycles pour pouvoir effectuer l'analyse des tendances. Cette dernière situation n'est pas si improbable. Par exemple, cette situation peut survenir suite à un changement dans le système d'éducation d'un pays qui ferait en sorte que l'année scolaire représentant 4 années d'éducation formelle (servant souvent de base pour définir une population) passe de la 4ième année à la 5ième année avec la venue du nouveau cycle (s'échelonnant sur 3 à 5 ans en général). Une réforme peut également modifier l'âge d'entrée à l'école. Présentement, il n'y a pas à notre connaissance de normes stipulant de manière explicite un pourcentage à atteindre dans ces situations (tel qu'un pourcentage minimum de la population commune aux deux cycles).

4.3 Plan de sondage

Nous discuterons ici des standards appliqués à la base de sondage, au mode de sélection des échantillons, à la taille des échantillons et à la mise en œuvre du plan d'échantillonnage.

4.3.1 Base de sondage

Dans les enquêtes en éducation, la base de sondage utilisée est souvent composée d'une liste d'écoles et la mesure de taille correspond aux nombres d'unités faisant partie de la population cible (professeurs ou élèves). À notre connaissance, il n'existe pas de standards établis et publiés pour en vérifier la qualité. Il y a par contre de bonnes pratiques (contrôles) mises en place qui correspondent aux vérifications habituellement faites sur toutes bases de sondage. Dans un premier temps, nous vérifions que la base de sondage soit la plus à jour possible. Ensuite, on s'assure que la base de sondage fournie par les participants donne une couverture complète de la population d'enquête, et qu'elle ne contienne pas de données erronées, de duplicats, ou d'éléments extérieurs à la population d'enquête/cible. De plus, autant que possible, une mesure de taille à jour pour chacune des unités constituant la base est exigée. Nous insistons également auprès des pays pour que la base de sondage fournie donne accès à l'ensemble de la population cible. Ceci permet de mieux estimer et documenter les exclusions. Enfin, nous utilisons des outils tels que le web, les informations des cycles précédents, les informations provenant d'autres pays pour valider les informations fournies par les représentants des pays. Par exemple, plusieurs pays ont des écoles internationales. Un

[7]l'*International Standard Classification of Education* (ISCED) est une norme de classification des systèmes d'éducation provenant de l'UNESCO.

[8]La fréquentation scolaire se définit comme le pourcentage de la cohorte d'âge d'une année scolaire donnée qui fréquente l'école.

pays en particulier peut avoir omis d'inclure ces écoles parce qu'il ne les considère pas comme des écoles faisant partie de son système d'éducation.

4.3.2 Mode de sélection

En ce qui a trait aux standards reliés au mode de sélection des échantillons, la norme est d'imposer un mode de sélection unique pour tous les pays participants. Les adaptations et/ou déviations sont permises mais elles doivent être approuvées par les gestionnaires de l'enquête avant la mise en œuvre et elles doivent aussi être documentées. L'imposition d'un mode de sélection unique nous permet d'établir et d'utiliser des programmes généralisés de sélection et de pondération des échantillons minimisant ainsi les risques d'erreurs. Cette approche facilite la répartition équitable des charges de travail pour chacun des participants, permettant d'uniformiser considérablement les tailles d'échantillons d'un pays à l'autre. Elle facilite également la validation des sélections et permet de minimiser le nombre de programmes de contrôle nécessaires pour la mise en œuvre. Enfin, l'adoption d'un mode unique permet d'uniformiser les opérations de collecte et donc de limiter le nombre de manuels d'opérations requis. Ainsi, les risques d'erreurs sont réduits, évitant que des différences dans les instructions aient des répercussions sur la qualité et la comparabilité des données. Enfin, le recours à un seul mode de sélection contribue à rassurer les participants sur la comparabilité des résultats : on s'attend à ce que les erreurs non dues à l'échantillonnage soient comparables. On s'attend également à ce que les erreurs d'échantillonnage soient de magnitude similaire – ce qui n'est pas nécessairement le cas mais la perception demeure. Notons qu'il y a des conséquences à ne pas satisfaire aux standards. Le risque que les données ne soient pas publiées ou qu'elles soient annotées dans les tableaux croît considérablement si le plan n'est pas approuvé ou que des anomalies sont observées.

4.3.3 Taille d'échantillon

Invariablement dans ce genre d'enquête, un standard définit la taille minimale de l'échantillon. La plupart du temps, celle-ci est fixée en fonction des marges d'erreur souhaitées et des contraintes intrinsèques à l'étude. Un autre standard souvent retenu concerne l'identification à priori d'écoles dites de remplacement. De façon générale, il y a un maximum de deux écoles de remplacement pour chaque école originalement choisie. Encore une fois, toute déviation doit être documentée et approuvée. Notons que les écoles de remplacement ne peuvent être utilisées que pour remplacer les écoles éligibles mais refusant de participer. L'utilisation de remplacement permet de satisfaire aux exigences concernant la taille des échantillons et peut contribuer à minimiser le risque de biais. Nous conservons cependant des exigences strictes en ce qui a trait au taux minimal de participation des écoles originalement sélectionnée (voir la section suivante).

4.3.4 Mise en oeuvre

Le standard le plus reconnu en ce domaine est l'obligation pour tous les pays de participer à une mise à l'essai. Celle-ci permet de tester les procédures sur le terrain et surtout de mettre en place des correctifs (comme c'est souvent le cas) avant la mise en œuvre de l'enquête. Cette participation à la mise à l'essai est en principe obligatoire sous peine de voir les résultats des pays fautifs exclus des publications internationales.

Un autre standard reconnu est le taux de réponse minimal. Nous pouvons en général distinguer trois zones :

1) La zone du minimum absolu ou zone rouge. À défaut de ne pas satisfaire à ces taux minimum, un pays se verra tout bonnement exclus de toutes publications internationales ;

2) À l'opposé, il y a la zone verte. Un pays se situe dans la zone verte si ces taux de participation sont au-delà d'un certain seuil, et ce sans le recours aux écoles de remplacement. On dit alors que le risque de biais dans les résultats découlant des données provenant de ce pays est négligeable. Si ce seuil est atteint seulement une fois que les écoles de remplacement sont prises en ligne de compte, alors les résultats de ce pays sont inclus dans les publications internationales mais comprennent une annotation afin d'alerter les utilisateurs d'un risque accru de biais ;

3) Finalement, il y a la zone grise. Si les taux de participation d'un pays sont au-delà du seuil minimum définit par la zone rouge, mais n'atteignent pas le seuil minimum définit par la zone verte, même après usage d'écoles de remplacement, alors une décision est habituellement prise au cas par cas. Les résultats peuvent se retrouver en fin de tableaux, en annexes, ou encore ne pas être publiés.

Il est impératif que les taux de réponse (parfois aussi appelés taux de participation) soient documentés afin que les analystes puissent évaluer la qualité des inférences et analyses issues des données.

En ce qui touche les contrôles, notons que durant la mise en œuvre, les pays ont souvent comme instruction de contacter les gestionnaires des enquêtes lorsqu'ils rencontrent une situation inhabituelle. Ceci permet de contrôler et d'agir avant que la collecte de données ne soit terminée et qu'il n'y ait plus de correctifs possibles. La présence d'une mesure de la taille des écoles sur la base de sondage permet un autre contrôle qui consiste à comparer cette dernière à la taille observée une fois sur le terrain. Il est alors possible de demander des justifications et une documentation plus pointue pour les cas où les différences sont importantes (omission de classes, erreur dans l'identification de l'école, changement dans la structure même de l'école visée, *etc.*). De plus, il n'est pas rare de procéder à une validation du statut des écoles non participantes une fois la collecte terminée afin de déterminer s'il n'a pas lieu de considérer un statut plus approprié (certains refus comme des exclusions par exemple). Toujours dans le domaine des contrôles, les erreurs-types sont calculées entre autres pour détecter les valeurs aberrantes, les valeurs influentes, les poids influents ou aberrants en fonction de la principale variable d'intérêt. Il est également possible d'effectuer une comparaison des estimations observées et attendues (comme par exemple, les taux d'exclusions à l'intérieur des écoles comparés à ceux des cycles précédents, les totaux de populations comparés aux totaux connus des cycles précédents, *etc.*). Tous ces contrôles nous permettent de détecter de possibles manquements aux règles dictées par le plan d'échantillonnage. Des justifications écrites auprès des participants sont habituellement exigées pour toutes anomalies détectées.

Comme dernier point ici, nous aimerions souligner l'importance d'effectuer une évaluation de la mise en œuvre. La revue des plans de sondage et de la mise en œuvre est en général effectuée en présence d'un expert extérieur au cercle de gestion de l'enquête. Cette évaluation et approbation indépendante apporte une crédibilité importante à l'enquête. Finalement, il est essentiel de clore le tout par la rédaction d'un rapport technique décrivant l'ensemble des procédures touchant le plan de sondage et sa mise en œuvre.

5. Conclusion

De notre expérience sur les enquêtes comparatives internationales, nous retenons ceci :

1) qu'il est difficile d'avoir des standards qui à la fois répondent à tous les besoins tout en conservant une certaine souplesse. À chaque enquête ou cycle, nous devons faire face à des situations particulières. Il devient donc important de mettre en place une équipe technique responsable de fournir un soutien aux participants et d'inviter ceux-ci à consulter cette équipe avant et pendant la mise en œuvre de l'enquête afin de palier les différents imprévus ;

2) que l'établissement de standards est nécessaire et primordial afin de dissiper tous doutes sur la pertinence des analyses découlant de l'enquête ;

3) qu'il est important d'exercer à un coût raisonnable un contrôle sur toutes les procédures menant à bonifier la qualité des données recueillies ;

4) qu'il est tout aussi important de quantifier et de documenter les actions prises afin d'assurer la qualité des données et leur comparabilité et ainsi renforcir la confiance envers les chargés de la mise en œuvre.

On peut certes prétendre qu'il y a toujours place à amélioration. Certains contrôles décrits plus haut pourraient facilement être raffermis pour devenir des standards. Par contre, comme gestionnaires d'enquêtes, on se doit de se montrer prudent et de conserver une certaine souplesse, toujours avec l'optique de garantir un niveau de qualité acceptable. Les standards sont contraignants pour les pays participants. Donc, devrait-on viser plus de standards au détriment d'une certaine souplesse et d'accommodements sur le terrain, ou plutôt son contraire, avec les risques de dérives possibles? La discussion reste ouverte…

Bibliographie

TIMSS & PIRLS International Study Center (2007), *TIMSS 2007 Technical Report*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.

TIMSS & PIRLS International Study Center (2005), *TIMSS 2007 school sampling manual*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.

SÉANCE 9B

LOGICIELS NORMALISÉS

Harmonisation des pratiques de désaisonnalisation grâce à l'élaboration du logiciel DEMETRA+

Jean Palate et Pascal Jacques[1]

Résumé

La désaisonnalisation est une étape importante de l'architecture opérationnelle des statistiques officielles, et l'harmonisation des pratiques s'est révélée un élément clé de la qualité des produits. Dans cet esprit, depuis les années 1990, Eurostat a joué un rôle dans la promotion, l'élaboration et le maintien d'une solution logicielle (Demetra), disponible sans frais pour la désaisonnalisation, en conformité avec les pratiques exemplaires établies.

En 2008, les lignes directrices du SSE (Système statistique européen) sur la désaisonnalisation ont été appuyées par le Comité des statistiques monétaires, financières et de la balance des paiements et le CPS (Comité du programme statistique) comme cadre pour la désaisonnalisation des PIEE (Principaux indicateurs économiques européens) et d'autres indicateurs économiques du SSE et du Système européen de banques centrales.

Les lignes directrices du SSE englobent les principales étapes du processus de désaisonnalisation et de calendrialisation et représentent une étape importante en vue de l'harmonisation des pratiques de désaisonnalisation et de calendrialisation à l'intérieur du SSE et d'Eurostat. Une politique commune pour la désaisonnalisation et la calendrialisation de toutes les statistiques infra annuelles améliorera la qualité et la comparabilité des données nationales, ainsi que la qualité globale du système européen, à condition que des outils de désaisonnalisation appropriés existent et soient disponibles.

Le Groupe directeur de la désaisonnalisation (groupe d'experts de haut niveau d'Eurostat Banque centrale européenne représentant les Institut nationaux de statistique et les Bureaux centraux nationaux, qui a produit les lignes directrices de désaisonnalisation du SSE) favorise le développement d'une solution logicielle souple pour la désaisonnalisation, qui doit être utilisée par le SSE.

Le groupe a porté son attention sur les technologies orientées objet utilisées par l'unité de recherche et développement du service des statistiques de la Banque nationale de Belgique pour élaborer une série de prototypes d'outils pour la désaisonnalisation. Cela a été considéré comme un cadre approprié pour l'élaboration en collaboration d'une nouvelle génération d'outils de désaisonnalisation durables, permettant la mise en œuvre des lignes directrices du SSE et le remplacement de l'ancien logiciel Demetra.

Le nouveau logiciel de désaisonnalisation (Demetra+) a été diffusé dans sa version .NET, C# comme outil officiel pour soutenir la mise en œuvre des lignes directrices. Les travaux se poursuivent concernant le remaniement de l'outil Demetra+ en JAVA, y compris des moteurs de base Tramo/Seats et X12/Arima et sa diffusion comme Open Source sur la plateforme OSOR. Cela mènera au déploiement d'un outil du SSE souple utilisant plusieurs plateformes pour la désaisonnalisation, qui facilitera la mise en œuvre de lignes directrices du SSE concernant la désaisonnalisation et répondra aux exigences des utilisateurs au profit de la collectivité chargée de la désaisonnalisation et d'un futur système de production des comptes nationaux.

[1]Jean Palate, National Bank of Belgium, et Pascal Jacques, Eurostat, Luxembourg.

Comment fonctionne le SCANCIR et peut-il être utile à un plus grand nombre d'utilisateurs?

Chunxiao (William) Liu, Sean Crowe et Asma Alavi[1]

Résumé

Le Système canadien de contrôle et d'imputation du recensement (SCANCIR) offre aux utilisateurs une méthodologie de contrôle et d'imputation souple, efficace et guidée par les données qui permet :

- de spécifier de grands nombres de contrôles au moyen de tables de décision logique ;
- de procéder à l'imputation par donneur et à l'imputation déterministe, et de dériver de nouvelles variables ;
- de travailler simultanément avec divers types de variables ;
- de travailler avec des fichiers de données au niveau de l'unité ou de la sous unité ;
- de personnaliser les fonctions au moyen de paramètres ;
- de l'utiliser dans la plupart des environnements informatiques ;
- d'ajouter continuellement de nouvelles caractéristiques en vue de répondre aux besoins croissants des utilisateurs.

Le SCANCIR recherche les actions d'imputation (AI) avec changement minimal d'une manière très efficiente et efficace, fortement guidée par les données, en commençant par repérer les donneurs qui ressemblent le plus à l'unité nécessitant une imputation (les plus proches voisins), puis en déterminant les meilleures AI pour ces donneurs. L'utilisateur peut contrôler de manière très fine la définition des plus proches voisins au moyen de variables d'appariement, ainsi que la façon dont les AI sont déterminées à partir des plus proches voisins.

Le SCANCIR a pour fondement la méthode d'imputation par le plus proche voisin (MIPPV) élaborée par Michael Bankier, à Statistique Canada, en 1992. Il est utilisé pour traiter les données des recensements du Canada depuis 1996. Il a également été employé pour certaines enquêtes par sondage à Statistique Canada et par plusieurs organismes statistiques d'autres pays.

L'article traite de la méthodologie du SCANCIR, de son application et des possibilités d'utilisation plus générales.

Mots clés : Imputation par la méthode du plus proche voisin ; changement minimal ; guidé par les données ; table de décision logique ; actions d'imputation.

1. Introduction

Au départ, le Système canadien de contrôle et d'imputation du recensement (SCANCIR) a été développé pour exécuter les tâches de contrôle et d'imputation pour les données du Recensement du Canada de 1996. Il a été utilisé pour cinq domaines spécialisés dans le cas du Recensement du Canada de 2001 et pour presque la totalité des variables de recensement dans le cas du Recensement de 2006. Il sera de nouveau utilisé pour la totalité des variables du Recensement de 2011 et de l'Enquête nationale auprès des ménages (ENM) de 2011. Le SCANCIR a été conçu principalement pour travailler avec des fichiers textes et son code a été réécrit en langage machine C# (C sharp) dans l'environnement .NET pour le Recensement et l'ENM de 2011. Les éléments nouveaux comprennent la capacité de lire des fichiers en format EXCEL et de générer des fichiers de sortie en format HTML.

Traditionnellement, le Recensement du Canada comprenait deux types de questionnaire : l'un (appelé questionnaire abrégé) pour recueillir les réponses à quelques questions démographiques et l'autre (appelé questionnaire complet) comprenant plus de 50 questions, y compris celles du questionnaire abrégé, portant sur divers sujets. Le questionnaire abrégé était envoyé à 80 % des ménages canadiens, tandis que le questionnaire détaillé était envoyé à un échantillon d'un ménage canadien sur cinq. La réponse aux deux types de questionnaire était obligatoire, sauf en 2011, année où le questionnaire complet du recensement a été remplacé par une enquête distincte à participation volontaire menée auprès d'un ménage canadien sur trois et nommée Enquête nationale auprès des ménages.

[1]Chunxiao (William) Liu, Sean Crowe, et Asma Alavi, Statistique Canada, Immeuble R. H. Coats, 15e étage, 100 promenade pré Tunney, Ottawa (Ontario) Canada, K1A 0T6, asma.alavi@statcan.ca.

À l'heure actuelle, les opérations de contrôle et d'imputation des données du recensement et de l'ENM du Canada sont subdivisées en plusieurs processus, en fonction des caractéristiques et des liens internes des questions. Dans le SCANCIR, un processus de contrôle et d'imputation est une suite de deux types de module, à savoir des modules de dérivation et des modules d'imputation par donneur. Chacun de ces modules accomplit une tâche particulière afin d'exécuter la stratégie de contrôle et d'imputation établie pour le processus.

Dans le présent article, nous expliquons comment le SCANCIR applique l'approche du changement minimal guidée par les données pour imputer simultanément tous les types de variables en spécifiant facilement un grand nombre de contrôles. Cela étant compris, nous discuterons des possibilités qu'a le système d'être utile à un plus grand groupe d'utilisateurs.

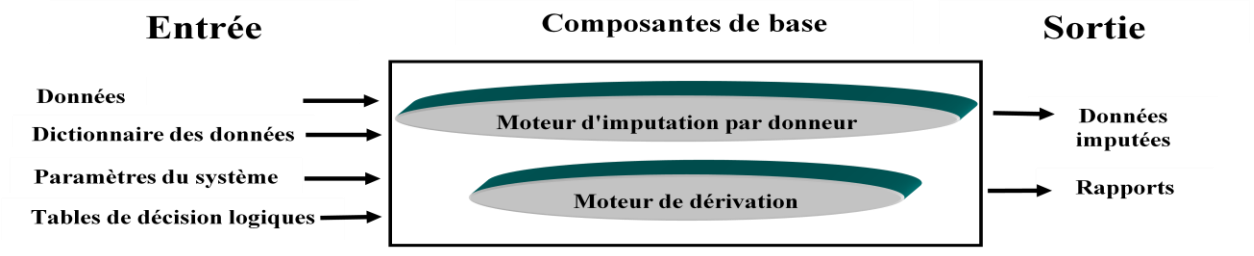
2. Aperçu du SCANCIR

Le SCANCIR est fondé sur la méthodologie d'imputation par la méthode du plus proche voisin (MIPPV). Proposée comme alternative à la méthodologie bien connue de Fellegi Holt (Fellegi, I.P. et Holt, D., 1976), la MIPPV consiste à trouver d'abord des donneurs potentiels, puis à déterminer le nombre minimal de variables à imputer sur la base de chaque donneur potentiel. L'inversion des étapes d'imputation est conforme à l'approche guidée par les données et donne à la MIPPV d'importants avantages sur le plan des calculs, tout en réalisant les objectifs qui consistent à imputer le plus petit nombre possible de variables et à préserver dans la mesure du possible les distributions des sous populations. Notons que l'expression « plus petit nombre possible de variables » est une simplification excessive car, par exemple, le SCANCIR préférerait, en général, modifier les valeurs de deux variables d'une très petite quantité que modifier la valeur d'une seule variable d'une très grande quantité (plus de détails seront données à la section 2.4).

Étant donné ces avantages sur le plan des calculs, le SCANCIR peut traiter de très gros fichiers de données, c'est-à-dire un grand nombre d'enregistrements et de variables, et permet aux utilisateurs de spécifier de nombreux contrôles. Le SCANCIR est capable de traiter simultanément des variables catégoriques, numériques et alphanumériques, offrant ainsi la plus grande portée possible pour la recherche des donneurs les plus appropriés.

Le système est constitué de deux composantes, à savoir le moteur de dérivation et le moteur d'imputation par donneur. Le moteur de dérivation exécute l'imputation déterministe et crée de nouvelles variables, tandis que le moteur d'imputation par donneur exécute, comme son nom l'indique, l'imputation par donneur. Autour de ces deux composantes de base s'articulent quatre types de fichiers d'entrée fournis par les utilisateurs et plusieurs fichiers de sortie produits par le système, comme l'illustre la figure 1 qui suit.

Figure 1 : Composantes du SCANCIR



Nous commençons par décrire les fichiers d'entrée et de sortie du SCANCIR, puis les composantes de base du système.

2.1 Fichiers d'entrée

Le fichier de données d'entrée contient les enregistrements qui doivent être traités. L'objet que doit traiter le SCANCIR est l'unité. Une unité peut être un enregistrement unique ou comprendre plusieurs enregistrements. Dans ce dernier cas, les enregistrements qui forment une unité sont appelés sous unités, par exemple les membres du ménage dans un ménage. Il convient de souligner que le SCANCIR ne peut traiter ensemble que des unités contenant des nombres identiques de sous unités.

Afin d'augmenter l'efficacité et l'efficience de la recherche de donneurs, les unités semblables doivent être placées aussi près l'une de l'autre que possible dans le fichier en se fondant sur l'expérience des spécialistes du domaine. Dans le cas du traitement des données du Recensement du Canada par exemple, lors du traitement des données démographiques, les ménages sont disposés dans le fichier en fonction de leur proximité géographique, étant donné l'hypothèse émise depuis longtemps que des familles similaires vivent dans un même quartier. Cependant, dans d'autres situations, une disposition différente pourrait être préférable. Par exemple, dans le cas du traitement du revenu, les unités sont classées en fonction de leur revenu total pour l'imputation des composantes du revenu. À leur tour, les sous unités sont classées dans une unité en fonction d'un ordre choisi, et l'évaluation de la similarité de deux unités est faite en évaluant la similarité des sous unités correspondantes.

Le dictionnaire de données fournit des renseignements qui seront utilisés dans le module, tels que les noms et types de variable, ainsi que toutes les valeurs possibles et valides des variables. Les utilisateurs peuvent assez facilement construire un dictionnaire de données, en format texte ou Excel, en se servant des modèles fournis. Le SCANCIR emploie des étiquettes textes définies par l'utilisateur pour les variables catégoriques au lieu des codes numériques connexes, par exemple « HOMME » et « FEMME » plutôt que « 1 » et « 2 ».

Le processus de contrôle et d'imputation est entièrement contrôlé par les paramètres du système. Ces derniers jouent un rôle important dans la souplesse, l'adaptabilité et l'efficacité du SCANCIR en permettant à l'utilisateur de contrôler entièrement un grand éventail de choix disponibles. Pour n'en nommer que quelques uns, l'utilisateur peut exercer un contrôle sur la vérification, la recherche de donneurs, le réarrangement des sous unités, la piste de vérification du processus de contrôle et d'imputation, etc.

Dans le SCANCIR, des tables de décision logique (DTL) sont utilisées pour spécifier les règles de contrôle pour les modules de dérivation ainsi que les modules d'imputation par donneur. Les règles de contrôle définies dans le SCANCIR sont appelées règles de contrôle de conflit, car le système recherche les anomalies dans les données et non celles qui sont à l'intérieur des normes. Autrement dit, un conflit a lieu lorsqu'il existe des incohérences entre les variables. Dans les modules de dérivation, certaines règles de contrôle spécifient des conditions, qui sont les situations de conflit dans les données en ce qui a trait aux variables d'intérêt, et des actions, qui sont les solutions aux conflits. Ces actions correspondent à la façon dont les TDL de dérivation exécutent l'imputation déterministe. D'autres règles de contrôle figurant dans les TDL de dérivation spécifient la création de nouvelles variables sous certaines contraintes. Par ailleurs, dans les modules d'imputation par donneur, les TDL types spécifient des situations de conflit qui doivent être résolues sans qu'une action soit spécifiée par l'utilisateur. Aucune action n'est spécifiée, puisque l'objectif est de trouver un donneur afin de résoudre les incohérences.

Les figures 2 et 3 sont des exemples de TDL à utiliser dans les modules de dérivation et d'imputation par donneur, respectivement. Dans les deux figures, les propositions (c'est-à-dire les lignes des figures) de la première règle de contrôle (c'est-à-dire la deuxième colonne dans les figures) indiquent l'incohérence (c'est-à-dire le conflit) que représente une personne de moins de 15 ans définie comme ayant pour ÉTMAT (c'est-à-dire l'état matrimonial) la valeur DÉJÀ_MARIÉ(E)(E) (définie comme étant marié(e), divorcé(e) ou veuf(ve)). La TDL de dérivation prescrit la solution qui consiste à modifier de manière déterministe l'état matrimonial pour qu'il devienne « célibataire ». Par ailleurs, la TDL d'imputation par donneur prescrirait qu'il faut trouver un donneur permettant d'imputer l'ÂGE ou l'ÉTMAT, ou les deux variables afin de résoudre le conflit. Il convient de souligner qu'en général, les variables ÂGE et ÉTMAT ne seront pas imputées ensemble puisque cela violerait le principe du changement minimum.

Dans ces TDL, nous pouvons voir deux autres caractéristiques du SCANCIR. La première est l'expression CLASSE (DÉJÀ_MARIÉ(E)), qui se rapporte au groupe de certaines valeurs de la variable ÉTMAT, définies par l'utilisateur dans le dictionnaire des données. Si cette caractérisque n'était pas utilisée, il serait nécessaire d'établir des règles de contrôle distinctes pour chacune de ces valeurs. Une deuxième caractéristique consiste à n'écrire qu'une seule fois

les règles de contrôle communes à toutes les sous unités d'une unité. Le symbole « #1 » représente une « position de sous unité » et énonce que la règle est appliquée à toutes les sous unités. Sans cette caractéristique, il serait nécessaire de répéter les mêmes règles de vérification pour chaque sous unités.

Figure 2 : TDL de dérivation

Propositions	Règle de contrôle 1	Règle de contrôle 2
AGE(#1) < 15	Y	Y
ETMAT(#1) = CLASSE(DEJA_MARIE)	Y	
REVENU(#1) > 0		Y
ETMAT(#1) = CELIBATAIRE	X	
REVENU(#1) = 0		X

2.2 Fichiers de sortie

À la fin du processus de contrôle et d'imputation, le système produit les données contrôlées et (ou) imputées, ainsi que certains rapports. En particulier, parmi les fichiers de sortie, la piste de vérification d'un module d'imputation par donneur fournit un rapport étape par étape très utile qui montre comment a été trouvé le meilleur donneur pour une unité particulière. Un autre fichier de sortie renseigne sur le nombre de fois qu'un donneur particulier a été utilisé. Le nombre de fichiers de sortie et le niveau de détail des renseignemenùts fournis sont contrôlés par l'utilisateur à l'aide des paramètres du système.

Figure 3 : TDL d'imputation par donneur

Propositions	Règle de contrôle 1	Règle de contrôle 2
AGE(#1) < 15	Y	Y
ETMAT(#1) = CLASSE(DEJA_MARIE)	Y	
REVENU(#1) > 0		Y

2.3 Moteur de dérivation

Le moteur de dérivation a la capacité de créer de nouvelles variables ou d'effectuer une imputation déterministe pour corriger des erreurs systématiques en se fondant sur l'expérience des spécialistes du domaine. Pour cela, le SCANCIR fournit les mêmes fonctionnalités, dans ses TDL, que la plupart des langages machines, telles que des fonctions (par exemple, attribution de valeurs aléatoires, recherche de la valeur maximale, *etc*.), des commandes « GO TO », des « boucles DO » et des « appels d'autres TDL ».

2.4 Moteur d'imputation par donneur

Le moteur d'imputation par donneur exécute l'imputation par donneur et possède aussi des fonctions de contrôle des données pour faciliter l'exécution de l'imputation. Le contrôle est effectué et géré grâce à la spécification de classes de validité (valeurs acceptables pour les variables) et de règles de contrôle de conflit définies dans les TDL. Chaque unité est évaluée afin de repérer les valeurs invalides (valeurs non comprises dans la classe de validité) et les incohérences (unité concordant avec les règles de contrôle de conflit). Les unités ne présentant pas de valeurs invalides ni d'incohérences sont acceptées au contrôle et sont appelées unités acceptées, tandis que les unités comportant une ou plusieurs valeurs invalides et (ou) une ou plusieurs incohérences sont rejetées au contrôle et sont

appelées unités rejetées. Conceptuellement, le SCANCIR traite les unités rejetées au contrôle successivement, dans l'ordre de leur apparition, mais en pratique, le système est capable de traiter simultanément, de manière indépendante, plusieurs unités rejetées.

Même si, pour que cela soit plus commode pour l'utilisateur, les contrôles peuvent être spécifiés dans un grand nombre de petites tables de décision logique, le SCANCIR crée une version unifiée des contrôles en combinant tous les contrôles individuels et en supprimant les redondances, afin de s'assurer qu'après l'imputation, les enregistrements soient acceptés à tous les contrôles. Lorsque l'on procède à l'imputation d'un enregistrement particulier, il est souvent possible de laisser tomber de nombreux contrôles, parce que le donneur potentiel est repéré en premier lieu. Par exemple, si la ménage rejeté au contrôle ainsi que le ménage donneur potentiel examinés ne comprennent ni l'un ni l'autre des grands parents, le SCANCIR écarte tous les contrôles ayant trait aux grands parents, parce qu'ils ne sont pas pertinents. La personnalisation de l'ensemble de contrôles pour chaque paire d'unités rejetée acceptée réduit considérablement le temps de traitement.

Pour chaque unité rejetée, le système prend en considération de nombreuses unités acceptées au contrôle et repère les plus proches voisins parmi celles ci. La détermination des plus proches voisins est effectuée sur plusieurs dimensions en comparant un ensemble donné de variables d'appariement. Comme nous l'avons mentionné plus haut, un effort est fait en vue de regrouper à l'avance les unités semblables dans le fichier de données d'entrée. Cela permet au système de trouver les voisins les plus proches dans le temps le plus court possible. La détermination des voisins les plus proches est fondée sur la similarité des valeurs pour les unités rejetées et les unités acceptées pour chaque variable d'appariement, ainsi que sur le poids (importance relative) de chaque variable d'appariement. Ces concepts sont appliqués de manière quantitative grâce à l'instrument de mesure D_{fp} qui est la mesure de distance entre l'unité rejetée et une unité acceptée :

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi}) \qquad (1)$$

où V_{fi} et V_{pi} sont les valeurs de la i^e variable d'appariement pour les unités rejetée et acceptée, respectivement, w_i est le poids attribué à la i^e variable d'appariement dans le dictionnaire des données et D_i est la fonction de distance choisie pour la variable d'appariement afin d'évaluer la similarité de V_{fi} et V_{pi}. Le poids attribué à une variable d'appariement reflète habituellement la valeur de celle ci en tant que prédicteur des variables qui doivent être imputées. Le choix de poids différents pour des variables différentes est fondé principalement sur l'expérience des spécialistes du domaine. À l'heure actuelle, le SCANCIR compte dix fonctions de distance qui lui permettent de traiter tous les types de variables. Chaque fonction de distance possède également des paramètres définis par l'utilisateur afin d'accroître encore davantage la souplesse. Dans l'équation (1), la mesure de distance permet de regrouper tous les types de variables sur une même échelle, car les valeurs de D_i sont comprises entre 0 (valeurs « essentiellement égales ») et 1 (valeurs « totalement différentes »), raison pour laquelle le SCANCIR peut traiter divers types de variables simultanément. Pour une unité rejetée donnée, la liste des voisins les plus proches est retenue en vue d'une évaluation plus approfondie.

Pour un module d'imputation par donneur particulier, chaque variable qui figure dans le fichier d'entrée est définie par l'utilisateur comme étant imputable ou non imputable. Tant les variables imputables que non imputables pourraient être utilisées dans le module comme variables d'appariement et (ou) dans la spécification des contrôles des TDL au moyen de propositions. Durant la procédure d'imputation, les variables imputables possédant des valeurs invalides sont toujours imputées immédiatement, car cela est nécessaire pour que toute action d'imputation (AI), qui spécifie quelles valeurs seront tirées du voisin le plus proche potentiel, soit fructueuse. Dans la situation la plus simple, l'unité rejetée est alors acceptée à tous les contrôles et une AI potentielle a été trouvée. Sinon, si des incohérences persistent, pour chacune des variables imputables qui restent (pour lesquelles $V_{pi} \neq V_{fi}$), le choix demeure d'imputer ou non. Afin de faciliter une recherche efficace des AI potentielles, le système utilise un arbre binaire dans lequel chaque nœud donne deux branches représentant la décision d'imputer ou de ne pas imputer pour une variable particulière. Il convient de souligner qu'il est possible, pour un voisin le plus proche donné, de générer plus d'une AI potentielle ou de n'en générer aucune, par exemple, les AI consistant à imputer l'âge ou l'état matrimonial pour résoudre le conflit dans le premier contrôle de la TDL de la figure 3. Chaque voisin le plus proche figurant sur la liste des meilleurs voisins les plus proches est ensuite évalué, par ordre croissant de la valeur de D_{fp}, pour déterminer les AI potentielles.

Afin de décider quelle est la meilleure AI potentielle, le SCANCIR utilise l'outil de mesure quantitative D_{fpa}.

$$D_{fpa} = \alpha\, D_{fa} + (1-\alpha)\, D_{ap} \qquad (2)$$

où, « a » représente l'unité imputé, D_{fa} et D_{ap} sont définies comme dans l'équation (1), et α est un paramètre du système défini par l'utilisateur dont la valeur est comprise dans l'intervalle (0,5, 1]. En général, nous nous attendons à ce que les unités acceptées les plus semblables à l'unité rejetée, ce qui implique la distance D_{fp} la plus faible, soient celles qui donneront le plus vraisemblablement la meilleure AI ou l'AI avec changement minimal. La valeur de D_{fa} mesure la ressemblance entre l'unité imputée et l'unité rejetée, donc mesure l'aspect changement minimal de l'AI. Par ailleurs, la valeur de D_{ap} mesure la ressemblance entre l'unité imputée et l'unité acceptée, donc mesure l'aspect plausibilité de l'AI, parce que le donneur potentiel est formé entièrement de données réelles. Soulignons que, pour appliquer le principe du changement minimal, le paramètre α doit avoir une valeur supérieure à 0,5.

Le SCANCIR tient une liste des AI potentielles les meilleures, c'est-à-dire celles dont la valeur de D_{fpa} est la plus faible, pour une unité rejetée à mesure que les voisins les plus proches sont examinés. L'utilisateur spécifie le nombre maximal d'AI à retenir sur la liste et dans quelle mesure une AI peut être moins bonne que l'AI absolument la meilleure. Une fois que tous les proches voisins ont été examinés afin de déterminer toutes les AI potentielles, une AI est choisie au hasard sur la liste de toutes les AI potentielles, pour être l'action d'imputation effectivement appliquée à l'unité rejetée.

3. Contrôle et gains d'efficacité supplémentaires

L'utilisateur peut contrôler finement la plupart des aspects du processus de contrôle et d'imputation à l'aide de divers paramètres. Ces aspects comprennent, entre autres, la recherche par étape des plus proches voisins, la détection et le contrôle des valeurs aberrantes, le réordonnancement des sous-unités, l'utilisation des unités rejetées comme donneurs, ainsi que la capacité de personnaliser les poids et les fonctions de distance basé sur les caractéristiques des unités rejetées.

La recherche de donneurs est habituellement effectuée par étape pour deux raisons. Premièrement, il n'est pas forcément nécessaire ou pratique d'évaluer toutes les unités acceptées pour déterminer les AI potentielles. Deuxièmement, les meilleures unités donneuses sont souvent celles qui sont physiquement proches de l'unité rejetée, surtout quand les unités qui figurent dans le fichier de données ont été classées à l'avance de manière appropriée, en ce qui a trait à une caractéristique d'intérêt. Au moyen des paramètres du système, l'utilisateur peut exercer un contrôle sur le nombre d'unités qui sont considérées comme des voisins les plus proches et des donneurs potentiels à chaque étape et sur le nombre maximal d'étapes qui peuvent être effectuées. Avant de procéder à une étape supplémentaire, SCANCIR détermine si la qualité des AI potentielles a été considérablement améliorée à l'étape précédente. Dans la négative, aucune autre étape ne sera exécutée. L'utilisateur peut aussi spécifier un niveau de qualité minimal pour le plus proche voisin utilisé pour générer l'AI, afin que les unités acceptées dont la valeur de D_{fp} est trop élevée comparativement à celle de l'unité rejetée ne soient même pas prises en considération, ce qui économise du temps en évitant l'analyse de donneurs inacceptables pour les AI, tout en s'assurant que l'AI choisie finale soit conforme à un niveau de qualité minimal.

En se basant sur les meilleures AI trouvées jusque là, le SCANCIR peut conclure que cela ne vaut pas la peine de générer des branches supplémentaires à partir d'un nœud d'un arbre binaire, puisque toute AI obtenue à partir de ces branches ne serait pas assez bonne. De même, le SCANCIR peut terminer son évaluation des plus proches voisins pour les AI à une étape donnée s'il conclut que, comparativement à celle de l'unité rejetée, la D_{fp} des unités qui sont les plus proches voisines est devenue trop élevée pour produire des AI de qualité acceptable comparativement à la meilleure AI découverte jusque-là. Ces deux mécanismes et de nombreux autres augmentent considérablement l'efficacité du système.

4. Le SCANCIR peut-il être utile à un plus grand nombre d'utilisateurs?

Compte tenu des caractéristiques décrites aux sections précédentes et de notre expérience, il semble qu'outre les recensements et autres enquêtes sociales, le SCANCIR pourrait être utilisé par un plus grand nombre d'utilisateurs. Il importe de se rappeler qu'à l'heure actuelle, le SCANCIR peut procéder à l'imputation déterministe et à l'imputation par donneur selon la méthode du plus proche voisin, mais n'a pas la capacité d'exécuter d'autres types d'imputation, tel que l'imputation directe par la régression ou l'imputation historique. Cependant, il convient de souligner que le SCANCIR peut utiliser des variables qui sont corrélées aux variables d'intérêt, ce qui permettrait d'entrer un modèle de régression et des versions historiques des variables d'intérêt comme variables d'appariement auxiliaires pour trouver un donneur.

Le SCANCIR présente des limites dans un environnement informatique dont la mémoire et le nombre d'unités centrales est faible, tel qu'un ordinateur personnel. Toutefois, moyennant des ressources suffisantes, disons un serveur équipé de plusieurs processeurs de taille raisonnable, le SCANCIR, en effectuant un traitement multifilières, peut traiter des situations comprenant un très grand nombre d'enregistrements, de variables et de contrôles, et ce dans un laps de temps raisonnable. Nous insistons donc sur le fait que le SCANCIR pourrait être fort utile à des praticiens du contrôle et de l'imputation qui demandent :

- un système capable d'effectuer des imputations déterministes et par enregistrement donneur, et de dériver de nouvelles variables ;
- la capacité de traiter simultanément des variables catégoriques, numériques et alphanumériques ;
- la possibilité de définir facilement de grands nombres de contrôles ;
- la capacité de traiter rapidement et efficacement de grands fichiers de données ;
- la souplesse de pouvoir contrôler finement tous les aspects du processus à l'aide de simples paramètres définis par l'utilisateur ;
- un logiciel qui peut être utilisé directement sur la plupart des plateformes informatiques sans nécessiter l'installation complexe de programmes personnalisés.

Remerciements

Les auteurs remercient Marcel Bureau, Mike Sirois et Daniel Finch, qui travaillent tous à Statistique Canada, de leurs commentaires et suggestions utiles qui les ont aidés à améliorer l'article.

Bibliographie

Bankier, M., Lachance, M. et P. Poirier (1999), « A generic implementation of the nearest neighbour imputation method », *Proceedings of the Survey Research Methods Section*, American Statistical Association, p. 548 à 553.

Bankier, M., Poirier, P. et M. Lachance (2001), « Efficient Methodology Within the Canadian Census Edit and Imputation System (CANCEIS) », ASA Joint Statistical Meetings, Atlanta.

Bankier, M. (2011), « Imputing Numeric and Qualitative Variables Simultaneously », A Technical Report Detailing the Methodology of CANCEIS, rapport interne, Statistique Canada.

Fellegi, I.P. et D. Holt (1976), « A Systematic Approach to Automatic Edit and Imputation », *Journal of the American Statistical Association*, mars 1976, volume 71, no. 353, p. 17 à 35.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Développement de l'environnement de traitement des enquêtes sociales

Larry MacNabb[1]

Résumé

En 2009, Statistique Canada a entrepris l'élaboration d'un ensemble d'outils à l'appui des principales étapes du cycle de vie des enquêtes, allant de la période précédant la collecte à la diffusion. Selon le principe que les métadonnées devraient être à l'origine du processus, ces outils facilitent l'utilisation et le partage efficaces d'information entre les enquêtes et permettent la création efficiente de questionnaires d'enquête, de documentation d'enquête et de données d'enquête pleinement traitées. Le présent article portera sur les points clés suivants : historique et justification du développement de l'environnement de traitement générique; objectifs généraux du projet ; principes et pratiques exemplaires de soutien qui ont guidé l'élaboration de l'environnement et des outils de soutien. Un aperçu de l'environnement de traitement générique sera présenté, y compris une description des flux de métadonnées entre les diverses étapes de traitement. Les avantages du nouvel environnement au niveau méthodologique seront abordés, ainsi que les nombreux défis auxquels a fait face l'équipe de développement. Enfin, les résultats jusqu'à maintenant seront présentés, tout comme les prochaines étapes du projet.

Mots clés : Traitement ; outils génériques ; enquête.

1. Contexte

En 2005, un groupe de travail interdivisionnaire à Statistique Canada a mené un examen des activités de traitement pour toute une gamme d'enquêtes auprès des ménages dans le domaine de la statistique sociale. Les résultats ont montré que les processus opérationnels et les outils de soutien étaient optimisés dans les différents environnements de traitement. Même si ces résultats semblent optimaux lorsqu'ils sont examinés du point de vue d'une division particulière, cette optimisation à l'intérieur du domaine de traitement présente de nombreux défis lorsqu'elle est envisagée dans une perspective globale.

Les principaux inconvénients de cette approche de développement comprenaient : le développement de plusieurs outils pour exécuter des tâches similaires, les difficultés liées à la mise à jour de la technologie, les nombreux systèmes à tenir à jour, les inefficacités dans la formation du personnel et, enfin, les difficultés de gestion des priorités pour le développement de systèmes dans les secteurs de programme.

Par suite de cet examen, le groupe de travail a conclu que le développement d'une approche de traitement et d'outils de soutien génériques contribuerait dans une large mesure à résoudre nombre des problèmes observés.

2. Principes directeurs du développement

Au début de 2009, le Secteur de la statistique sociale, de la santé et du travail de Statistique Canada a lancé un projet ambitieux, en vue d'élaborer un ensemble d'outils de traitement génériques pour appuyer toutes les activités du cycle de vie des enquêtes. Le cycle de vie des enquêtes est défini comme comprenant toutes les activités liées à la pré collecte, à la collecte, au traitement et à la diffusion. Au moment du lancement de ce projet. Plusieurs principes ont été établis pour guider le développement de ce système.

De façon plus particulière, le système développé devait :

- optimiser la réutilisation des logiciels et des systèmes existants ;
- réduire les processus inefficaces ;

[1]Larry MacNabb, Statistique Canada, 100, promenade pré Tunneys, Ottawa (Ontario) Canada, K1A 0T6 (larry.macnabb@statcan.gc.ca).

- intégrer complètement les métadonnées dans le processus ;
- permettre l'intégration dans l'ensemble des processus opérationnels et entre eux ;
- faciliter le partage entre les différents domaines d'enquête.

3. Explication des processus fondés sur les métadonnées

Un processus fondé sur les métadonnées en est un qui s'autodocumente essentiellement. L'un des principaux défis liés aux approches antérieures de traitement était que l'étape de documentation des enquêtes n'était pas menée tant que toutes les activités de traitement n'étaient pas complètes. Cela se produit en dépit du fait qu'une part importante des renseignements requis pour préparer la documentation d'enquête fait partie intégrante des processus opérationnels liés à la tâche de traitement des données d'enquête recueillies. Cette approche a comme résultat final que la documentation des enquêtes est très inefficace et que la préparation de la documentation d'enquête nécessite le retour aux étapes de traitement antérieures pour extraire et assimiler les données d'enquête pertinentes.

Pour mieux expliquer le concept, on peut observer que la création de spécifications de questionnaire pour le développement d'un instrument de collecte représente essentiellement l'étape initiale du traitement d'une enquête. Les informations requises pour définir un questionnaire comprennent le texte des questions, les codes de réponse (1 = homme, 2 = femme), les codes de non réponse (Sans objet, Ne sait pas, Refus, Non déclaré), l'enchaînement des questions et les champs de contrôle appliqués aux données. Ces renseignements servent à leur tour à préparer des clichés d'enregistrement pour les fichiers reçus des opérations sur le terrain. Grâce à ces renseignements de base, on peut préparer les entrées nécessaires pour procéder aux étapes ultérieures du traitement, une fois que les données commencent à être revenir du terrain. Ces données servent aussi d'informations minimales essentielles pour décrire les variables recueillies.

Une fois que les données recueillies commencent leur déplacement dans le système, les métadonnées sont utilisées pour définir les activités de traitement subséquentes liées à la vérification des données renvoyées, à l'application de contrôles de cohérence et à la correction ainsi que la création de variables dérivées. Tous ces renseignements servent ultimement comme renseignements supplémentaires pour décrire un élément d'information recueilli. Une fois le traitement terminé, le système peut extraire les éléments pertinents de métadonnées, à partir du développement du questionnaire, pour documenter pleinement un ensemble de données recueillies. Cela se fait généralement avec la production d'un dictionnaire de données ou d'un répertoire de codes exhaustif.

Pour arriver au scénario idéal d'entrée unique et de réutilisation au besoin, un système de traitement des métadonnées doit faciliter la saisie des métadonnées à l'étape appropriée du traitement et permettre à ces renseignements d'accompagner les données et d'être utilisés aux étapes de traitement subséquentes. La façon optimale d'y arriver est d'utiliser les métadonnées pour contrôler le traitement proprement dit des données d'enquête recueillies.

4. Pratiques exemplaires actuelles

L'élaboration d'une vision globale pour un système de traitement générique à l'intérieur du Secteur de la statistique sociale, de la santé et du travail nécessitait un examen initial des pratiques exemplaires existantes et des outils des secteurs touchés par le traitement des enquêtes auprès des ménages. Cet examen a fait ressortir deux secteurs des systèmes et approches de traitement qui semblent très prometteurs du point de vue de l'élaboration de l'environnement de traitement générique.

La Division de la statistique de la santé (DSS) a élaboré, au moyen de Microsoft Access, un système exhaustif pour la création et la mise à jour de spécifications de questionnaires d'enquête liées à ses systèmes de traitement. Cela a servi en dernier ressort de prototype pour le développement d'un système de traitement axé sur les métadonnées.

La Division des enquêtes spéciales (DES) a élaboré le Système de traitement généralisé (STG), qui comprend un ensemble de modules et de macros de traitement en SAS, au moyen de modèles d'entrée en Excel, appuyés par une structure de répertoire commun. Il a été démontré que ce système peut traiter une gamme variée d'instruments

d'enquête et de modes de collecte. À cette fin, il possède des avantages démontrables du point de vue de la représentation d'une méthode de traitement très robuste et efficace.

En combinant l'approche axée sur les métadonnées élaborée par la DSS et la méthode de traitement efficace présente dans le STG, l'équipe de développement disposait d'une base solide pour entreprendre le développement d'un ensemble générique d'outils de traitement axés sur les métadonnées.

5. Vue d'ensemble du système

5.1 Vue d'ensemble du système générique

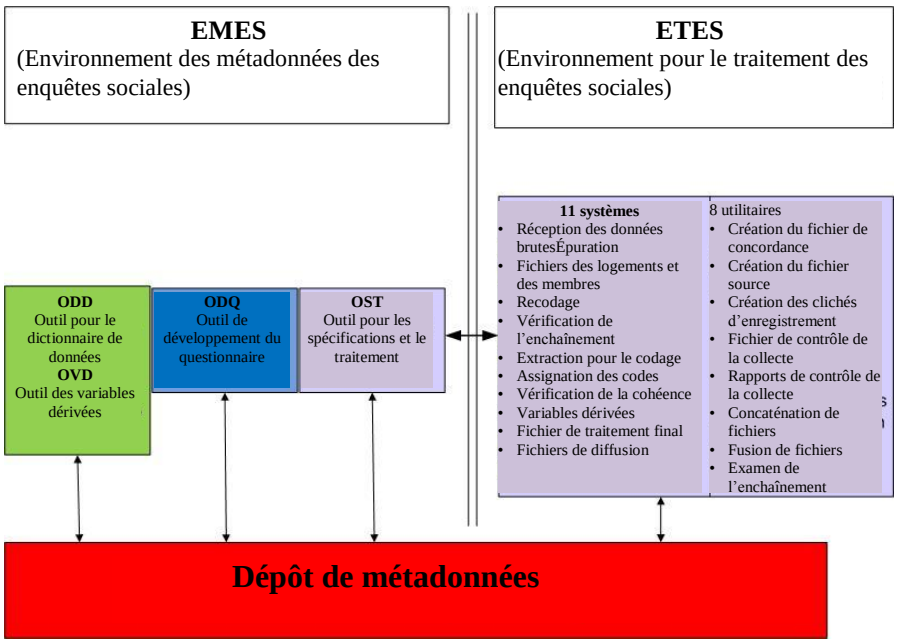
La figure 5.1 1 fournit une représentation schématique de l'ensemble d'outils envisagés pour appuyer le cycle de vie des enquêtes. Le dépôt de métadonnées constitue le fondement du système, c'est à dire la base de données dans laquelle toutes les métadonnées sont entreposées, et il est utilisé par les divers outils de soutien pour accéder aux données requises pour le traitement des enquêtes et les entreposer. Ce dépôt est appuyé par des programmes communs utilisés pour gérer les métadonnées et assurer l'uniformité interne de la base de données.

L'Environnement des métadonnées des enquêtes sociales (EMES) comprend un ensemble d'outils utilisés pour élaborer et manipuler les métadonnées d'enquête, notamment :

- l'outil de développement du questionnaire (ODQ), utilisé pour développer le questionnaire ;
- l'outil pour le dictionnaire de données (ODD), utilisé pour créer des répertoires de codes des enquêtes ;
- l'outil des variables dérivées (OVD), utilisé pour gérer et documenter la création des variables dérivées ; et
- l'outil pour les spécifications et le traitement (OST), qui sert à relier l'environnement de métadonnées et le système de traitement.

L'Environnement pour le traitement des enquêtes sociales (ETES) est au cœur du système de traitement. Dans sa forme actuelle, il est constitué d'un ensemble de 11 étapes de traitement individuelles associées à des tâches de traitement particulières et de huit utilitaires de soutien qui peuvent être utilisés par tous les systèmes à l'intérieur de l'ETES.

Figure 5.1-1
Aperçu des systèmes de l'environnement de traitement générique



5.2 Environnement de traitement des enquêtes sociales

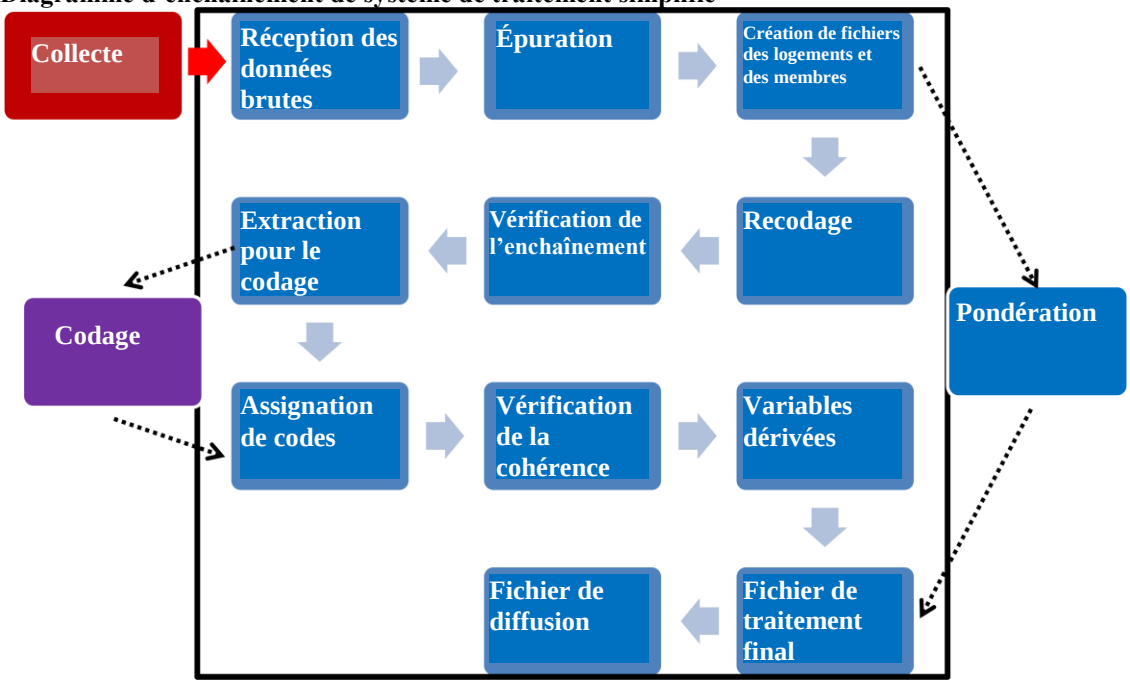
La figure 5.2 1 fournit un schéma de base d'un modèle de traitement simplifié, qui utilise l'ensemble des modules compris dans l'ETES. Même si la description de chaque étape du traitement dépasse la portée du présent article, plusieurs aspects clés du système de traitement justifient des explications plus poussées.

Tout d'abord, le système est conçu de façon à ce que les produits d'une étape de traitement servent d'entrée pour les étapes subséquentes. Cette approche permet d'assurer l'uniformité entre les étapes de traitement, ainsi qu'une vérification cohérente des erreurs, d'une étape à l'autre.

En deuxième lieu, le système est conçu pour que l'intégration d'une étape dans la suivante puisse être uniformisée sur la base des besoins de traitement individuels, et que chaque étape puisse être exécutée plus d'une fois, selon les besoins. Par exemple, dans certaines circonstances, il peut être avantageux d'exécuter les vérifications de l'enchaînement avant de passer aux étapes du codage, puis de reprendre l'étape de vérification de l'enchaînement pour s'assurer que le codage n'a pas eu de répercussions sur l'intégrité des enchaînements du questionnaire.

Enfin, le système est conçu pour permettre les mouvements de données à l'intérieur de l'ETES, en vue de l'utilisation d'autres systèmes, comme ceux servant à la pondération ou au codage des enquêtes. À long terme, cette approche permettra une intégration harmonieuse et efficace des systèmes génériques nouveaux et améliorés, au fur et à mesure de leur évolution à Statistique Canada. À court et à moyen termes, cela permettra aux secteurs d'enquête individuels de faire graduellement la transition au nouvel environnement et d'utiliser les outils existants qui ne faisaient pas partie de l'ensemble initial d'outils génériques.

Figure 5.2-1
Diagramme d'enchaînement de système de traitement simplifié



6. Avantages méthodologiques

De nombreux avantages méthodologiques sont liés à l'utilisation de l'ensemble de systèmes génériques de l'ETES.

Le système appuie l'utilisation d'un processus opérationnel commun, ce qui facilite l'établissement de pratiques exemplaires et les comparaisons entre les secteurs d'enquête. Cela permet à tous les secteurs d'adopter des approches

révisées et mises à jour, créées à l'intérieur des secteurs spécialisés individuels. Par exemple, la création de spécifications de contrôle pour le traitement des données sur le revenu des ménages par la Division de la statistique du revenu peut facilement être mise à la disposition de tous les secteurs d'enquête, assurant ainsi un traitement uniforme des variables de revenu dans tous les programmes d'enquêtes sociales.

La pratique qui consiste à exécuter une tâche seulement par étape de traitement particulière permet une approche plus ciblée et réduit les possibilités d'erreur. Cela a pour résultat final que, lorsque les données sont traitées correctement à chaque étape individuelle, avant de passer à la suivante, les secteurs de traitement consacrent moins de temps à réviser les étapes de traitement précédentes pour chercher des erreurs auxquelles ils pourraient se heurter aux étapes ultérieures du traitement.

Le système est essentiellement autodocumenté et utilise une structure de dépôt commune et des modèles uniformes d'entrée. Cette approche facilitera les mouvements de personnel entre les secteurs d'enquête et réduira les besoins de formation, permettant une plus grande mobilité du personnel. Cela permettra en outre l'application uniforme des étapes de traitement au fil du temps, ce qui donnera lieu à des produits uniformes par les secteurs d'enquête pour les cycles subséquents d'une enquête donnée comportant un contenu commun.

Le système utilise aussi de façon exhaustive les rapports automatisés des erreurs, ce qui simplifie la tâche de contrôle de la qualité des données et assure l'uniformité des produits. Cela permettra aux secteurs d'enquête de consacrer moins de temps au traitement et plus de temps à l'analyse et à la diffusion. Cela permettra en outre au personnel moins expérimenté de mener à bien une étape de traitement donnée, contribuant ainsi à améliorer l'efficacité opérationnelle globale des activités de traitement à Statistique Canada.

7. Défis en matière de développement

Le développement de l'environnement de traitement générique pose de nombreux défis. Au cours du processus d'établissement des exigences, des efforts considérables sont requis pour comprendre les besoins des secteurs individuels et déterminer si les différences sont le résultat d'une terminologie différente ou, en fait, de variations dans les exigences. Par exemple, après une analyse plus poussée, les termes épuration et validation utilisés pour décrire les étapes de traitement par deux secteurs de traitement différents se sont révélés représenter la même activité de traitement.

Au moment de l'élaboration des règles opérationnelles comprises dans le système, l'équipe de développement a continuellement dû établir un équilibre entre la nécessité d'appliquer les normes et d'augmenter la complexité du système, et la nécessité de maintenir la simplicité d'utilisation et d'entretien du système. La gestion de cet équilibre a nécessité l'évaluation des coûts-avantages de l'ajout d'autres fonctions et niveaux de complexité au système. L'équipe a dû aussi s'assurer que les règles opérationnelles appliquées aux étapes initiales du développement n'avaient pas d'influence négative sur sa capacité de mettre en œuvre efficacement les nouvelles règles opérationnelles aux étapes ultérieures du développement.

Enfin, une fois que le système a démontré son efficacité auprès de la collectivité des utilisateurs, l'équipe de développement a dû accorder beaucoup d'attention à la gestion du dépassement des paramètres. Cela était nécessaire, non seulement pour gérer les attentes des utilisateurs, mais aussi pour s'assurer que les améliorations prévues comportaient un ordre de priorité approprié. Essentiellement, le système a été la victime de son propre succès, et l'enthousiasme qu'il a suscité dans les collectivités d'utilisateurs a rendu nécessaires une gestion stricte et l'établissement d'un ordre de priorité pour les nouvelles améliorations, afin d'atteindre les objectifs de développement originaux en respectant l'enveloppe budgétaire prévue.

8. Conclusion

Même si le développement se poursuit, des progrès considérables ont été réalisés dans le Secteur de la statistique sociale, de la santé et du travail de Statistique Canada en vue de créer des processus et des systèmes réellement génériques à l'appui du cycle de vie des enquêtes. Le système a déjà commencé à faire ses preuves du point de vue

des économies réalisées au chapitre du développement des questionnaires et de la capacité de Statistique Canada d'intégrer davantage les systèmes de traitement dans les différents secteurs d'activité.

Les efforts pour l'avenir consisteront principalement à aider les secteurs d'enquête à procéder à une transition réussie et à mettre en œuvre les nouveaux outils dans leurs programmes respectifs. À l'avenir, le développement sera élargi, en vue d'inclure d'autres modes de collecte, outre les interviews téléphoniques et les interviews sur place assistées par ordinateur, comme les questionnaires électroniques et papier.

Bibliographie

Rapport interne (2005), « Household Surveys Processing Working Group, Generic Household Surveys Processing, Initial Report », rapport non publié, Ottawa, Canada, Statistique Canada.

Rapport interne (2009), « Rapport du Groupe de travail sur l'architecture opérationnelle du Bureau », rapport non publié, Ottawa, Canada, Statistique Canada.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Mise en œuvre de changements ou d'améliorations méthodologiques dans un système de traitement normalisé, ou comment un groupe consultatif peut-il faciliter le changement?

Katherine J. Thompson[1]

Résumé

Le présent article décrit le contexte de la création du groupe consultatif sur la méthodologie du StEPS (StEPS Methodology Advisory Group ou SMAG), une équipe permanente de méthodologistes prodiguant un soutien au système standard de traitement des enquêtes économiques (Standard Economic Processing System ou StEPS) du U.S. Census Bureau. L'exposé porte principalement sur les procédures opérationnelles, afin de démontrer comment ce groupe interdivisionnaire de statisticiens a élaboré et supervisé la mise en œuvre de plusieurs améliorations majeures du système existant, et survole brièvement le rôle du groupe dans l'établissement des exigences relatives au système remanié StEPS II.

Mots clés : Groupe consultatif ; exigences techniques ; améliorations méthodologiques.

1. Introduction

En mai 1995, la Direction économique du U.S. Census Bureau a entrepris le développement d'un système standard de traitement des enquêtes économiques (StEPS) destiné à être utilisé pour les activités de traitement postérieures à l'échantillonnage, y compris la collecte des données, le traitement postérieur à la collecte et la totalisation. Au cours de la dernière décennie, plus de 100 enquêtes mensuelles, trimestrielles et annuelles ont utilisé ce système de traitement (Ahmed et Tasky, 2001). Ces enquêtes couvrent plusieurs secteurs de l'économie américaine, dont la construction, le commerce de détail, le transport, le commerce de gros, les services et la fabrication.

Le développement d'un système de traitement unique pour servir une telle variété de programmes pose une foule de défis. Les unités d'échantillonnage diffèrent. Par exemple, l'enquête sur la construction repose sur l'échantillonnage des permis de bâtir des logements résidentiels, l'enquête annuelle sur le commerce de détail s'appui sur l'échantillonnage de sociétés ou d'entités fiscales, et l'enquête trimestrielle sur l'utilisation de la capacité des usines échantillonne des usines de fabrication. Le nombre d'éléments pour lesquels des données sont recueillies par questionnaire varie, de même que le nombre maximal prévu d'éléments par questionnaire. Dans certains cas, le programme requiert que l'unité d'échantillonnage fournisse une liste de sous-unités, qui fournissent chacune un ensemble complet de données. Les besoins en matière de collecte des données diffèrent, allant de l'envoi et du retour du questionnaire par la poste à l'interview sur place en passant par la collecte par la poste, par télécopieur ou en ligne. Les programmes doivent parfois faire la distinction entre les unités déclarantes (établies pour la collecte des données par l'unité d'échantillonnage) et les unités de totalisation (établies pour l'estimation par les gestionnaires du programme). Dans le cas des enquêtes-entreprises, la composition de l'unité d'échantillonnage peut évoluer au cours du temps en raison de fusions, de cessions ou d'acquisitions. Et naturellement, le traitement méthodologique des données attribuées aux unités de totalisation, à savoir la vérification, l'imputation, la pondération, la correction des valeurs aberrantes, l'estimation et l'estimation de la variance, diffère également. Malgré ces défis, un système de traitement consolidé offre de nombreux avantages. L'utilisation d'un même système facilite le partage des connaissances entre les divers secteurs et réduit le besoin de personnel spécialisé. Il est plus facile de maintenir un ensemble centralisé de programmes que plusieurs systèmes de traitement distincts. Plusieurs programmes peuvent bénéficier d'une seule amélioration du traitement.

Trois enquêtes utilisaient le StEPS au moment de son lancement en 1998. Depuis, plus de 100 enquêtes utilisent ou ont utilisé ce système; nous prévoyons à l'heure actuelle que 23 enquêtes utiliseront le StEPS durant le prochain

[1]Katherine J. Thompson, Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau, Washington, DC 20233 (Katherine.J.Thompson@census.gov).

exercice. Chaque nouveau transfert d'enquête a facilité la révision des modules existants du StEPS. Les révisions qu'il est proposé d'apporter au StEPS sont approuvées par un comité de contrôle du changement (Change Control Board ou CCB) constitué de spécialistes du domaine. Les changements qu'il est proposé d'apporter aux processus de collecte et de vérification des données ont souvent peu d'incidence sur les autres processus d'enquête. En revanche, cela est rarement le cas pour les changements qu'il est proposé d'apporter aux modules postérieurs à la collecte, c'est-à-dire ceux du contrôle, de l'imputation, de l'examen et de la correction interactifs des données, de l'estimation, de l'estimation de la variance et de la prévention de la divulgation. Par conséquent, en 2006, la direction a créé le groupe consultatif sur la méthodologie du StEPS (StEPS Methodology Advisory Group ou SMAG), dont l'objectif est triple :

1. Examiner les changements qu'il est proposé d'apporter au StEPS dans une perspective méthodologique ;

2. Recommander des modifications du StEPS en vue d'améliorer la méthodologie, et élaborer et valider des exigences non fonctionnelles (techniques) détaillées relatives à la méthodologie ;

3. Élaborer des procédures normalisées dans une perspective de pratiques exemplaires tenant compte de toutes les exigences en matière de recherche et de méthodologie des procédures en question.

Le présent article fournit des renseignements contextuels sur la composition du SMAG, en mettant principalement l'accent sur les procédures opérationnelles. L'article ne décrit pas les fonctions du StEPS. À ce sujet, le lecteur est invité à consulter Ahmed et Tasky (2001), Sigman (2000) et Sigman (2001).

2. Évolution de l'administration du StEPS

Au sein de ses secteurs de programme, la Direction économique a établi des **divisions spéci**alisées ayant pour mission de produire des estimations exactes, efficaces et à jour pour des programmes économiques **spécifiques**. Leurs responsabilités comprennent la conception des enquêtes, la collecte et le traitement des données, la totalisation et la diffusion. Trois divisions spécialisées utilisent le StEPS, à savoir la Division de la statistique des entreprises (Company Statistics Division ou CSD), la Division de la fabrication et de la construction (Manufacturing and Construction Division ou MCD), et la Division de la statistique sur le secteur des services (Services Sector Statistics Division ou SSSD). Des **divisions de soutien** fournissent un soutien à l'échelle de la Direction dans des domaines de services particuliers. La Division de la planification et de la coordination économique (Economic Planning and Coordination Division ou EPCD) fournit un soutien au StEPS à l'échelle de la Direction et s'occupe du transfert des enquêtes, du contrôle des changements, de la mise à l'essai, de l'élaboration des exigences, de la formation et de la gestion de projets. La Division de la programmation économique (Economic Programming Division ou EPD) regroupe les programmeurs de la Direction. Enfin, le Bureau des méthodes statistiques et de la recherche pour les programmes économiques (Office of Statistical Methods and Research for Economic Programs ou OSMREP) soutient les programmes de la Direction en procédant à des études générales des méthodes statistiques, en collaborant à la mise en œuvre des méthodes recommandées pour les programmes courants, en offrant une formation technique et en donnant des conseils « d'expert » sur demande. Dans les divisions spécialisées, des statisticiens-mathématiciens fournissent un soutien en matière de recherche et de méthodologie aux programmes désignés. Par contre, les employés de l'OSMREP sont répartis par secteur spécialisé, tels que les méthodes d'enquête par sondage, la prévention de la divulgation et les méthodes concernant les séries chronologiques.

Au départ, le StEPS a été l'œuvre d'une petite équipe désignée de programmeurs, de spécialistes des domaines et de méthodologistes faisant partie de l'EPCD. Le processus de développement était souple, et la portée du projet était flexible. Initialement, les modifications des modules du StEPS étaient intégrées selon la demande en suivant le flux. Mais à mesure que le nombre de programmes utilisant le StEPS a augmenté, cette approche réactive est devenue impossible. La première version d'un processus officiel de contrôle du changement appliqué au StEPS a été lancée en 1998. Elle comportait l'entrée de toutes les demandes de changement dans le système de solution (Remedy system) du Census Bureau et la liste complète de ces demandes était tenue à jour par l'EPCD. Une fois par mois, le comité d'examen des utilisateurs – un groupe de gestionnaires de programme et de programmeurs chargés du développement du StEPS – se réunissait pour établir l'ordre de priorité des demandes de changement et faire le suivi de leur résolution.

En 2001, 90 enquêtes utilisaient le StEPS (Ahmed et Tasky, 2001). Des améliorations du StEPS étaient généralement produites au moment du transfert des enquêtes; des corrections des procédures du StEPS étaient introduites simultanément. Entre-temps, les utilisateurs ont établi des « groupes d'utilisateurs » spécialisés : le groupe d'utilisateurs du StEPS a été créé en 2000; le groupe d'utilisateurs de l'estimation du StEPS a été créé en 2001 et le groupe d'utilisateur de l'imputation du StEPS a été créé en 2004. Ces groupes offraient un forum pour la discussion et la résolution de problèmes de traitement et avaient tendance à se concentrer sur l'usage et l'amélioration des logiciels et modules existants. En 2004, le comité de contrôle du changement (Change Control Board ou CCB) du StEPS a été établi en vue de mettre en œuvre un processus officiel de contrôle du changement. Le CCB du StEPS est présidé et coordonné par l'EPCD. Les membres sont choisis par nomination et chaque division compte un nombre fixe de représentants.

Le comité de gouvernance du StEPS (StEPS Governing Board ou SGB) a été créé en novembre 2006 pour appliquer une approche systématique représentative à l'encadrement et à l'orientation du StEPS (Russell, 2011). Le SGB comprend des gestionnaires de programme de haut niveau (chefs divisionnaires adjoints). Le SGB fournit un encadrement de haut niveau et une surveillance du développement et de la maintenance du StEPS, du système existant et des futures versions, détermine les plans et les priorités en ce qui concerne les améliorations du StEPS et résout les problèmes et élabore les politiques en vue d'améliorer les opérations du StEPS. En créant le SGB, la direction économique a désigné un « propriétaire » pour ce système de traitement à très grande échelle et pour les projets connexes. Dans la perspective de contrôle du changement, cela a mis un terme aux problèmes de résolution des demandes de changement contestées. Le SGB est l'entité qui assume la responsabilité en fin de compte. Le SGB a besoin de documentation et d'études pour prendre toutes les grandes décisions. Les plans annuels des opérations décrivent dans les grandes lignes les projets et fournissent une justification pour les projets approuvés. En outre, la présentation des demandes de changement contestées au SGB doit être accompagnée de données probantes à l'appui ou à l'encontre de la demande de changement (par exemple avantage technique, nombre de programmes concernés, exigences-effort de développement et de programmation).

L'établissement du SGB a mené indirectement à la création du SMAG. En général, les membres du CCB du StEPS sont des spécialistes des domaines et non des méthodologistes. Dans ce forum, il n'était pas facile d'aborder les demandes de changement en rapport avec la méthodologie. Chaque demande de changement nécessite une justification minutieuse et le CCB doit déterminer le niveau d'effort nécessaire pour mettre en œuvre la demande. Conscient de l'« effet de ricochet » que peut avoir la modification d'un module général qui s'applique à plusieurs programmes, personne ne voulait « approuver sans discussion » les demandes de changement méthodologiques. Les membres du CCB du StEPS ne possédant pas les compétences requises, par conséquent, en 2006 le directeur adjoint des programmes économiques a créé un groupe consultatif sur la méthodologie du StEPS (SMAG) distinct. Toutes les divisions qui utilisent le StEPS pour le traitement des données sont représentées au sein de ce groupe. Le tableau 2-1 énumère les responsabilités des membres du SMAG. En partenariat avec le CCB du StEPS, le SMAG examine tout changement que l'on propose d'apporter au StEPS lorsqu'il comporte une modification des calculs ou de la méthodologie, et fait des recommandations tout en fournissant la documentation à l'appui. Les représentants sont tenus de consulter les méthodologistes appropriés et d'informer les spécialistes du domaine des conséquences éventuelles des changements ou améliorations méthodologiques apportés au StEPS pour les programmes de leur division. Les membres du SMAG sont nommés par les chefs divisionnaires adjoints respectifs et sont des gestionnaires ou sont autorisés par les gestionnaires à prendre des décisions concernant les sujets à l'étude, et se situent généralement à un niveau hiérarchique de supervision ou de cadre intermédiaire.

En tant que comité consultatif, le SMAG élabore des « pratiques exemplaires ». Au moment de sa création en 2006, le SMAG comprenait des représentants des divisions spécialisées qui n'utilisaient pas le StEPS pour le traitement de leurs données. Au cours du temps, ces divisions ont choisi de ne pas participer aux activités du SMAG : une fois qu'une pratique exemplaire est établie, le SMAG se concentre sur sa mise en œuvre dans le StEPS. La documentation du SMAG est mise à la disposition de la Direction, et d'autres divisions peuvent, et ont, participé à des sous-groupes du SMAG quand le sujet est d'intérêt général.

Tableau 2-1
Composition du SMAG

Division	Responsabilités
OSMREP	• Président • Fonctions administratives • Fournit les chefs des sous-groupes et des groupes d'utilisateurs
CSD, MCD, SSSD	• Fournissent l'expérience du domaine spécialisé en ce qui concerne l'application des méthodes aux programmes permanents.
EPCD et EPD	• Assurent la liaison avec leurs divisions et avec le CCB. • Fournissent des ressources pour l'élaboration des exigences et l'approbation de la conception.

3. Procédures opérationnelles du SMAG

La charte du SMAG établit la raison d'être, la portée des activités, les exigences en matière de ressources et les hypothèses opérationnelles pour le SMAG; les règlements administratifs du SMAG décrivent les procédures opérationnelles de ce dernier et comprennent des renseignements détaillés sur les membres, les comités et sous-équipes, les réunions, le vote et la documentation. Même si le SMAG a établi des procédures opérationnelles officielles, il est important de souligner que celles-ci constituent le squelette. L'étoffe du groupe consultatif, c'est-à-dire la façon dont fonctionne le SMAG, est considérablement plus dynamique. Comme les membres sont nommés, la composition du SMAG a tendance à être constante. Surtout, les membres du SMAG partagent une même vision. L'objectif du SMAG est toujours de mettre en œuvre une méthodologie qui profite à tous les secteurs et qui ne nuit à aucun d'eux. Nous nous efforçons d'obtenir l'approbation unanime des représentants des divisions pour les méthodes proposées. La discussion de groupe est favorisée quel que soit le sujet, les courriels de groupe n'étant utilisés que comme un complément et non comme un remplacement de la discussion. Cela dit, le temps des discussions de groupe est limité étant donné la composition du SMAG.

Les procédures opérationnelles qui ont été codifiées ont évolué au cours du temps par essais et erreurs. Le SMAG utilise trois forums pour discuter des projets, à savoir la prise de décision de groupe, les sous-groupes et les comités/groupes d'utilisateurs.

3.1 Processus de prise de décisions de groupe

Le processus de prise de décisions de groupe s'appuie sur des réunions planifiées pour élaborer et approuver tous les produits livrables. En général, un animateur dirige la discussion de groupe, et une personne est désignée pour en faire le compte rendu. Les tâches assignées en dehors de la réunion sont limitées et se résument généralement à recueillir de l'information, quoique l'animateur puisse avoir des tâches supplémentaires. Les exemples qui suivent illustrent le recours à ce processus.

Taux de réponse au niveau de l'unité Le SMAG s'est vu confier son premier projet par l'Office of Management and Budget (OMB) Standards and Guidelines for Statistical Surveys (2006), qui requiert que les programmes permanents produisent et publient des taux de réponse en utilisant des formules normalisées. Partant de la documentation provisoire existante, le SMAG était chargé :

- D'établir une définition des « répondants » reconnue à l'échelle de la Direction. Cet exercice incluait la supervision de la révision des définitions des répondants au niveau des programmes et la détermination de la manière de mettre ces définitions en application dans le StEPS ;

- D'examiner et de modifier les règles de signalisation du StEPS au niveau de la question et au niveau de l'unité ;

- De rédiger des exigences pour tous les calculs et de collaborer à l'élaboration, à la mise à l'essai et à la documentation de ces exigences ;

- D’élaborer et de prodiguer une formation à l’échelle de la Direction ;

- D’élaborer des cas d’utilisation et de mettre à l’essai le code du StEPS.

L’achèvement de ce projet a demandé environ cinq mois de réunions hebdomadaires. En outre, de petites équipes des membres du SMAG se sont réunies en dehors de ces réunions pour rédiger des propositions et le président du SMAG a collaboré directement avec les programmeurs chargés du développement du StEPS à la mise en œuvre et à la mise à l’essai des nouvelles règles de signalisation au niveau de l’unité et au niveau de la question.

Taux pondérés de réponse au niveau de l’élément de données (taux de réponse pour la quantité totale et taux de réponse pour la quantité). Ces taux de réponse requis au niveau de l’élément de données mesurent la proportion pondérée d’une estimation clé reposant sur des données déclarées par les unités répondantes et sur des sources de qualité équivalente; les taux de réponse pour la quantité totale permettent d’utiliser les deux types de données dans le numérateur, tandis que les calculs des taux de réponse pour la quantité sont restreints aux données obtenues directement auprès des répondants. Les responsabilités du SMAG en ce qui concerne l’élaboration de ces mesures étaient presque identiques à celles énoncées pour les taux de réponse au niveau de l’unité. Cependant, des aspects techniques supplémentaires ont dû être pris en considération, tels que la détermination des ajustements de la pondération qui devraient être inclus dans les calculs, l’élaboration de règles pour repérer les éléments de données « admissibles », et la détermination du traitement approprié des éléments à valeur réelle dans les calculs. Par conséquent, dans le cadre du processus d’élaboration des exigences, le SMAG a organisé une série de séminaires à l’échelle de la Direction afin d’obtenir des renseignements auprès des divisions spécialisées. L’achèvement de ce projet a demandé environ six mois.

Variation en pourcentage historique (niveau de l’unité). Un programme d’enquête sur des indicateurs souhaitait pouvoir connaître la variation en pourcentage pour un élément de données particulier au niveau de l’unité d’échantillonnage dans les examens et les corrections. Cette mesure est assez utile pour les éléments de données à valeur positive comme les ventes ou l’emploi, mais ne l’est pas pour les éléments de données à valeur réelle comme le revenu, et elle ne peut pas être calculée pour des accroissements ou des décroissements par rapport à zéro. Le SMAG a élaboré des exigences détaillées pour le calcul et l’affichage de cette mesure. La mesure mise en œuvre est à la disposition de tous les programmes qui utilisent le StEPS.

Limites robustes (Resistant Fences). Bechtel (2011) décrit comment le SMAG a établi les exigences concernant la méthode de détection des valeurs aberrantes au moyen des limites robustes et la mise en application dans le StEPS.

Les avantages d’un processus entièrement démocratique sont évidents. La discussion ouverte aboutit à des exigences entièrement approuvées qui tiennent compte des points de vue de toutes les divisions. Le processus de prise de décisions par consensus aide le SMAG à formuler des recommandations unanimes. Pour les projets décrits plus haut, toutes les recommandations ont été adoptées par le CCB du StEPS, dont la majorité des discussions ont porté sur le moment de la mise en œuvre et l’affectation des ressources pour cette dernière.

Le processus de prise de décisions en groupe est approfondi, mais il peut être lent, surtout quand les spécialistes du domaine ne sont pas dans la même pièce. Il peut aussi « paralyser » les réunions quand les sujets abordés sont limités à un domaine d’expertise que ne partagent peut-être pas tous les membres du SMAG. Enfin, certains estiment que les exigences techniques ne devraient pas être établies sans exigences fonctionnelles (d’utilisation) connexes, mais cette dernière responsabilité a été affectée à une équipe distincte de spécialistes du domaine. Par conséquent, après avoir achevé le projet des limites robustes (Bechtel, 2011), le SMAG a décidé de créer des « sous-groupes » (sous-équipes) pour les projets à court terme.

3.2 Sous-groupes/sous-équipes

Les sous-groupes ou sous-équipes du SMAG sont des comités qui sont créés pour des projets uniques dont le calendrier est établi. Il s’agit de projets approuvés comportant des jalons précis et des produits livrables préapprouvés. Les membres d’un sous-groupe du SMAG sont choisis au cas par cas par les divisions qui parrainent le projet et la participation à un sous-groupe n’est pas limitée aux membres du SMAG, quoique l’OSMREP fournisse toujours le président du sous-groupe et exécute toutes les fonctions de soutien administratif. Les sous-groupes du SMAG se réunissent indépendamment et produisent régulièrement des rapports d’étape à l’intention du SMAG. Les

recommandations du sous-groupe doivent être approuvées par le SMAG avant d'être transmises en vue d'un suivi ultérieur. Les exemples qui suivent illustrent l'utilisation de sous-équipes par le SMAG.

Estimation par rapport en chaîne (link relative estimator). L'estimateur par rapport en chaîne combine une valeur repère obtenue périodiquement avec une estimation par sondage restreinte de la variation pour produire une estimation pour la période courante (Madow et Madow, 1978). Cet estimateur est utilisé pour deux indicateurs économiques : le programme de la fabrication, des livraisons et des stocks (Manufacturing, Shipments and Inventories ou M3), qui s'appuie sur un échantillon non probabiliste, et l'enquête mensuelle sur le commerce de détail (Advance Monthly Retail Trade Survey ou MARTS), qui s'appuie sur un échantillon probabiliste. Le programme MARTS a été transféré au StEPS en 2010 et le programme M3, en 2003. Par conséquent, il existe dans le StEPS des programmes de saisie des données et d'estimation nécessaires pour le traitement des données de l'enquête MARTS. Une équipe constituée de deux membres de l'OSMREP et de spécialistes du domaine de chaque programme (méthodologistes et gestionnaires de programme) s'est réunie pendant trois mois pour élaborer les exigences complètes pour l'estimation des rapports de mois à mois, l'estimation de la variance, ainsi que l'examen et la correction des données d'entrée. L'équipe a rédigé les exigences techniques pour les modifications devant être apportées au logiciel existant, a exécuté le codage et les mises à l'essai, et élaboré la formation à l'intention des méthodologistes, des programmeurs et des analystes.

Équipe des tableaux de vérification. Les programmes permanents de la Direction économique doivent participer au programme de vérification de la qualité, qui a pour objectif de vérifier le respect des normes de l'OMB. Le but de ce projet approuvé était d'établir les exigences pour la production automatisée dans le StEPS de certains tableaux qui donnent des preuves du respect des normes de qualité de l'OMB. Pour cela, le SGB a mis sur pied une équipe constituée des trois membres de l'OSMREP, de trois méthodologistes provenant des CSD, MCD et SSSD, et de quatre spécialistes du domaine (provenant des mêmes divisions). Cette équipe s'est réunie toutes les deux semaines pendant un an. Elle avait pour objectif de rédiger des propositions de tableaux et de graphiques produits automatiquement. Chaque tableau proposé était assorti d'exigences de haut niveau pour la mise en œuvre dans le StEPS. Le SGB a appuyé la recommandation finale de l'équipe, mais l'inclusion des tableaux et des graphiques a été reportée jusqu'à l'achèvement du remaniement du StEPS.

Équipe de la base d'imputation. En réponse à plusieurs demandes de changement provenant de diverses sources en vue de modifier le module d'imputation du StEPS, le SMAG a formé une équipe constituée de méthodologistes spécialisés afin de décrire les pratiques exemplaires pour la création d'une base d'imputation ou la sélection des données d'origine pour les méthodes d'imputation disponibles dans le StEPS. Ce groupe s'est réuni pendant 14 mois. Les recommandations de l'équipe ont été présentées à la Direction en janvier 2010 et le rapport officiel a été diffusé le 20 avril 2010.

En général, le recours à des sous-groupes spéciaux s'est avéré assez fructueux. Ces groupes offrent des possibilités de perfectionnement professionnel à des personnes non membres du SMAG et garantissent que toutes les parties prenantes soient bien représentées. Les membres sélectionnés pour faire partie de l'équipe reconnaissent qu'ils travaillent sur un projet de court terme dont les obligations sont clairement définies et dont la portée n'est pas négociable. En outre, les membres du sous-groupe ont un intérêt direct pour le résultat spécifique du projet. Néanmoins, la gestion du projet peut poser des défis. L'OSMREP fournit le président du sous-groupe. Les membres de l'équipe proviennent d'autres divisions et ont des tâches de production concurrentes. La « gestion matricielle » est la règle et non l'exception, et le chef d'équipe doit assigner les tâches avec beaucoup de prudence afin de s'assurer que le projet soit achevé dans les délais établis.

3.3 Groupes d'utilisateurs

Les groupes d'utilisateurs du SMAG sont des comités « permanents » ayant un mandat thématique unique. Les groupes d'utilisateurs du SMAG sont au nombre de deux, à savoir le groupe des utilisateurs de l'imputation générale et le groupe consultatif sur la méthodologie de l'entrepôt de données analytiques sur les séries chronologiques (Time Series Analytical Repository ou TSAR). Les membres de ces deux groupes sont choisis par les divisions qui les parrainent. Les présidents des groupes d'utilisateurs appartiennent à l'OSMREP et produisent des rapports réguliers à l'intention du SMAG. Comme le SMAG, les deux groupes d'utilisateurs sont dotés d'une charte de projet; les règlements administratifs sont facultatifs. Les groupes d'utilisateurs sont chargés d'examiner toutes les demandes de changement en rapport avec le thème. Si le groupe d'utilisateurs recommande de procéder au changement, il est

également chargé d'élaborer les exigences techniques. Les groupes d'utilisateurs proposent également l'apport au StEPS d'améliorations méthodologiques en rapport avec leur thème et sont en partie responsables des essais, y compris de l'élaboration de cas de mise à l'essai et de scénarios de mise à l'essai, ainsi que l'exécution d'essais d'utilisation de versions améliorées ou nouvelles du code du StEPS.

Comme les sous-groupes du SMAG, les groupes d'utilisateurs offrent des avantages professionnels à leurs membres. Mais surtout, les produits livrés par les comités profitent à la Direction car elle est certaine que les recommandations sont formulées par des « experts ». Les contraintes en vue d'être accepté comme membre d'un comité sont moins importantes que pour être membres du SMAG ou de l'un de ses sous-groupes; à condition qu'une personne ait l'approbation de sa division, elle peut participer aux réunions des comités. Les présidents de ces comités désignés par l'OSMREP doivent relever les défis que suscite la « gestion matricielle » pour attribuer les tâches. En outre, certaines questions se posent quant à l'existence de membres de réserve, puisque ces comités sont permanents, mais que les employés ne le sont pas.

4. Conclusion

Le SMAG a été créé en 2006 sous l'autorité du directeur adjoint des programmes économiques avec l'appui total et consensuel des divisions clientes. À l'époque, le SMAG comblait un vide, en offrant un forum officiel pour l'approbation des demandes de changement. Les employés désignés pour être membres du SMAG ont également dû faire face à une somme assez importante de travail supplémentaire. Arriver à gérer cette charge de travail de façon que les projets soient exécutés dans un délai raisonnable sans surcharger indûment les participants constitue un défi constant.

Le SMAG s'est avéré efficace pour de nombreuses raisons. En premier lieu, ce groupe consultatif a beaucoup bénéficié de l'appui de la haute direction, y compris le directeur adjoint, le SGB et les divisions clientes. En deuxième lieu, la structure organisationnelle de la Direction économique permet au SMAG de fonctionner sans heurts. L'OSMREP assure la stabilité organisationnelle du SMAG et de ses sous-équipes/comités en fournissant le président et en exécutant toutes les fonctions administratives. Cela permet aux autres membres du SMAG et aux participants affiliés au SMAG de restreindre leurs obligations à examiner la documentation pertinente et à fournir des commentaires. Enfin, le SMAG jouit d'une excellente dynamique de groupe. En 2009, le directeur adjoint des programmes économiques a autorisé le remaniement du StEPS. Ce remaniement comprendra une nouvelle structure de base de données et une nouvelle interface utilisateur. Bien que le nombre d'enquêtes dont les données sont traitées par le StEPS ne soit plus que de 23, il ne s'agit en rien d'un petit projet. La participation du SMAG au développement du nouveau système est, quant à elle, une constante.

Remerciements

L'article vise à informer les parties intéressées des résultats de recherche et à encourager la discussion. Les opinions exprimées sont celles de l'auteur et ne représentent pas forcément celles du U.S. Census Bureau. Je remercie Anne Russell et Xijian Liu de leur examen minutieux et de leurs commentaires constructifs au sujet de versions antérieures du présent manuscrit, ainsi que les membres du SMAG de leurs commentaires et suggestions concernant l'exposé.

Bibliographie

Ahmed, S.A. et D.L. Tasky (2001), « Are generalized systems the way of the future: A case study on the Standard Economic Processing System (StEPS)? », *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Bechtel, L.T. (2011), « Retro-fitting a simpler outlier detection procedure into a complex generalized system », *Proceeding of the 2011 Statistics Canada Symposium*.

Federal Register Notice (2006), OMB Standards and Guidelines for Statistical Surveys.

Madow, L.H. et W.G. Madow (1978), « On link relative estimation », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, p. 534 à 539.

Russell, A.S. (prévu pour 2011), « StEPS Governing Board Charter », unpublished U.S. Census Bureau mémoire, disponible sur demande.

Sigman, R.S. (2000), « Estimation and variance estimation in a standardized economic processing system », *Proceedings of the Second International Conference on Establishment Surveys*, p. 677 à 686.

Sigman, R.S. (2001), « Editing and imputation in a standard economic processing system », *Proceedings of Statistics Canada Symposium 2001*.

SÉANCE 10A

CADRES DE TRAVAIL

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Élaboration d'une méthodologie pour le Cadre canadien pour les statistiques culturelles

Mary K. Allen[1]

Résumé

Pendant au moins 20 ans, des efforts ont été déployés aux niveaux national et international, en vue d'élaborer des mesures uniformes pour les arts et la culture. Cela a mené au Cadre de l'UNESCO pour les statistiques culturelles, et au Canada, au Cadre canadien pour les statistiques culturelles de 2004. Ces cadres ont été élaborés pour répondre à des besoins statistiques particuliers de chercheurs dans le domaine de la culture, qui n'étaient pas en mesure d'obtenir des mesures comparables et uniformes à partir des données existantes fondées sur les systèmes de classifications types.

Le Cadre de 2004 a fourni une structure pour la mesure de la culture, mais était défini selon des systèmes de classification existants et dépendait de ces systèmes, comme le Système de classification des industries de l'Amérique du Nord (SCIAN) de 2002 et la Classification type des biens (CTB). Cela signifiait que, même s'il était utile pour la recherche, le Cadre se limitait aux données et aux normes existantes.

En 2011, le Cadre canadien pour les statistiques culturelles a été remanié selon une approche plus conceptuelle. Plutôt que de définir la culture du point de vue des classifications types existantes, le Cadre de 2011 définit la culture et ses composantes de façon conceptuelle et comporte une méthodologie pour guider la mise en correspondance des systèmes de classification existants et futurs avec le Cadre. Cette nouvelle méthodologie appuie la création de listes types de codes de classification pouvant être reproduits, à l'appui de la recherche sur la culture et comme guide pour l'élaboration future des données, y compris un examen des possibilités de combler les lacunes dans les données. On dispose ainsi d'une méthode type uniforme pour déterminer ce qui, par exemple, est un produit culturel ou une industrie culturelle.

Cet article présente la méthodologie élaborée pour le Cadre canadien pour les statistiques de la culture de 2011.

Mots clés : Culture ; cadre.

Bibliographie

Statistique Canada (2011), « Cadre conceptuel pour les statistiques de la culture 2011 », Cadre canadien pour les statistiques de la culture, n° 001, n° 87-542-X au catalogue de Statistique Canada, Ottawa, Ontario (Canada).

Statistique Canada (2011), « Guide de classification pour le Cadre canadien pour les statistiques de la culture 2011 », Cadre canadien pour les statistiques de la culture, n° 002, n° 87-542-X au catalogue de Statistique Canada, Ottawa, Ontario (Canada).

[1]Mary K. Allen, Statistique Canada, 100, promenade pré Tunney, Ottawa, Ontario (Canada), K1A 0T6 (mary.allen@statcan.gc.ca).

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Combien de Canadiens vivent dans une ville? Conceptualisation, définition et diffusion proposée de normes de rechange

Ray D. Bollman et Peter Murphy[1]

Résumé

Statistique Canada n'assemble pas ses données sur les villes dans un format permettant de répondre à la question : Combien de Canadiens vivent dans une ville? Les objectifs du présent article sont les suivants :

1. décrire trois moyens différents de définir une « ville » ;
2. présenter une série de données chronologiques sur la population du Canada vivant dans une « ville » pour a) chacune des trois définitions et pour b) divers seuils de taille de population possibles pour classer une localité comme étant une « ville » ;
3. discuter du niveau d'« urbanisation » qu'implique la part de la population du Canada vivant dans des villes de diverses tailles pour différentes définitions ;
4. montrer, pour chaque définition d'une ville, quelle partie de la distribution de l'urbanisation affiche la croissance la plus rapide (c'est-à-dire quelle catégorie de taille de ville croît, ou s'urbanise, au taux le plus élevé) ;
5. déterminer si ce taux d'urbanisation est attribuable à des facteurs démographiques (naissances, décès, immigration ou migration nette à l'intérieur du Canada) ou bien à une reclassification ou une fusion (c'est-à-dire déterminer si la localité a changé de catégorie de taille entre deux recensements et (ou) si les limites de la localité ont changé entre deux recensements).

Mots clés : Ville ; urbain ; rural ; population.

1. Introduction

Les utilisateurs des données posent à Statistique Canada la question : Combien de Canadiens vivent dans une ville? Il n'existe aucun moyen simple de trouver la réponse à cette question, car Statistique Canada ne produit et ne publie pas de données sur les niveaux de population et sur les taux d'urbanisation pour a) différentes définitions des villes et pour b) différents seuils de taille pour chaque définition d'une « ville ».

Les objectifs du présent article sont les suivants :

1. décrire trois moyens différents de définir une « ville » ;
2. présenter une série de données chronologiques sur la population du Canada vivant dans une « ville » pour a) chacune des trois définitions et pour b) divers seuils de taille de population possibles pour classer une localité comme étant une « ville » ;
3. discuter du niveau d'« urbanisation » qu'implique la part de la population du Canada vivant dans des villes de diverses tailles pour différentes définitions ;
4. montrer, pour chaque définition d'une ville, quelle partie de la distribution de l'urbanisation affiche la croissance la plus rapide (c'est-à-dire quelle catégorie de taille de ville croît, ou s'urbanise, au taux le plus élevé) ;
5. déterminer si ce taux d'urbanisation est dû à des facteurs démographiques (naissances, décès, immigration ou migration nette à l'intérieur du Canada) ou bien à une reclassification ou une fusion (c'est-à-dire déterminer si la localité a changé de catégorie de taille entre deux recensements et (ou) si les limites de la localité ont changé entre deux recensements).

[1]Ray Bollman (RayD.Bollman@sasktel.net) a pris sa retraite récemment en tant que chef, Groupe de la recherche rurale, Statistique Canada et Peter Murphy (Peter.Murphy@statcan.gc.ca) est chef, Division de la géographie, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6.

2. Trois moyens de conceptualiser une « ville »

Les démographes ont utilisé trois moyens différents de conceptualiser une « ville » [voir, par exemple, Puderer (2009)]. Ces trois mesures ont toujours pour point de départ la densité de population d'une localité, tandis que les limites sont déterminées de façons différentes. Les trois classifications sont présentées ci après.

2.1 Forme ou morphologie

Ce concept décrit la zone bâtie d'un établissement de population et peut être considéré simplement comme la « vue par le pare-brise ». Pendant votre promenade en voiture dominicale, à quel moment arrivez vous hors de la « ville »? On pourrait soutenir qu'il s'agit de la clientèle cible pour un planificateur du transport urbain.

2.2 Unité administrative

Ce concept représente la vue du maire. Combien de personnes vivent dans mon unité administrative? Du point de vue du résident, la question est : À qui mes impôts sont ils versés et quelle administration est responsable de fournir les services locaux? Durant votre promenade en voiture dominicale, vous ne pouvez habituellement pas voir les limites de la cité administrative – quoiqu'un panneau disant « Bienvenue dans notre ville de xxx xxx habitants » soit souvent installé à la limite.

2.3 Zone fonctionnelle

Ce concept est fondé sur la notion selon laquelle « nous sommes tous concernés ». Autrement dit, tant du point de vue des citoyens que de celui des investisseurs, tout le monde partage les résultats des bons projets de développement de même que ceux des mauvais projets de développement. En ce sens, la population agglomérée fonctionne à l'unisson comme une zone fonctionnelle, indépendamment de la forme des limites administratives.

3. Application de ces concepts à Statistique Canada

3.1 Forme ou morphologie

Statistique Canada délimite la forme ou morphologie d'une zone bâtie pour tous les **centres de population** possédant un noyau urbain dont la densité de population est égale ou supérieure à 400 habitants par km2 et dont la population totale est supérieure à 1 000 habitants. Anciennement, les centres de population étaient appelés « régions urbaines de recensement ». Pour des renseignements plus détaillés, voir Statistique Canada (2007). Une analyse comparable de la façon de définir et de délimiter les zones bâties a été publiée par Hofmann et coll. (2010a, 2010b). Leur méthodologie leur a permis d'affiner la délimitation de la quantité de terres dans les zones bâties, mais n'a produit qu'une correction relativement faible de la population délimitée dans les centres de population.

3.2 Unité administrative

Statistique Canada délimite une **subdivision de recensement (SDR)** pour chaque ville ou municipalité constituée en personne morale. En collaboration avec chaque province et territoire, les terres (et les résidents) qui ne font pas partie d'une municipalité constituée en personne morale sont délimitées en SDR à des fins statistiques. Pour plus de renseignements, voir Statistique Canada (2007).

3.3 Zone fonctionnelle

Statistique Canada délimite des régions métropolitaines de recensement (RMR) et des agglomérations de recensement (AR) pour tout noyau urbain de 10 000 habitants ou plus et englobe dans celles ci toutes les SDR voisines dans lesquelles 50 % ou plus des résidents ayant un emploi font la navette pour aller travailler dans le noyau urbain de la SDR. Donc, les taux de navettage sont utilisés pour mesurer ou pour approximer les zones entourant le noyau urbain qui « fonctionnent ensemble ».

4. Résultats

4.1 Centres de population représentant la forme ou la morphologie

En ce qui concerne le niveau d'urbanisation, trois centres de population (Montréal, Toronto et Vancouver) avaient une population de plus de 1 million d'habitants et représentaient 32 % de la population du Canada (tableau 4 1.1). En 2006, 895 centres de population avaient une population de 1 000 habitants ou plus.

Tableau 4.1-1
Population selon la taille du centre de population, Canada, 1991 à 2006

Catégorie de taille de population du centre de population[1]	Population totale et nombre d'établissements[1]												Distribution en pourcentage de la population								Variation en pourcentage de la population dans des limites constantes et une classification constante			
	Dans les limites et la classification de 1991			Dans les limites et la classification de 1996			Dans les limites et la classification de 2001			Dans les limites et la classification de 2006			Dans les limites et la classification de 1991		Dans les limites et la classification de 1996		Dans les limites et la classification de 2001		Dans les limites et la classification de 2006					
	1986	1991	Nbre de centres de population[1]	1991	1996	Nbre de centres de population[1]	1996[3]	2001	Nbre de centres de population[1]	2001	2006	Nbre de centres de population[1]	1986	1991	1991	1996	1996	2001	2001	2006	1986 à 1991	1991 à 1996	1996[3] à 2001	2001 à 2006
1 million et plus	7,158,005	7,865,789	3	7,962,741	8,539,938	3	…	9,412,027	3	9,372,715	10,022,987	3	28	29	29	30	…	31	31	32	9.9	7.2	…	6.9
500 000 à 999 999	3,623,967	3,905,298	6	3,934,353	4,079,970	6	…	4,369,921	6	4,378,914	4,660,213	6	14	14	14	14	…	15	15	15	7.8	3.7	…	6.4
100 000 à 499 999	2,705,621	2,888,861	15	3,135,825	3,274,783	17	…	3,728,333	20	3,774,438	3,973,453	20	11	11	11	11	…	12	13	13	6.8	4.4	…	5.3
50 000 à 99 999	1,345,360	1,479,768	21	1,561,942	1,662,590	24	…	1,481,831	22	1,549,340	1,653,109	23	5	5	6	6	…	5	5	5	10.0	6.4	…	6.7
30 000 à 49 999	956,787	1,015,237	26	957,602	992,847	26	…	966,319	26	1,113,989	1,197,050	31	4	4	4	3	…	3	4	4	6.1	3.7	…	7.5
10 000 à 29 999	1,445,804	1,515,559	93	1,480,773	1,555,786	97	…	1,597,682	100	1,496,398	1,563,022	100	6	6	5	5	…	5	5	5	4.8	5.1	…	4.5
5 000 à 9 999	820,029	866,470	127	897,370	946,935	136	…	946,880	133	935,749	960,734	136	3	3	3	3	…	3	3	3	5.7	5.5	…	2.7
2 500 à 4 999	725,056	745,143	209	720,444	748,677	210	…	786,829	222	694,921	695,905	198	3	3	3	3	…	3	2	2	2.8	3.9	…	0.1
1 000 à 2 499	605,212	612,630	380	648,645	659,684	410	…	618,389	381	629,495	624,270	378	2	2	2	2	…	2	2	2	1.2	1.7	…	-0.8
Moins de 1 000[2]	5,920,489	6,402,101	..	5,997,161	6,385,548	..	…	6,098,883	..	6,061,135	6,262,154	..	23	23	22	22	…	20	20	20	8.1	6.5	…	3.3
Toutes les régions	25,306,330	27,296,856	880	27,296,856	28,846,758	929	…	30,007,094	913	30,007,094	31,612,897	895	100	100	100	100	…	100	100	100	7.9	5.7	…	5.4
Population dans les centres de population ayant une taille de population de :																								
1 million et plus	7,158,005	7,865,789	3	7,962,741	8,539,938	3	…	9,412,027	3	9,372,715	10,022,987	3	28	29	29	30	…	31	31	32	9.9	7.2	…	6.9
500 000 et plus	10,781,972	11,771,087	9	11,897,094	12,619,908	9	…	13,781,948	9	13,751,629	14,683,200	9	43	43	44	44	…	46	46	46	9.2	6.1	…	6.8
100 000 et plus	13,487,593	14,659,948	24	15,032,919	15,894,691	26	…	17,510,281	29	17,526,067	18,656,653	29	53	54	55	55	…	58	58	59	8.7	5.7	…	6.5
50 000 et plus	14,832,953	16,139,716	45	16,594,861	17,557,281	50	…	18,992,112	51	19,075,407	20,309,762	52	59	59	61	61	…	63	64	64	8.8	5.8	…	6.5
30 000 et plus	15,789,740	17,154,953	71	17,552,463	18,550,128	76	…	19,958,431	77	20,189,396	21,506,812	83	62	63	64	64	…	67	67	68	8.6	5.7	…	6.5
10 000 et plus	17,235,544	18,670,512	164	19,033,236	20,105,914	173	…	21,556,113	177	21,685,794	23,069,834	183	68	68	70	70	…	72	72	73	8.3	5.6	…	6.4
5 000 et plus	18,055,573	19,536,982	291	19,930,606	21,052,849	309	…	22,502,993	310	22,621,543	24,030,568	319	71	72	73	73	…	75	75	76	8.2	5.6	…	6.2

1. Un centre de population possède une population d'au moins 1 000 habitants et une densité de population d'au moins 400 habitants par kilomètre carré. Pour 2001 et 2006, les données sont fondées sur les chiffres de population du recensement courant et pour les recensements antérieurs, elles étaient fondées sur le chiffre de population du recensement précédent.
2. En 1991, comprend 12 119 personnes réparties entre 13 établissements ayant une population de 834 à 999 habitants. En 1996, comprend 16 477 personnes réparties entre 17 établissements ayant une population de 858 à 999 habitants.
3. Les chiffres de population pour 1996 à l'intérieur des limites de 2001 ne sont pas disponibles. « Étant donné que les îlots de 2001 ne respectaient pas nécessairement les limites des secteurs de dénombrement de 1996, on n'a pas pu recréer parfaitement les régions urbaines de 1996 d'après les îlots de 2001. » Voir Matier, Kelly (2008). **Délimitation des régions urbaines de 2006 : Les défis et réalisations** (Ottawa : Statistique Canada, Série de documents de travail de la géographie n° 2008001, n° 92F0138 au catalogue), p. 5.
Source : Statistique Canada, Recensement de la population, 1986 à 2006.

En nous fondant sur les observations de Mendelson et Lefebvre (2003), qui ont constaté que les localités possédant un noyau de population d'au moins 50 000 habitants présentaient des fonctions « métropolitaines », nous pouvons également noter qu'en 2006, il existait 52 centres de population comptant au moins 50 000 habitants et qu'ils représentaient 64 % de la population canadienne.

D'autres vues du niveau d'urbanisation (pour les analystes qui privilégient les « centres de population » pour définir une « ville ») sont présentées au tableau 4.1-1. La part de la population vivant dans des centres de population d'au moins 1 million d'habitants est passée de 29 % en 1991 à 32 % en 2006. Les centres de population de 50 000 habitants ou plus représentaient 59 % de la population canadienne en 1991 et 64 % en 2006. Étant donné que certaines provinces (par exemple la Saskatchewan) attribuent le terme de « ville » aux localités de 5 000 habitants ou plus, nous montrons qu'en 2006, il existait 319 centres de population comptant au moins 5 000 résidents et qu'ils représentaient 76 % de la population canadienne, proportion en hausse par rapport à 72 % en 1991.

Les quatre dernières colonnes du tableau 4.1-1 donnent le taux de variation de la population (c'est-à-dire le taux d'urbanisation) attribuable à des facteurs démographiques (naissance, décès et migration nette) qui est calculé dans des limites constantes et en utilisant des catégories de taille constantes (fondées les unes et les autres sur la situation à la fin de la période de cinq ans). La plus grande partie de la différence de taux de croissance entre les catégories de taille de population est due à la migration nette. Donc, nous voyons dans quels endroits les gens préfèrent vivre ou déménager. De 2001 à 2006, la croissance de la population a été la plus rapide dans les centres de population de 30 000 à 49 999 habitants (une augmentation de 7,5 %). Dans les centres de population de 1 000 à 4 999 habitants, pratiquement aucun changement n'est observé. De 1991 à 1996, ce sont les centres de population de 1 million d'habitants et plus qui ont connu la croissance la plus rapide. De 1986 à 1991, les centres de 50 000 à 99 999 habitants sont ceux qui ont affiché la plus forte croissance. Donc, le taux d'urbanisation dû aux facteurs démographiques n'est pas forcément le plus élevé dans les centres de population les plus grands.

Cependant, la variation totale de la population vivant dans une catégorie de taille de centre de population donnée est également déterminée par la reclassification. Cette dernière peut découler a) d'un changement démographique faisant en sorte que la localité doit être reclassée dans une catégorie plus grande ou plus petite de centres de population, ou b) de la fusion éventuelle de deux centres de population faisant en sorte que la population du nouveau centre ainsi créé doit être reclassée dans une autre catégorie de taille. Donc, entre les périodes « t » et « t+1 », il se produit un changement du nombre de localités et un changement du nombre de Canadiens qui connaissent les avantages et les coûts de la vie dans une localité de taille donnée. En raison de ces reclassifications, nous observons une variation du nombre total de Canadiens résidant dans un centre de population appartenant à une catégorie de taille donnée (c'est-à-dire le taux d'urbanisation correspondant au taux de variation du nombre de personnes qui vivent dans une « ville » d'une taille donnée, qu'il soit dû au changement démographique ou à la reclassification). Un calcul fondé sur les données du tableau 4.1-1 (mais non présenté ici) indique que la catégorie de taille pour laquelle la croissance de la population a été la plus importante était celle de 30 000 à 49 999 habitants de 2001 à 2006 et celle de 100 000 à 499 999 habitants de 1991 à 1996. Fait important, la plupart (au moins les trois quarts) de la variation totale était due à des changements démographiques et moins du quart, à la reclassification.

4.2 Subdivisions de recensement représentant des régions administratives

En 2006, si l'on considère le niveau d'urbanisation, il existait deux SDR comptant au moins 1 million d'habitants (Montréal et Toronto) et elles représentaient 13 % de la population canadienne (tableau 4.2-1). En 1991, il n'existait qu'une seule SDR (Montréal) qui représentait 4 % de la population canadienne. En 2006, on dénombrait 48 SDR ayant une population de 100 000 habitants ou plus qui représentaient 52 % de la population canadienne (en hausse par rapport à 36 SDR représentant 38 % de la population canadienne).

En général, les grandes SDR (mais non les plus grandes) ont un taux d'urbanisation dû à la croissance démographique plus élevé que les SDR plus petites (tableau 4.2-1). De 2001 à 2006, la croissance démographique a été la plus importante dans les SDR de 250 000 à 499 999 habitants (hausse de 11,6 %) ; de 1996 à 2001, la croissance la plus forte a eu lieu dans la catégorie de 500 000 à 999 999 habitants (hausse de 8,3 %) ; de 1991 à 1996, elle a eu lieu dans la catégorie des 250 000 à 499 999 habitants et de 1986 à 1991, dans celle de 100 000 à 249 999 habitants.

Cependant, en raison des importantes fusions municipales qui ont eu lieu ces dernières décennies, de la moitié aux trois quarts de la variation totale de la population vivant dans une SDR d'une taille donnée est due à la reclassification. De 2001 à 2006, l'accroissement le plus important de l'urbanisation pour ce qui est de la variation de la population totale a eu lieu dans la catégorie de 100 000 à 249 999 habitants, de 1996 à 2001, la hausse la plus importante a été observée dans la catégorie de 1 million et plus (la SDR de Toronto ayant dépassé ce seuil durant cette période) et de 1991 à 1996, la croissance la plus importante a eu lieu dans la catégorie des 500 000 à 999 999 habitants.

Tableau 4.2-1
Population selon la taille de la subdivision de recensement, Canada, 1991 à 2006

Catégorie de taille de population de la subdivision de recensement (ville ou municipalité constituée en personne morale)	Population totale et nombre de subdivisions de recensement (villes ou municipalités constituées en personne morale)												Distribution en pourcentage de la population								Variation en pourcentage de la population dans des limites constantes et une classification constante			
	Dans les limites et la classification de 1991			Dans les limites et la classification de 1996			Dans les limites et la classification de 2001			Dans les limites et la classification de 2006			Dans les limites et la classification de 1991		Dans les limites et la classification de 1996		Dans les limites et la classification de 2001		Dans les limites et la classification de 2006					
	1986	1991	N[bre] de subdivision de recensement	1991	1996	N[bre] de subdivision de recensement	1996	2001	N[bre] de subdivision de recensement	2001	2006	N[bre] de subdivision de recensement	1986	1991	1991	1996	1996	2001	2001	2006	1986 à 1991	1991 à 1996	1996 à 2001	2001 à 2006
1 million et plus	1,015,420	1,017,666	1	1,017,669	1,016,376	1	3,401,797	3,521,028	2	4,065,084	4,123,974	2	4	4	4	4	12	12	14	13	0.2	-0.1	3.5	1.4
500 000 à 999 999	3,458,638	3,666,765	6	4,601,246	4,863,602	8	3,782,391	4,097,182	6	4,587,587	4,915,294	7	14	13	17	17	13	14	15	16	6.0	5.7	8.3	7.1
250 000 à 499 999	2,268,988	2,495,274	7	2,048,115	2,203,177	7	2,039,440	2,202,176	6	2,396,848	2,675,280	7	9	9	8	8	7	7	8	8	10.0	7.6	8.0	11.6
100 000 à 249 999	2,888,246	3,236,429	22	3,509,845	3,740,267	28	3,613,238	3,862,586	27	4,520,530	4,827,672	32	11	12	13	13	13	13	15	15	12.1	6.6	6.9	6.8
50 000 à 99 999	3,192,641	3,527,719	49	3,306,687	3,505,619	50	3,471,260	3,606,808	51	2,974,961	3,212,519	45	13	13	12	12	12	12	10	10	10.5	6.0	3.9	8.0
30 000 à 49 999	1,692,013	1,880,042	46	1,860,459	1,972,461	50	1,732,283	1,782,950	46	1,488,090	1,582,105	41	7	7	7	7	6	6	5	5	11.1	6.0	2.9	6.3
10 000 à 29 999	3,643,807	4,024,391	249	3,921,631	4,211,771	261	4,094,033	4,230,425	262	3,762,657	3,979,858	247	14	15	14	15	14	14	13	13	10.4	7.4	3.3	5.8
5 000 à 9 999	2,154,167	2,316,074	333	2,135,152	2,274,139	328	2,257,857	2,303,986	327	2,095,159	2,171,479	311	9	8	8	8	8	8	7	7	7.5	6.5	2.0	3.6
2 500 à 4 999	1,853,697	1,957,740	554	1,869,067	1,958,326	550	1,710,631	1,700,611	479	1,526,596	1,546,831	444	7	7	7	7	6	6	5	5	5.6	4.8	-0.6	1.3
1 000 à 2 499	1,823,180	1,874,269	1,200	1,783,300	1,844,056	1,174	1,549,505	1,535,309	980	1,441,071	1,453,236	922	7	7	7	6	5	5	5	5	2.8	3.4	-0.9	0.8
500 à 999	848,576	847,852	1,174	793,549	811,781	1,121	748,339	733,174	1,023	697,045	692,586	958	3	3	3	3	3	2	2	2	-0.1	2.3	-2.0	-0.6
250 à 499	333,976	331,271	874	326,863	329,392	880	323,461	312,991	843	318,270	308,209	832	1	1	1	1	1	1	1	1	-0.8	0.8	-3.2	-3.2
Moins de 250	135,982	121,367	1,491	123,276	115,794	1,526	122,526	117,868	1,548	133,196	123,854	1,570	1	0	0	0	0	0	0	0	-10.7	-6.1	-3.8	-7.0
Toutes les régions	25,309,331	27,296,859	6,006	27,296,859	28,846,761	5,984	28,846,761	30,007,094	5,600	30,007,094	31,612,897	5,418	100	100	100	100	100	100	100	100	7.9	5.7	4.0	5.4
Population dans les subdivisions de recensement ayant une taille de population de :																								
1 million et plus	1,015,420	1,017,666	1	1,017,669	1,016,376	1	3,401,797	3,521,028	2	4,065,084	4,123,974	2	4	4	4	4	12	12	14	13	0.2	-0.1	3.5	1.4
500 000 et plus	4,474,058	4,684,431	7	5,618,915	5,879,978	9	7,184,188	7,618,210	8	8,652,671	9,039,268	9	18	17	21	20	25	25	29	29	4.7	4.6	6.0	4.5
250 000 et plus	6,743,046	7,179,705	14	7,667,030	8,083,155	16	9,223,628	9,820,386	14	11,049,519	11,714,548	16	27	26	28	28	32	33	37	37	6.5	5.4	6.5	6.0
100 000 et plus	9,631,292	10,416,134	36	11,176,875	11,823,422	44	12,836,866	13,682,972	41	15,570,049	16,542,220	48	38	38	41	41	45	46	52	52	8.1	5.8	6.6	6.2
50 000 et plus	12,823,933	13,943,853	85	14,483,562	15,329,041	94	16,308,126	17,289,780	92	18,545,010	19,754,739	93	51	51	53	53	57	58	62	62	8.7	5.8	6.0	6.5
30 000 et plus	14,515,946	15,823,895	131	16,344,021	17,301,502	144	18,040,409	19,072,730	138	20,033,100	21,336,844	134	57	58	60	60	63	64	67	67	9.0	5.9	5.7	6.5
10 000 et plus	18,159,753	19,848,286	380	20,265,652	21,513,273	405	22,134,442	23,303,155	400	23,795,757	25,316,702	381	72	73	74	75	77	78	79	80	9.3	6.2	5.3	6.4
5 000 et plus	20,313,920	22,164,360	713	22,400,804	23,787,412	733	24,392,299	25,607,141	727	25,890,916	27,488,181	692	80	81	82	82	85	85	86	87	9.1	6.2	5.0	6.2

Le terme général « subdivision de recensement (SDR) » désigne les municipalités (villes ou municipalités rurales constituées en personne morale) définies par les lois provinciales/territoriales ou les régions considérées comme l'équivalent d'une municipalité à des fins statistiques (p. ex. les réserves indiennes, les établissements indiens et les territoires non organisés).
Source : Statistique Canada, Recensement de la population, 1986 à 2006.

4.3 Régions métropolitaines de recensement et agglomérations de recensement représentant des zones fonctionnelles

En 2006, pour ce qui est du niveau d'urbanisation, 45 % de Canadiens vivaient dans une RMR d'au moins 1 million d'habitants, c'est à dire une hausse par rapport aux 29 % observés en 1991 (tableau 4.3-1).

Durant chaque période de cinq ans entre 1981 et 2006, les RMR de 1 million d'habitants ou plus ont affiché le taux d'urbanisation le plus élevé en ce qui concerne la croissance démographique. Cependant, durant certaines périodes, d'autres catégories des tailles de RMR ou d'AR ont connu le changement d'urbanisation le plus important en ce qui a trait à la variation de la population totale : de 1996 à 2001, il s'agissait des RMR de 100 000 à 249 999 habitants, de 1986 à 1991, des AR de 50 000 à 99 999 habitants, et de 1981 à 1986, des AR de 10 000 à 49 999 habitants. Cependant, la plupart (de la moitié aux trois quarts) du taux d'urbanisation était imputable à la croissance démographique.

Tableau 4.3-1
Population selon la catégorie de taille de la zone fonctionnelle de marché du travail, Canada, 1981 à 2006

Catégorie de taille de la zone fonctionnelle de marché du travail	Population										Répartition en pourcentage de la population												Variation en pourcentage					
	Dans les limites de 1981		Dans les limites de 1986		Dans les limites de 1991		Dans les limites de 1996		Dans les limites de 2001		Dans les limites de 2006		Dans les limites de 1981		Dans les limites de 1986		Dans les limites de 1991		Dans les limites de 1996		Dans les limites de 2001		Dans les limites de 2006		De 1976 à 1981	De 1981 à 1986	De 1986 à 1991	De 1991 à 1996	De 1996 à 2001	De 2001 à 2006
	1976	1981	1981	1986	1986	1991	1991	1996	1996	2001	2001	2006	1976	1981	1981	1986	1986	1991	1991	1996	1996	2001	2001	2006						
Grandes RMR (1 million et plus)	6,771,966	7,095,479	7,260,861	7,729,254	7,734,067	8,622,790	9,652,307	10,432,430	10,420,589	11,159,875	13,078,028	14,110,317	28	29	30	31	31	32	35	36	36	37	44	45	4.8	6.5	11.5	8.1	7.1	7.9
RMR moyennes (de 500 000 à 999 999)	3,370,701	3,670,790	3,822,645	4,061,654	4,050,342	4,412,478	3,500,925	3,647,688	3,647,567	3,905,672	2,025,564	2,103,094	14	15	16	16	16	16	13	13	13	13	7	7	8.9	6.3	8.5	4.2	7.1	3.8
Petites RMR (de 100 000 à 499 999)	2,767,796	2,892,675	3,224,726	3,364,585	3,364,195	3,630,092	3,633,886	3,784,538	4,110,441	4,231,373	5,017,869	5,295,164	11	12	13	13	13	13	13	13	14	14	17	17	4.5	4.3	7.9	4.1	2.9	5.5
Régions métropolitaines de recensement (total partiel)	**12,910,463**	**13,658,944**	**14,308,232**	**15,155,493**	**15,148,604**	**16,665,360**	**16,787,118**	**17,864,646**	**18,178,597**	**19,296,926**	**20,121,461**	**21,508,575**	**53**	**56**	**59**	**60**	**60**	**61**	**61**	**62**	**63**	**64**	**67**	**68**	**5.8**	**5.9**	**10.0**	**6.4**	**6.2**	**6.9**
Grandes AR (de 50 000 à 99 999)	1,447,751	1,523,607	1,903,808	1,973,431	2,069,418	2,277,832	2,407,087	2,578,276	2,349,659	2,423,726	1,847,219	1,947,917	6	6	8	8	8	8	9	9	8	8	6	6	5.2	3.7	10.1	7.1	3.2	5.5
AR moyennes (de 30 000 à 49 999)	1,069,788	1,128,815	1,035,003	1,057,094	1,137,284	1,194,341	1,018,807	1,056,633	1,057,158	1,064,817	1,118,738	1,157,978	4	5	4	4	4	4	4	4	4	4	4	4	5.5	2.1	5.0	3.7	0.7	3.5
Petites AR (de 10 000 à 29 999)	701,844	711,176	1,038,437	1,029,089	903,473	929,681	927,144	950,300	1,069,278	1,053,617	997,280	1,017,087	3	3	4	4	4	3	3	3	4	4	3	3	1.3	-0.9	2.9	2.5	-1.5	2.0
Agglomérations de recensement (total partiel)	**3,219,383**	**3,363,598**	**3,977,248**	**4,059,614**	**4,110,175**	**4,401,854**	**4,353,038**	**4,585,209**	**4,476,095**	**4,542,160**	**3,963,237**	**4,122,982**	**13**	**14**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**15**	**13**	**13**	**4.5**	**2.1**	**7.1**	**5.3**	**1.5**	**4.0**
Grands centres urbains (total partiel)	**16,129,846**	**17,022,542**	**18,285,480**	**19,215,107**	**19,258,779**	**21,067,214**	**21,140,156**	**22,449,855**	**22,654,692**	**23,839,086**	**24,084,698**	**25,631,557**	**66**	**70**	**75**	**76**	**76**	**77**	**77**	**78**	**79**	**79**	**80**	**81**	5.5	5.1	9.4	6.2	5.2	**6.4**
Zone d'influence métropolitaine forte	..	..	..	..	1,435,028	1,574,359	1,458,448	1,564,700	1,470,493	1,524,579	1,289,265	1,350,098	..	..	..	..	6	6	5	5	5	5	4	4	..	..	9.7	7.3	3.7	4.7
Zone d'influence métropolitaine modérée	..	..	..	..	2,280,052	2,335,157	2,289,911	2,365,175	2,307,387	2,285,538	2,203,563	2,224,347	..	..	..	..	9	9	8	8	8	8	7	7	..	..	2.4	3.3	-0.9	0.9
Zone d'influence métropolitaine faible	..	..	..	..	1,952,122	1,951,974	2,041,871	2,078,342	2,027,488	1,969,211	2,077,950	2,049,190	..	..	..	..	8	7	7	7	7	7	6	..	..	..	0.0	1.8	-2.9	-1.4
Zone sans influence métropolitaine	..	..	..	..	334,560	315,813	316,281	332,604	330,616	333,847	296,785	297,984	..	..	..	..	1	1	1	1	1	1	1	1	..	..	-5.6	5.2	1.0	0.4
RRPV dans les territoires	..	..	..	..	48,790	52,342	50,192	56,085	56,085	54,833	54,833	59,712	..	..	..	..	0	0	0	0	0	0	0	0	..	..	7.3	11.7	-2.2	8.9
Régions rurales et petites villes (RRPV) (total partiel)	**6,862,759**	**7,320,635**	**6,057,697**	**6,094,222**	**6,050,552**	**6,229,645**	**6,156,703**	**6,396,906**	**6,192,069**	**6,168,008**	**5,922,396**	**5,981,340**	**28**	**30**	**25**	**24**	**24**	**23**	**23**	**22**	**21**	**21**	**20**	**19**	**6.7**	**0.6**	**3.0**	**3.9**	**-0.4**	**1.0**
Total	**22,992,605**	**24,343,177**	**24,343,177**	**25,309,329**	**25,309,331**	**27,296,859**	**27,296,859**	**28,846,761**	**28,846,761**	**30,007,094**	**30,007,094**	**31,612,897**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	5.9	4.0	7.9	5.7	4.0	**5.4**
Population dans les zones fonctionnelles de marché du travail ayant une population de :																														
1 million et plus	6,771,966	7,095,479	7,260,861	7,729,254	7,734,067	8,622,790	9,652,307	10,432,430	10,420,589	11,159,875	13,078,028	14,110,317	29	29	30	31	31	32	35	36	36	37	44	45	4.8	6.5	11.5	8.1	7.1	7.9
500 000 et plus	10,142,667	10,766,269	11,083,506	11,790,908	11,784,409	13,035,268	13,153,232	14,080,118	14,068,156	15,065,548	15,103,592	16,213,411	44	44	46	47	47	48	48	49	49	50	50	51	6.1	6.4	10.6	7.0	7.1	7.3
100 000 et plus	12,910,463	13,658,944	14,308,232	15,155,493	15,148,604	16,665,360	16,787,118	17,864,646	18,178,597	19,296,926	20,121,461	21,508,575	56	56	59	60	60	61	62	63	64	67	68	5.8	5.9	10.0	6.4	6.2	6.9
50 000 et plus	14,358,214	15,182,551	16,212,040	17,128,924	17,218,022	18,943,192	19,194,205	20,442,922	20,528,256	21,720,652	21,968,680	23,456,492	62	62	67	68	68	69	70	71	71	72	73	74	5.7	5.7	10.0	6.5	5.8	6.8
30 000 et plus	15,428,002	16,311,366	17,247,043	18,186,018	18,355,306	20,137,535	20,213,012	21,499,555	21,585,414	22,785,469	23,087,418	24,614,477	67	67	71	72	73	74	74	75	75	76	77	78	5.7	5.4	9.7	6.4	5.6	6.6
10 000 et plus	16,129,846	17,022,542	18,285,480	19,215,107	19,258,779	21,067,214	21,140,156	22,449,855	22,654,692	23,839,086	24,084,698	25,631,557	70	70	75	76	76	77	77	78	79	79	80	81	5.5	5.1	9.4	6.2	5.2	6.4

Les régions métropolitaines de recensement (RMR) possèdent un noyau urbain de 100 000 habitants ou plus et englobent toutes les villes et municipalités voisines dont au moins 50 % de la population active fait la navette pour aller travailler dans le noyau urbain.
Les agglomérations de recensement (AR) ont un noyau urbain de 10 000 à 99 999 habitants et englobent toutes les villes et municipalités voisines dont au moins 50 % de la population active fait la navette pour aller travailler dans le noyau urbain.
Les zones d'influence métropolitaines (ZIM) sont classées en fonction de la part de leur population active qui fait la navette pour aller travailler dans une RMR ou une AR (ZIM forte : de 30 à 49 %; ZIM modérée : de 5 à 29 %; ZIM faible : de 1 à 5 %; zone sans IM : pas de navetteurs.
Les données pour les ZIM de 1991 et de 1996 ont été rajustées afin qu'elles concordent avec le protocole de 2001 en vertu duquel les villes et les municipalités autres qu'une RMR ou une AR dans les territoires n'ont pas été réparties en fonction de la classification des ZIM.
La désignation des ZIM pour 1991 et 1996 a été tirée de Sheila Rambeau et Kathleen Todd (2000), **Zones d'influence des régions métropolitaines de recensement et des agglomérations de recensement (ZIM) accompagnées de données du recensement** (Ottawa : Statistique Canada, série de documents de travail de la géographie n° 2000-1, n° 92F0138MIF au catalogue) (www.statcan.ca/cgi-bin/downpub/listpub.cgi?catno=92F0138MIF). Il convient de souligner que, pour désigner les ZIM de 1991, Rambeau et Todd se sont servies des délimitations provisoires des RMR/AR de 1996, tout en conservant les limites de 1991. Pour le présent tableau, nous avons réimposé la délimitation des RMR/AR de 1981 et avons attribué la catégorie « ZIM forte » en 1991 aux villes et aux municipalités codées comme une RMR/AR pour 1996.
Pour 2001, la désignation des ZIM a été tirée de Statistique Canada, **GéoSuite, Recensement de 2001** (Ottawa : Statistique Canada, n° 92F0150GIF au catalogue).
Source : Statistique Canada, Recensement de la population, 1981 à 2006.

5. Discussion : Donc, combien de Canadiens vivent dans une ville?

La réponse à cette question est manifestement « Cela dépend! ».

Si l'on définit une « ville » comme étant une localité d'au moins 100 000 habitants, la réponse est (en 2006) :

- 59 % dans des zones bâties (centres de population) de 100 000 habitants ou plus ;
- 52 % dans des unités administratives (subdivisions de recensement) de 100 000 habitants ou plus ;
- 68 % dans des zones fonctionnelles (RMR) de 100 000 habitants ou plus (tableau 5.1-1).

La réponse dépend également des personnes à qui l'on s'adresse :

- 59 % si l'on s'adresse à des planificateurs des transports en commun ;
- 52 % si l'on s'adresse à des maires ;
- 68 % si l'on s'adresse à des analystes du développement économique.

Pour une catégorie de taille donnée, nous classons nos définitions des villes en fonction de celle qui donne le niveau d'urbanisation le plus élevé. Le classement est le suivant : a) zones fonctionnelles (RMR et AR), b) zones bâties (centres de population) et c) unités administratives (subdivisions de recensement).

De même, pour une catégorie de taille donnée, nous pouvons classer nos définitions des villes en fonction de celle qui produit le taux d'urbanisation le plus élevé. Le classement est le suivant : a) unités administratives (subdivisions de recensement) (en raison des fusions), et b) zones bâties (centres de population) et zones fonctionnelles (RMR et AR).

Tableau 5.1-1
Combien de Canadiens vivent dans une ville?

. . . Selon la catégorie de taille de population déterminant une « ville »							
. . . . Selon la façon de définir une « ville »							
Selon la catégorie de taille de population déterminant une « ville »	Selon la façon de définir une « ville »	1981	1986	1991	1996	2001	2006
		Pourcentage de la population totale					
1 million et plus	Centres de population[1]	..	..	29	30	31	32
	Subdivisions de recensement[2]	..	..	4	4	12	13
	RMR et AR[3]	29	31	32	36	37	45
500 000 et plus	Centres de population[1]	..	..	43	44	46	46
	Subdivisions de recensement[2]	..	..	17	20	25	29
	RMR et AR[3]	44	47	48	49	50	51
100 000 et plus	Centres de population[1]	..	..	54	55	58	59
	Subdivisions de recensement[2]	..	..	38	41	46	52
	RMR et AR[3]	56	60	61	62	64	68
50 000 et plus	Centres de population[1]	..	..	59	61	63	64
	Subdivisions de recensement[2]	..	..	51	53	58	62
	RMR et AR[3]	62	68	69	71	72	74
30 000 et plus	Centres de population[1]	..	..	63	64	67	68
	Subdivisions de recensement[2]	..	..	58	60	64	67
	RMR et AR[3]	67	72	74	75	76	78
10 000 et plus	Centres de population[1]	..	..	68	70	72	73
	Subdivisions de recensement[2]	..	..	73	75	78	80
	RMR et AR[3]	70	76	77	78	79	81

1. Centre de population : Forme ou morphologie ou zone bâtie (toute localité ayant une densité de population de 400 habitants par km^2 ou plus – délimité pour des localités dont la population totale est d'au moins 1 000 habitants.
2. Subdivision de recensement : Unité administrative (une ville ou une municipalité constituée en personne morale).
3. RMR ou AR : Unité fonctionnelle de marché du travail (régions métropolitaines de recensement et agglomérations de recensement – possédant un noyau urbain de 10 000 habitants ou plus et englobant toutes les subdivisions de recensement voisines dans lesquelles 50 % ou plus des résidents ayant un emploi font la navette pour aller travailler dans le noyau urbain).

6. Conclusion

En utilisant trois moyens différents de montrer la densité (ou la taille) de population, nous constatons que :

1. la réponse à la question « Combien de Canadiens vivent dans une ville? » dépend de la personne à laquelle on s'adresse, ainsi que du seuil de taille utilisé pour considérer une localité comme une « ville » dans le contexte du sujet discuté ;
2. la croissance démographique dicte le taux d'urbanisation pour les zones bâties (centres de population) et pour les zones fonctionnelles (RMR et AR) ;
3. la reclassification (y compris les fusions) sous tend la majeure partie de l'urbanisation pour les régions administratives (subdivisions de recensement).

Bibliographie

Hofmann, N. et coll. (2010), « *Présentation d'un nouveau concept et d'une nouvelle méthodologie de délimitation des zones habitées : un projet de recherche sur les zones habitées au Canada* », série de documents analytiques et techniques sur les comptes et la statistique de l'environnement, document technique no 11, no 16-001-M au catalogue de Statistique Canada, http://www.statcan.gc.ca/pub/16-001-m/16-001-m2010011-fra.pdf.

Hofmann, N. et coll. (2010), « Un nouveau projet de recherche sur les zones habitées au Canada : premiers résultats géographiques », *EnviroStats*, vol. 4, no 1, no 16-002-X au catalogue de Statistique Canada, http://www.statcan.gc.ca/pub/16-002-x/16-002-x2010001-fra.pdf.

Mendelson, R. et J. Lefebvre (2003), « *Examen des régions métropolitaines de recensement (RMR) et des agglomérations de recensement (AR) au Canada selon la fonctionnalité métropolitaine* », série de documents de travail de la géographie, no 92F0138MIF au catalogue - no 001 au catalogue de Statistique Canada, http://www.statcan.gc.ca/pub/92f0138m/92f0138m2003001-fra.pdf.

Puderer, H.A. (2009), « *Perspectives et mesures de l'urbain* », série de documents de travail de la géographie, no 92F0138M - no 2009001 au catalogue de Statistique Canada, http://www.statcan.gc.ca/pub/92f0138m/92f0138m2009001-fra.htm.

Statistique Canada (2007), *Dictionnaire du Recensement de 2006*, no 92-566-XWF au catalogue de Statistique Canada, http://www12.statcan.gc.ca/census-recensement/2006/ref/dict/index-fra.cfm.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Rôle des normes de qualité des données dans l'uniformisation des méthodes et des outils d'enquête

John L. Eltinge[1]

Résumé

Cette communication explore l'interface entre les normes de qualité des données et l'uniformisation des méthodes et des outils d'enquête. En premier lieu, elle aborde la méthodologie d'enquête statistique comme une forme de technologie et utilise le cadre conceptuel qui en découle pour explorer plusieurs façons d'évaluer les coûts et avantages éventuels des normes de qualité et de l'uniformisation des méthodes pour les programmes statistiques. Le cadre conceptuel met principalement l'accent sur les types de normes, les méthodes de calage, les méthodes de mise en œuvre et d'application, et les questions spéciales relatives aux programmes statistiques parrainés par le gouvernement. Ce cadre mène à un examen des avantages éventuels des normes, y compris l'amélioration de la qualité des données ainsi que la réduction des coûts quantifiables et la réduction des risques pour les intervenants. En parallèle avec les avantages, la communication passe aussi en revue les coûts et les risques possibles liés aux normes d'enquête, y compris les coûts directs, les coûts indirects et l'affectation inefficace des ressources.

En deuxième lieu, cette communication évalue les répercussions de l'uniformisation des méthodes et des outils sur le processus d'enquête. Parmi les répercussions possibles figurent les modifications, tant positives que négatives, concernant les éléments suivants : a) les composantes fixes et variables de la structure de coûts globale de l'enquête ; b) la robustesse de l'enquête dans le contexte de changements dans les besoins des intervenants, la disponibilité des ressources ou l'environnement opérationnel externe ; c) la mesure dans laquelle l'enquête en découlant pourrait répondre à certaines normes de qualité des données.

La communication se termine par des commentaires sur les répercussions pratiques de ces concepts généraux pour : l'élaboration, la mise en œuvre et l'application de normes ; l'élaboration, la mise en œuvre et la mise à jour de méthodes et d'outils uniformisés ; la communication avec les intervenants de l'extérieur.

[1]John L. Eltinge, Bureau of Labor Statistics, États-Unis.

Travaux liés à l'architecture intégrée à Statistics Sweden

Martin Axelson, Jakob Engdahl, Ylva Fossan, Eva Holm, Ingegerd Jansson, Boris Lorenc et Lars Göran Lundel[1]

Résumé

Le document présente les travaux en cours liés à l'architecture intégrée et à ses composantes à Statistics Sweden. Il fait état des fondements de l'architecture, à savoir les éléments moteurs, les normes et les cadres de l'architecture, ainsi que d'une vision pour la production de statistiques. La modélisation est considérée comme la principale contribution des architectes intégrés, en général et grâce à deux principaux domaines d'application, à savoir, la plateforme Triton pour la conception, la collecte des données et le post-traitement, ainsi qu'une vision élaborée d'une stratégie d'entreposage des données et d'une plateforme pour la production de statistiques intégrées à partir de données administratives et de données obtenues au moyen de la collecte de données primaires. La modélisation à l'extérieur du processus opérationnel est aussi mentionnée. Le document se termine par certaines remarques générales.

Mots clés : Architecture intégrée ; architecture opérationnelle ; The Open Group Architecture Framework (TOGAF) ; modélisation de l'information ; modélisation des processus.

« - Voudriez-vous me dire, s'il vous plaît, quel chemin je dois prendre pour m'en aller d'ici?
- Cela dépend beaucoup de l'endroit où tu veux aller, répondit le chat.
- Peu m'importe l'endroit... dit Alice.
- En ce cas, peu importe la route que tu prendras, répliqua-t-il.
- ... pourvu que j'arrive quelque part, ajouta Alice en guise d'explication
- Oh, tu ne manqueras pas d'arriver quelque part, si tu marches assez longtemps. »
Lewis Carroll : *Alice au pays des merveilles*

1. Bases

1.1 Éléments moteurs

À Statistics Sweden, les travaux liés à l'architecture intégrée et à ses composantes sont motivés par les objectifs suivants :

- augmenter l'efficacité du processus de production et réduire les coûts ;
- augmenter la qualité des processus et des statistiques produites ;
- réduire le fardeau administratif des fournisseurs de données découlant de la collecte de données primaires ;
- réagir plus rapidement aux changements dans les besoins des utilisateurs et dans l'environnement (y compris l'environnement des technologies de l'information -TI) ;
- augmenter la transparence et simplifier la gouvernance du processus de gestion des changements ;
- améliorer la planification à long terme, ainsi que l'aperçu et la gestion des travaux de développement.

Au niveau interne, nous percevons ces éléments moteurs, qui sont partagés par de nombreux autres instituts nationaux de statistique (INS), comme des composantes d'une équation plus simple qui comprend uniquement la qualité et le coût. L'objectif d'un producteur de statistiques est de produire des statistiques de la plus grande qualité possible, selon un certain coût fixe, ou encore, de réduire le coût de la production de statistiques, selon une certaine qualité déterminée. La normalisation (par exemple, Hofman et coll., 2011 ; Lopdell et Dunnet, 2011) représente une façon d'atteindre cet objectif, mais ne constitue par un objectif en soi ; il s'agit d'un objectif uniquement si cela mène à une amélioration globale (selon une certaine évaluation) de la fonction conjointe de coût/qualité.

[1]Martin Axelson, Jakob Engdahl, Ylva Fossan, Eva Holm, Ingegerd Jansson, Boris Lorenc et Lars Göran Lundel, Statistics Sweden, 701 89 Örebro, Suède.

1.2 Normes

Dans ses travaux liés à l'architecture intégrée et à ses composantes, Statistics Sweden tient compte des normes et des approches courantes concernant la réglementation de certaines des questions principales liées à la production de statistiques, et participe à l'élaboration de certaines d'entre elles. Il s'agit notamment du Generic Statistical Business Process Model (METIS 2009), du Generic statistical information model (OCMIMF 2011), du Statistical Data and Metadata exchange (SDMX 2011), de la Data Documentation Initiative (Vardigan et coll., 2008), et CORE (Scannapieco, 2011).

L'importance de l'existence de ces normes et de la dépendance à l'égard de celles ci dépend de l'amélioration de la communication entre les différents INS, de la réduction des travaux qui font double emploi et de la recherche d'une possibilité de réduction des coûts, grâce à une division du travail. Par ailleurs, la comparabilité entre les pays (c'est-à-dire ceux de l'Union européenne - UE) ou entre les différents producteurs de statistiques dans un environnement statistique national décentralisé (par exemple, aux États Unis) nécessite une grande similitude entre les processus et les définitions, à tout le moins au niveau de la production, afin d'assurer la comparabilité des statistiques produites.

1.3 Cadre d'architecture intégrée

Afin de structurer les travaux de Statistics Sweden concernant l'architecture intégrée, nous utilisons la notion de niveaux d'architecture du cadre de Zachman (1987), qui est constitué des éléments suivants : I. Architecture intégrée (AI), II. Architecture opérationnelle (AO), III. Architecture de solutions, IV. Architecture d'applications, V. Architecture d'infrastructure des TI, y compris le matériel. Même si la majeure partie de nos travaux pratiques se déroule au niveau de l'AO ou à un niveau inférieur, cela ne semble pas suffire pour atteindre les objectifs de la section 1.1. Une perspective plus large est requise, qui englobe les coûts (considérations budgétaires), la compétence du personnel, un cadre de qualité, *etc*. Par conséquent, Statistics Sweden est d'avis qu'une approche d'architecture intégrée est nécessaire pour pouvoir, dans une perspective à long terme, obtenir un système de production de statistiques raisonnable. Par architecture intégrée, nous entendons « un ensemble cohérent de principes, de méthodes et de modèles, qui servent à la conception et à la réalisation d'une structure organisationnelle intégrée, de processus opérationnels, de systèmes d'information et d'une infrastructure (Lankhorst et coll., 2009) ».

De façon plus particulière, nous explorons TOGAF (The Open Group, 2009) comme cadre susceptible d'appuyer l'approche holistique dont il est question dans le paragraphe précédent. Un des aspects attrayants de TOGAF est qu'il intègre une méthode pour le développement de composantes d'architecture, appelée méthode de développement d'architecture.

1.4 Vision

Nous reconnaissons qu'il n'est ni simple, ni judicieux, de préciser de façon trop détaillée les propriétés exactes qu'un système de production statistique aura, disons, dans dix ans. Toutefois, le fait d'avoir une vision comporte des avantages clairs. Prenons l'exemple de l'objectif de meilleure intégration des données administratives et des données d'enquête dans la production statistique. Il est peu probable que l'on atteigne cet objectif dans un avenir prévisible simplement en approuvant les propositions de développement au fur et à mesure qu'elles se présentent. Cet objectif doit figurer au premier plan, et les étapes vers sa réalisation doivent être entreprises activement pour qu'il ait une chance raisonnable d'être atteint.

De façon plus particulière, la communication de la Commission européenne sur la « méthode de production de statistiques dans l'UE » (Communauté européenne, 2009), appelée Vision 2020, est digne de mention. Certaines composantes de la vision s'appliquent uniquement au niveau supranational, mais celle ci nécessite en général la transition d'un modèle de production statistique cloisonné à des systèmes de production statistique, à une utilisation accrue des données administratives dans la production, *etc*., et est par conséquent source d'inspiration pour tout producteur de statistiques officielles.

Même si Statistics Sweden travaille toujours à sa vision à long terme, certaines composantes de celle ci sont présentes dans les deux plateformes dont il est question à la section 4. D'autres composantes ont aussi trait à l'atteinte des objectifs de la section 1.1. L'une d'elles, par exemple, comprend la fourniture de services conçus pour

n'imposer que des exigences légères aux ressources humaines des TI. Cet objectif et d'autres font ressortir la nécessité de mettre en place des mesures quantitatives des progrès, en vue de la réalisation des composantes de la vision, y compris le suivi des coûts (ou plus précisément, du rendement des investissements).

2. Mise en place organisationnelle

Un *groupe de l'architecture* a été créé en 2008 dans le département de recherche et développement (R-D) de Statistics Sweden. Le groupe est constitué d'environ dix méthodologistes principaux et experts principaux des TI, qui répartissent leur temps entre des travaux d'architecture et d'autres fonctions. L'environnement immédiat du groupe de l'architecture, à l'intérieur du département de R-D, est constitué du *groupe de la qualité* et du *groupe du chef de projet*.

En vue de normaliser les processus utilisés pour la production de statistiques, un département des processus a aussi été créé en 2008 et regroupe la majeure partie des compétences en TI et en méthodologie de Statistics Sweden. Par ailleurs, un nouveau groupe a été établi au sein du département des processus : les propriétaires de processus pour les principales étapes du MSGPO. En 2011, une évaluation des changements apportés en 2008 a entraîné une réorganisation du département des processus. Un nouveau département des TI a été créé à partir des unités de TI qui le constituaient, en vue d'améliorer la gouvernance du secteur des TI. De ce fait, deux forums d'architecture sont sur le point d'être mis en place pour coordonner les activités d'architecture entre les départements des TI, des processus et de R-D.

Un groupe de gestion de projet (GGP), présidé par le directeur général adjoint responsable de la R-D, est en place depuis 2008 et veille à ce que les efforts de développement fassent l'objet d'un ordre de priorité, d'un point de vue global, et à ce que les initiatives locales ne s'écartent pas des objectifs généraux. Une évaluation récente du GGP a mené à une proposition de ressource de *gestion du portefeuille des projets de développement*.

3. Mise en œuvre de l'AI à Statistics Sweden

Les architectes opérationnels de Statistics Sweden utilisent quatre composantes de base pour structurer les travaux relatifs à l'AO et à l'AI : objectifs opérationnels, processus, information et applications. Les objectifs opérationnels se situent dans un espace où sont réalisés les processus opérationnels et l'information nécessaires pour les mener à bien. Un rapport de compatibilité mutuelle doit exister entre les processus et l'information. Une spécification exacte des objectifs, des processus et de l'information sert de contexte aux solutions d'application. Cela a pour conséquence que les règles opérationnelles, les organigrammes de processus et les structures d'information sont représentés dans les modèles de façon systématique et cohérente.

3.1 Modélisation des processus

La description formelle du modèle de processus opérationnel à Statistics Sweden coïncide dans une large mesure avec celle comprise dans le MSGPO. Toutefois, une AI complète fournira aussi des modèles correspondants pour les processus à l'appui des activités, par exemple, la gestion des ressources humaines et la gestion du cadre juridique.

La modélisation des processus se fait à un niveau qui permet de déterminer les sous processus opérationnels communs. L'objectif de la modélisation est de produire une représentation des organigrammes de processus. À l'heure actuelle, nous disposons de modèles pour les sous processus suivants :

- à l'étape de la conception : i) choisir la ou les sources de données et la ou les méthodes de collecte des données, ii) choisir une stratégie de contact et déterminer les groupes de population pertinents, iii) décider du niveau de contrôle pendant la collecte des données et choisir la méthode d'entrée des données, iv) concevoir le flux de production, v) vérifier les programmes administratifs, vi) planifier et réserver les ressources ;

- à l'étape de la collecte : i) mettre à jour l'échantillon, ii) préparer la distribution des questionnaires, iii) préparer le scannage, iv) préparer la collecte des données sur le Web, v) fournir du soutien au fournisseur de données, vi) gérer les déclarations en double, vii) gérer les rappels, viii) scanner et vérifier ;
- à l'étape du traitement (sous processus de contrôle) : i) alerte d'erreur, ii) contrôle automatique, iii) vérification manuelle.

3.2 Modélisation de l'information

La modélisation de l'information aide à officialiser la description de l'information utilisée au niveau intégré, avec comme objectif d'assurer la compatibilité avec l'infrastructure des TI qui sert au traitement de l'information. Les modèles sont conceptuels, formels et indépendants des couches inférieures d'application et d'infrastructure, et peuvent être interprétés sans équivoque par celles ci. Nous faisons une distinction entre deux niveaux de modèles : modèles de groupe d'objet et modèles d'objet (détaillés).

Les modèles de groupe d'objet peuvent servir à déterminer la propriété de l'information. Les modèles d'objet servent de base à un répertoire de concepts communs au niveau intégré. Les modèles font aussi partie de l'espace des exigences générales et aident à le préciser. Par exemple, aucun objectif opérationnel ne peut être mis en œuvre dans les modèles de données matérielles sans être d'abord intégré dans des modèles d'objet. Ainsi, l'interprétation des objectifs opérationnels est officialisée et consignée, plutôt que de se faire de façon arbitraire (et non documentée) par le personnel responsable de la mise en œuvre des applications.

Parmi les groupes de modèles d'information élaborés jusqu'à maintenant à Statistics Sweden figure, par exemple, l'*enquête statistique* (avec des modèles d'objet constitués des éléments suivants : *cycle d'enquête*, *échantillon*, *article de rapport*, *envoi de documents*, *élément de données*, *commentaire*, *etc.*) ; à l'extérieur du MSGPO, nous avons des modèles de groupe d'objet pour, par exemple, le *personnel* et la *structure organisationnelle*. Les principes architecturaux en comprennent un qui permet d'énoncer les types particuliers d'information qui ne devraient pas être créés et/ou entreposés à plus d'un endroit, par exemple, l'*enquête statistique*, la *règle de contrôle* et le *code d'état*. Cela doit toutefois être pondéré par rapport à d'autres préoccupations liées à l'architecture, comme le risque de créer des dépendances trop grandes entre les systèmes (et, par conséquent, un système instable).

3.3 Modélisation de concepts

Pour qu'un système puisse être partagé par plusieurs, les concepts utilisés doivent être bien définis. Les concepts sont présents de façon implicite dans les modèles de processus et les modèles d'information ; toutefois, il se peut que des concepts cruciaux doivent être officialisés dans des modèles de concept bien définis. Nous avons élaboré de tels modèles pour les sous processus de la vérification et du contrôle de la divulgation du MSGPO.

Un résultat très utile de la modélisation de concept et de la modélisation de l'information effectuées à Statistics Sweden a été, jusqu'à maintenant, de rendre explicites les différentes façons dont les concepts et l'information étaient interprétés dans l'organisation ; cela a par la suite mené, grâce à des discussions, à l'adoption de définitions unifiées des termes utilisés.

Nous avons l'intention de déterminer si la formulation des modèles doit être dépendante du niveau de l'architecture (dont il est question à la section 1.3) auquel ils doivent s'appliquer.

3.4 Capacité de l'architecture

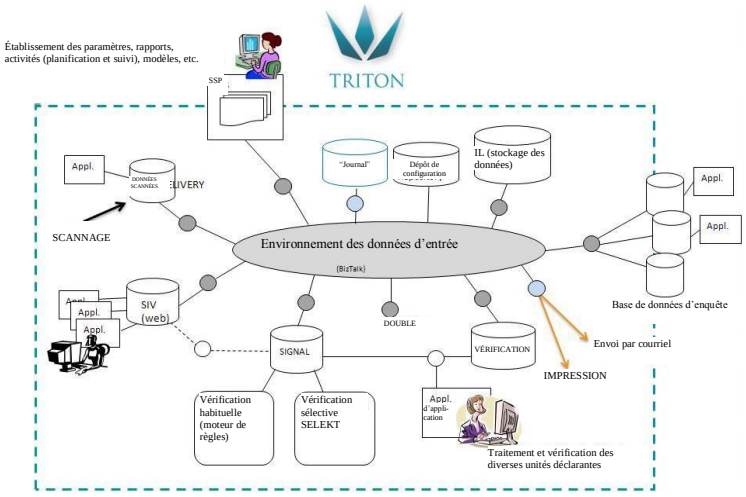
Compte tenu de la nécessité de poursuivre les travaux relatifs à l'AI et de la structure des groupes d'architecture différents concernés, Statistics Sweden met en œuvre les rôles suivants liés à l'architecture : i) architecte organisationnel, ii) architecture de solution, et iii) architecte de logiciel. Parmi les autres compétences figurent notamment les analystes des besoins, le personnel d'essai, les spécialistes de l'infrastructure et les plateformes techniques particulières.

4. Exemples de travaux liés à l'AI

4.1 Plateforme générale de collecte des données

Statistics Sweden est en voie d'élaborer une plateforme générale pour la collecte et la vérification des données (Engdahl, 2010 ; Ericson, 2011). Cela vient appuyer les étapes de construction et de collecte du MSGPO, ainsi que les sous processus 2.3 et 2.6 de l'étape de conception, et 5.1 à 5.4 de l'étape de traitement. Contrairement aux solutions traditionnelles, qui sont axées sur l'entreposage des données, la plateforme est conçue pour appuyer des processus opérationnels au moyen des données et des métadonnées nécessaires (figure 4.1-1). On utilise une approche axée sur les événements, dans laquelle les données et les métadonnées sont transférées entre les processus au moyen d'objets entreprise (dans le cadre d'un modèle d'information organisationnelle). En pratique, une chaîne de valeur, prenant la forme d'une mise en correspondance entre les modèles de processus et les modèles d'information, est créée pour chaque type d'objet entreprise, en déterminant avec soin où l'objet entreprise est créé et quels processus/services devraient être utilisés pour ajouter de la valeur à l'objet. Une plateforme de communication appuie le flux d'information.

Figure 4.1-1
Triton, une plateforme pour la conception d'enquêtes, la collecte de données et le post traitement

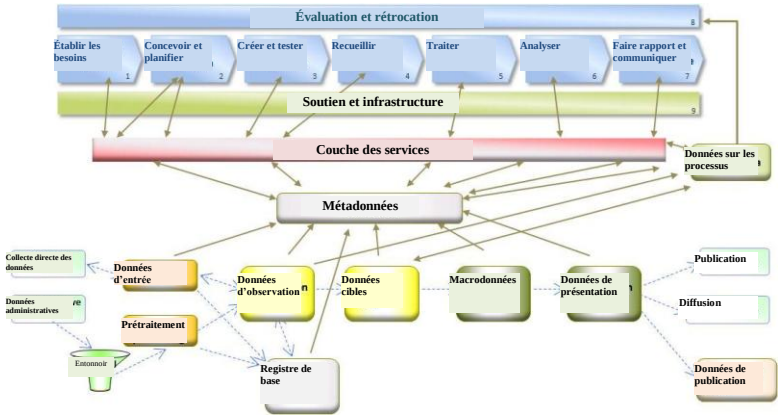


Une des composantes liées à Triton est le système de soutien de la production. Il a vu le jour sous la forme d'un ensemble des lignes directrices en HTML pour l'exécution des activités courantes. Toutefois, il s'est transformé graduellement en environnement de production interactif. Dans cet environnement, les méthodes, outils et approches communs seront disponibles sous forme de services, par l'entremise d'interfaces communes dans lesquelles les paramètres établis pour chaque enquête détermineront les fonctions qui sont appliquées (Bergdahl et Blomqvist, 2011).

4.2 Stratégie d'entreposage des données pour la production de statistiques intégrées

Statistics Sweden en est actuellement aux étapes initiales de la conception d'une plateforme statistique exhaustive, à partir d'une approche faisant intervenir un entrepôt de données uniformes, structurées et bien documentées (figure 4.2-1). L'objectif est d'avoir les registres comme éléments centraux, et sa conception est notamment guidée, entre autres, par les principes de réduction du transfert des données et d'absence d'entreposage des données en double ou quasi en double. La plateforme est conçue afin d'assurer le soutien de méthodes de production axées sur les processus, la collecte efficace des données et la souplesse de la diffusion.

Figure 4.2-1
Aperçu conceptuel complet du système envisagé selon l'approche de l'entrepôt de données



Les travaux relatifs à la plateforme devraient jeter un nouvel éclairage sur le rapport entre le MSGPO et une approche d'entrepôt de données, ainsi que faire ressortir le rôle des métadonnées dans la production moderne de statistiques. Ces travaux comprennent aussi l'établissement plus précis des capacités d'une couche de services qui peut assurer la fonction, tant dans Triton que dans le système d'entrepôt de données.

4.3 Autres travaux liés à l'architecture

Nous procédons actuellement à la modélisation de l'étape de l'établissement des besoins et de l'étape de la diffusion du MSGPO. Au moment du remaniement du registre des entreprises, nous modélisons la collecte des données à partir de sources administratives. Par ailleurs, à l'extérieur du MSGPO, nous avons élaboré des modèles de processus pour le processus de liste de paye et pour les processus intégrés de planification et de contrôle à l'ensemble de l'entreprise. De concert avec trois autres organismes, nous avons élaboré des modèles communs de concept et d'information, afin de faciliter le partage de l'information recueillie.

5. Conclusion

Nous encourageons les INS à commencer simplement par appliquer un cadre d'AI qu'ils considèrent comme approprié pour leurs besoins : cadre de l'architecture intégrée fédérale (ou une autre AI gouvernementale), cadre de Gartner, TOGAF, ou autre chose. Les travaux relatifs à l'AI ne se font pas sur papier, mais plutôt par une intervention directe dans les projets et dans les domaines pertinents de développement.

Nous sommes fermement d'avis que les INS devraient utiliser les cadres d'architecture existants, plutôt qu'inventer leurs propres cadres, les premiers ayant en général été développés à partir de compétences beaucoup plus larges de l'AI que celles dont dispose habituellement un INS. Par ailleurs, nous proposons d'accorder la préférence aux cadres ouverts (par exemple, TOGAF), plutôt qu'aux cadres propriétaires, ainsi qu'aux normes acceptées pour la production de statistiques (par exemple, MSGPO), car cela permettra une intégration plus facile des systèmes de production dans les INS.

Remerciements

Les opinions exprimées dans le présent document sont celles des auteurs et ne reflètent pas forcément les politiques de Statistics Sweden. Les auteurs remercient Anders Holmberg et Hans Irebäck, tous deux de Statistics Sweden, pour leurs commentaires constructifs concernant une ébauche antérieure du document.

Bibliographie

Bergdahl, M. et K. Blomqvist (2011), « National Implementation of the MSGPO – The Swedish Experience », Workshop on Statistical Metadata, Genève, Suisse, 5 au 7 octobre 2011.

Communauté européenne (2009), « On the production method of EU statistics: a vision for the next decade », COM(2009) 404, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF.

Engdahl, J. (2010), « An Event-Driven Architecture for Data Collection », présenté à la réunion MSIS 2010, Daejeon, Corée du Sud, 26 au 29 avril 2010.

Eriksson, J. (2011), « Triton : un outil général de collecte et de microvérification des données », présenté au Symposium international de 2011 sur les questions de méthodologie de Statistique Canada, Ottawa, Canada.

Hofman, F., Renssen, R. et A. Camstra (2011), « Normalisation des processus », présenté au Symposium international de 2011 sur les questions de méthodologie de Statistique Canada, Ottawa, Canada.

Lankhorst, M. et coll. (2009), *Enterprise Architecture at Work: Modelling, Communication and Analysis.* Springer-Verlag New York Inc.

Lopdell, J. et G. Dunnet (2011), « La boîte à outils méthodologique standard de Statistics New Zealand », présenté au Symposium international de 2011 sur les questions de méthodologie de Statistique Canada, Ottawa, Canada.

METIS (2009), « Generic Statistical Business Process Model: Version 4.0 – avril 2009 », Secrétariat de la Commission économique des Nations Unies pour l'Europe, http://www1.unece.org/stat/platform/download/attachments/8683538/MSGPO+Final.pdf?version=1.

OCMIMF (2011), « Generic Statistical Information Model (GSIM), Common Reference Model, Version 0.1 – juin 2011 », disponible à http://www1.unece.org/stat/platform/download/attachments/62751291/GSIM_+Common+Reference+Model+V0_1.docx?version=1.

Scannapieco, M. (2011), « ESSnet CORE Intermediary Report », disponible à http://www.essnet-portal.eu/sites/default/files/79/Del1.2-intermediary_report_v4.doc.

Statistical Data and Metadata exchange (2011), « Framework For SDMX Technical Standards, Version 2.1 », disponible à http://sdmx.org/wp-content/uploads/2011/04/SDMX_2-1_SECTION_1_Framework.pdf.

The Open Group (2009), *The Open Group Architecture Framework (TOGAF), Version 9*, The Open Group: San Francisco, CA.

Vardigan, M., Heus, P. et W. Thomas (2008), « Data documentation initiative: Toward a standard for the social sciences », *The International Journal of Digital Curation*, vol. 3, n° 1.

Zachman, J.A. (1987), « A framework for information systems architecture », *IBM Systems Journal*, vol 26, n° 3, p. 276 à 292.

SÉANCE 10B

EFFETS CALENDRIER ET COHÉRENCE TEMPORELLE

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Étalonnage et prévision : une approche descendante pour combiner les prévisions faites à plusieurs fréquences

Michele A. Trovero, Ed Blair et Michael J. Leonard[1]

Résumé

Les prévisionnistes se servent souvent de données recueillies à divers intervalles de temps (par exemple, données mensuelles et données journalières). Une pratique courante consiste à faire les prévisions indépendamment aux deux intervalles de temps afin de choisir le meilleur modèle pour chaque série, mais elle peut produire des prévisions qui ne concordent pas.

Le présent article montre comment la procédure HPFTEMPRECON de SAS® High-Performance Forecasting utilise la prévision dont la fréquence est la plus faible comme valeur de référence pour ajuster la prévision de fréquence plus élevée de manière à tirer le meilleur parti possible des deux prévisions.

Mots clés : Prévision ; étalonnage ; fréquences multiples ; SAS/HPF ; PROC HPFTEMPRECON.

1. Introduction

Les prévisionnistes doivent souvent produire des prévisions pour une série chronologique particulière à plus d'une fréquence. Par exemple, une entreprise qui offre la réparation sous garantie d'appareils électroménagers pourrait souhaiter prévoir le nombre d'appels quotidiens pour planifier la dotation en personnel et les opérations, telles que la commande de fournitures. L'entreprise pourrait aussi souhaiter prévoir le nombre de demandes de service d'entretien à une fréquence mensuelle pour planifier son expansion à long terme ainsi que les ressources financières nécessaires, par exemple pour l'achat d'un plus grand nombre de véhicules ou le recrutement de nouveaux employés. Le présent article traite du problème de la prévision d'une série chronologique à diverses fréquences, en se concentrant sur les variables de stock. Pour une variable de stock, la série établie à fréquence faible est l'agrégation temporelle de la série à fréquence élevée. Le terme accumulation désigne l'agrégation temporelle et la distingue donc des autres formes, telles que l'agrégation de la série dans une sous-classe qui peut avoir lieu dans le contexte d'une prévision hiérarchique. Le problème de la prévision à plusieurs fréquences est facilement résolu dans un monde idéal où les données sont abondantes, où les séries se comportent de la manière attendue (ce qui signifie qu'elles comprennent principalement des valeurs non nulles et qu'elles sont facilement transformées en une série stationnaire en covariance), et où le modèle correct est choisi pour chaque série. Dans ces conditions, l'accumulation des prévisions faites à fréquence élevée est au moins aussi efficace que les prévisions générées par modélisation de la série à fréquence faible, en ce sens que l'erreur quadratique moyenne de la prédiction faite h pas vers l'avenir dans le premier cas est inférieure ou égale à l'erreur quadratique moyenne de la prédiction faite h pas vers l'avenir dans le second cas. Une description formelle de cet argument pour les processus ARIMA saisonniers figure dans Wei (1990, chapitre 16). L'idée est simple : une prévision (prédiction) est la projection linéaire sur l'espace de Hilbert généré par la série observée. L'espace couvert par les données à fréquence faible est un sous-ensemble de l'espace couvert par les données à fréquence élevée. Par conséquent, l'accumulation de la projection sur l'espace de niveau de détail plus fin produit par les données à fréquence élevée est au moins aussi « proche » de la valeur future réelle que la projection sur l'espace plus grossier couvert par les données à fréquence faible. Une autre façon d'exprimer le même concept, plus simple et ne nécessitant aucun jargon mathématique, consiste à dire que le processus d'accumulation est une forme de compression qui comporte une perte d'information. Les données à fréquence élevée originales ne peuvent pas être générées à nouveau en utilisant uniquement les données accumulées. Par conséquent, les prévisions générées par l'information restreinte contenue dans les données accumulées ne peuvent pas être meilleures que celles

[1]Michele A. Trovero, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Michele.Trovero@sas.com); Ed Blair, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Ed.Blair@sas.com); Michael J. Leonard, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Michael.Leonard@sas.com).

générées au moyen de l'information complète des données non accumulées. Cependant, la réalité prend rarement la forme d'un manuel. Considérons les exemples réels suivants (les noms des sociétés sont omis pour des raisons de confidentialité) :

> **Exemple 1**. La division des pièces de rechange d'une grande société a des activités de portée nationale et gère plus de 40 000 pièces de rechange. Des prévisions journalières trois mois d'avance sont nécessaires pour chaque code postal afin de réapprovisionner les camions de réparation et de prendre des décisions concernant la dotation en personnel. Très peu de pièces sont nécessaires régulièrement. Environ 10 % seulement d'entre elles font l'objet d'une demande plus ou moins régulière pour chaque code postal. Pour les autres pièces, la demande journalière est presque toujours nulle. Des prévisions mensuelles de long terme sont nécessaires pour la production des pièces, le recrutement et les investissements à long terme.
>
> **Exemple 2**. Une grande chaîne de magasins de détail recueille des données au point de vente pour chaque magasin. Des prévisions horaires sont nécessaires à moyen terme pour la dotation en personnel. Les données horaires sont gardées trois mois, après quoi elles sont supprimées en raison du coût de stockage d'une aussi grande quantité de données. Seules les données accumulées à une fréquence journalière sont gardées. Des prévisions mensuelles de long terme sont nécessaires pour planifier l'expansion et le financement.

Dans les deux exemples, des prévisions sont nécessaires à diverses fréquences à des fins différentes. Cependant, il y a tout lieu de penser que l'accumulation des prévisions à fréquence élevée ne donnera pas de bonnes prévisions pour les données à fréquence faible. Dans le premier exemple, la plupart des séries présentent un comportement intermittent. Les séries intermittentes sont constituées principalement d'une valeur unique, habituellement zéro. Les modèles pour des données intermittentes, tels que celui bien connu de Croston (1972), ne permettent pas de saisir des caractéristiques importantes telles que la tendance, la saisonnalité et la dépendance à l'égard d'événements ou d'autres variables externes. En outre, des composantes saisonnières multiples pourraient être présentes dans les données à fréquence élevée, qu'elles soient intermittentes ou non. La modélisation et l'estimation simultanées de plusieurs composantes saisonnières peuvent être complexes et nécessiter d'importants calculs.

Dans le deuxième exemple, la durée des données horaires (fréquence élevée) n'est pas suffisante pour produire des prévisions mensuelles (à fréquence faible) de toute valeur. En effet, on peut soutenir raisonnablement que l'information contenue dans l'historique plus long des données journalières peut être utilisée profitablement pour prévoir les données horaires. Par exemple, pour prendre des décisions de dotation en personnel pour la très importante saison des fêtes en hiver, le détaillant devrait utiliser l'information contenue dans les données journalières, qui couvrent les saisons des fêtes antérieures, au lieu de se fier entièrement aux prévisions des données horaires qui sont fondées uniquement sur les trois mois précédents. En pratique, les prévisions pour les deux fréquences ou plus d'intérêt sont souvent dérivées indépendamment les unes des autres en choisissant à chaque fréquence un modèle qui fournit les meilleurs résultats en fonction de critères, tels que l'erreur absolue moyenne en pourcentage (MAPE, pour *mean absolute percentage error*). Cependant, si les prévisions sont calculées indépendamment, le résultat de l'accumulation des prévisions à fréquence élevée diffère généralement des prévisions produites par le modèle appliqué aux données à fréquence faible.

En outre, comme dans l'exemple 2, on pourrait vouloir utiliser les prévisions à fréquence faible pour améliorer les prévisions à fréquence élevée. Le présent article décrit une méthode de révision des prévisions à fréquence élevée de manière que les prévisions à fréquence faible résultant de leur accumulation soient égales aux prévisions générées par le modèle sélectionné pour les données à fréquence faible. À la première section, nous décrivons la méthode en détail. À la deuxième section, nous présentons la procédure HPFTEMPRECON du moteur de prévision haute performance SAS® High-Performance Forecasting et montrons comment elle permet de rapprocher les prévisions mensuelles des prévisions journalières pour les données sur les compagnies aériennes de Box and Jenkins. À la troisième section, nous présentons les résultats de l'application de la méthode à un ensemble de données constitué de plusieurs séries chronologiques qui présentent un comportement intermittent. Enfin, à la dernière section, nous tirons les conclusions.

2. Méthode

Il arrive souvent, dans le domaine de la statistique des entreprises, de combiner une série de données recueillies à une fréquence élevée avec une série de données plus fiables, mais recueillies moins fréquemment. Par exemple, des enquêtes sont menées trimestriellement auprès d'un sous-échantillon de la population d'intérêt pour déterminer les variations interannuelles, tandis que des enquêtes plus complètes auprès de l'ensemble de la population ne sont menées qu'annuellement. Le processus d'ajustement des données plus fréquentes afin qu'elles concordent avec les données moins fréquentes, mais plus fiables, est appelé étalonnage dans la littérature spécialisée. Denton (1971) a proposé le premier cadre général pour l'étalonnage fondé sur la minimisation d'une fonction quadratique. Une revue récente et complète du sujet peut être consultée dans Dagum et Cholette (2006). Les prévisions à fréquence faible sont également appelées prévisions de référence, tandis que les prévisions à fréquence élevée sont aussi appelées indicateurs prévisionnels. Généralement parlant, les procédures d'étalonnage peuvent être appliquées à n'importe quelle paire de séries mesurées à des intervalles de temps différents. Par conséquent, dans le présent article, nous parlerons de manière plus générale de la série de référence et de la série indicateur pour indiquer les prévisions qui interviennent dans l'étalonnage. Désignons la série indicateur par x_t avec $t = 1, \dots, T$, où t est associé à une date. Désignons la série de référence par a_m, $m = 1, \dots, M$. Les prévisions de référence ont une date de début $t_{1;m}$ et une date de fin $t_{2;m}$, telles que $1 \leq t_{1;m} < t_{2;m} \leq T$. Nous voulons trouver une série étalonnée optimale θ_t, $t = 1, \dots, T$ telle que l'accumulation des séries étalonnées en vue d'obtenir la fréquence des prévisions à fréquence faible soit égale à la série de référence. Autrement dit,

$$\sum_{t=t_{1;m}}^{t_{2;m}} \theta_t = a_m$$

pour $m = 1, \dots, M$.
Le biais est défini comme l'écart prévu entre la série de référence et la série indicateur. Nous pouvons décider s'il faut ou non ajuster la série indicateur originale pour tenir compte du biais. Désignons la série indicateur corrigée du biais par s_t. Si aucune correction du biais n'est effectuée, $s_t = x_t$. La correction du biais additif est donnée par :

$$b = \frac{\sum_{m=1}^{M} a_m - \sum_{m=1}^{M} \sum_{t_{1;m}}^{t_{2;m}} x_t}{\sum_{m=1}^{M} \sum_{t_{1;m}}^{t_{2;m}} 1}.$$

Dans ce cas, l'indicateur corrigé du biais est $s_t = b + x_t$.
La correction du biais multiplicatif est donnée par :

$$b = \frac{\sum_{m=1}^{M} a_m}{\sum_{m=1}^{M} \sum_{t_{1;m}}^{t_{2;m}} x_t}.$$

Dans ce cas, la série corrigée du biais est $s_t = bx_t$. Notons que le biais multiplicatif n'est pas défini si le dénominateur est nul.

Soit $\boldsymbol{s} = [s_1, \dots, s_T]'$ le vecteur de la série indicateur corrigée du biais, et soit $\boldsymbol{\theta} = [\theta_1, \dots, \theta_T]'$ le vecteur de ses valeurs rapprochées. Soit $\boldsymbol{D}$ la matrice diagonale de dimensions *T* x *T* dont les éléments de la diagonale principale sont $d_{t,t} = |s_t|^{\lambda}$, $t = 1, \dots, T$. Indiquons par $\boldsymbol{V}$ la matrice symétrique tridiagonale dont les éléments de la diagonale principale sont $v_{1,1} = v_{T,T} = 1$ et $v_{t,t} = 1 + \rho^2$, $t = 2, \dots, T-1$, et dont les éléments de la sous-diagonale et de la surdiagonale sont $v_{t,t+1} = v_{t+1,t} = -\rho$, $t = 1, \dots, T-1$. Définissons $\boldsymbol{Q} := \boldsymbol{D}^{+}\boldsymbol{V}\boldsymbol{D}^{+}$ et $\boldsymbol{c} := -\boldsymbol{Q}\boldsymbol{s}$, où $\boldsymbol{D}^{+}$ indique la pseudo-inverse de Moore-Penrose de $\boldsymbol{D}$. La série étalonnée (rapprochée) est donnée par les valeurs $\boldsymbol{\theta}_t$, $t = 1, \dots, T$, qui minimisent la fonction quadratique

$$f(\boldsymbol{\theta}; \lambda, \rho) = \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{c}'\boldsymbol{\theta}$$

sous les contraintes

$$\sum_{t=t_{1;m}}^{t_{2;m}} \theta_t = a_m, \qquad m = 1, \dots, M$$

où $0 \leq \rho \leq 1$ et $\lambda \in \mathbb{R}$ sont les paramètres que nous avons sélectionnés. Quand $\boldsymbol{s}$ ne contient pas de zéros, la fonction cible est équivalente à celle proposée par Quenneville et coll. (2006).

Deux problèmes sont pris en considération lorsque l'on procède à l'étalonnage. Le premier consiste à préserver dans la mesure du possible le mouvement dans la série à fréquence élevée (préservation du mouvement). Le deuxième consiste à tenir compte de l'actualité des prévisions de référence, en ce sens que la référence pour la dernière période pourrait faire défaut si la série indicateur s'étend au-delà de la dernière valeur de référence. La correction du biais est un moyen d'améliorer l'actualité de la série de référence en ce sens qu'elle vise à réduire les écarts prévus entre la série de référence et la fonction indicateur. Le paramètre ρ est un paramètre de lissage qui contrôle la préservation du mouvement. Le mouvement de la série originale sera d'autant mieux préservé que la valeur de ρ sera proche de un. Le paramètre λ prend habituellement les valeurs 0, 0,5 ou 1. Pour $\lambda = 0$, nous obtenons le modèle d'étalonnage additif. Pour $\lambda = 0{,}5$ et $\rho = 0$, nous obtenons un modèle d'étalonnage proportionnel.

Dans l'application classique de l'étalonnage, l'objectif est de rétablir l'additivité d'une série désaisonnalisée par rapport à la série de référence. Dans le contexte du présent article, l'objectif est de trouver pour la série à fréquence élevée les prévisions optimales qui respectent la contrainte d'accumulation. Par conséquent, nous suggérons de choisir la correction du biais et les valeurs des paramètres ρ et λ de manière à optimiser le critère de sélection utilisé au départ pour choisir le modèle pour les données à fréquence élevée. Par exemple, si ce modèle est choisi de manière à minimiser la MAPE, les paramètres ρ et λ, et la correction du biais devrait également être choisie de manière à minimiser la MAPE pour les prévisions étalonnées.

Quand $0 \leq \rho < 1$, le problème de minimisation sous contrainte peut être dérivé du problème de régression sous contrainte

$$\begin{aligned} s_t &= \theta_t + c_t e_t \qquad t = 1, \dots, T \\ e_t &= \rho e_{t-1} + \epsilon_t \qquad t = 1, \dots, T \\ \sum_{t=t_{1;m}}^{t_{2:m}} \theta_t &= a_m, \qquad m = 1, \dots, M \end{aligned}$$

où ϵ_t est un bruit blanc de variance σ_ϵ^2, et c_t sont des poids proportionnels à $|s_t|^\lambda$. Par conséquent, quand $\lambda = 0$, le problème de minimisation est équivalent à un problème de régression sous contrainte où l'erreur entre l'indicateur corrigé du biais et la série étalonnée suit un processus AR(1) avec un paramètre autorégressif proportionnel à ρ.

Soit $\boldsymbol{a} = [a_1, a_2, \dots., a_M]']$. L'équation de contrainte peut être réécrite sous la forme

$$\boldsymbol{J\theta} = \boldsymbol{a}$$

où $\boldsymbol{J}$ est une matrice de valeurs 0 et 1 telle que $\boldsymbol{J\theta}$ est l'accumulation de la série étalonnée pour obtenir la fréquence de la série de référence. La solution du problème de minimisation devient alors

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{s} + \boldsymbol{C\Sigma}_e\boldsymbol{CJ}'(\boldsymbol{JC\Sigma}_e\boldsymbol{CJ}')^{-1}(\boldsymbol{a} - \boldsymbol{Js})$$

où $\boldsymbol{C}$ est une matrice diagonale dont les éléments de la diagonale principale sont c_t, et $\boldsymbol{\Sigma}_e$ est la matrice de covariance de e_t. Quand l'étalonnage peut être interprété comme un problème de régression, il est également possible de dériver la covariance des prévisions rapprochées. Voir Quenneville et coll. (2006) pour des renseignements plus détaillés.

Une autre interprétation de cette méthode est de la voir comme un moyen de combiner les prévisions faites à deux fréquences pour produire des prévisions pour la fréquence la plus élevée. Les poids pour la combinaison sont calculés en utilisant la solution du problème de minimisation. Un poids unitaire est attribué aux prévisions à fréquence faible, puisqu'elles fournissent le deuxième membre des équations de contrainte.

3. Procédure HPFTEMPRECON

En utilisant la méthode décrite à la section précédente, la procédure HPFTEMPRECON rapproche les prévisions à fréquence élevée des prévisions à fréquences faibles de façon telle que l'accumulation des prévisions à fréquence élevée rapprochées soit égale aux prévisions à fréquence faible. PROC HPFTEMPRECON rapproche les prévisions faites pour le même item à deux fréquences temporelles différentes dont les intervalles sont emboîtés l'un dans l'autre. Autrement dit, la procédure rapproche une hiérarchie de prévisions à deux niveaux dans la dimension

temporelle. Par exemple, elle rapproche les prévisions mensuelles pour les données sur les passagers des compagnies aériennes de Box et Jenkins (dans le jeu de données Sashelp.Air) des prévisions trimestrielles pour la même série. Par conséquent, la procédure HPFTEMPRECON nécessite non seulement deux jeux de données d'entrée pour les prédictions, mais aussi la spécification des deux fréquences des prévisions au moyen de deux instructions distinctes : l'instruction ID pour les données à fréquence élevée, et l'instruction BENCHID pour les données à fréquence faible.

Les procédures du moteur de prévision haute performance SAS High-Performance Forecasting sont utilisées pour générer les prévisions aux fréquences mensuelle et trimestrielle. Ces prévisions deviennent les données d'entrée de PROC HPFTEMPRECON. Une discussion complète de SAS High-Performance Forecasting dépasse le cadre du présent article. Des renseignements détaillés peuvent être consultés dans *SAS High-Performance Forecasting: User's Guide*.

Premièrement, la procédure HPFESMSPEC génère une spécification de modèle exponentiel de lissage qui est ensuite sélectionnée par la procédure HPFSELECT :

```
proc hpfesmspec
        rep=work.repo
        specname=myesm;
  esm;
run;

proc hpfselect
         rep=work.repo
         name=myselect;
  spec myesm;
run;
```

Puis, les prévisions sont générées au moyen de PROC HPFENGINE aux fréquences mensuelle et trimestrielle en utilisant la spécification sélectionnée du modèle :

```
proc hpfengine
        data=Sashelp.Air
        rep=work.repo
        globalselection=myselect
        out=OutMon
        outfor=OutForMon
        outmodelinfo=OutMod;
  id date interval=month;
  forecast air;
run;

proc hpfengine
        data=Sashelp.Air
        rep=work.repo
        globalselection=myselect
        out=OutQtr
        outfor=OutForQtr
        outmodelinfo=OutModQrt;
  id date interval=qtr accumulate=total;
  forecast air;
run;
```

Il convient de souligner que la variable « air » figure dans l'instruction FORECAST des deux instances de PROC HPFENGINE. L'option INTERVAL= diffère dans les instructions ID. Dans la première instance, l'intervalle de temps ID est le mois; dans la deuxième instance, il s'agit du trimestre. Les prévisions mensuelles sont mémorisées dans la variable PREDICT du jeu de données OutForMon, et les prévisions trimestrielles sont mémorisées dans la variable PREDICT du jeu de données OutForQtr.

Enfin, les prévisions mensuelles sont rapprochées des prévisions trimestrielles en utilisant PROC HPFTEMPRECON :

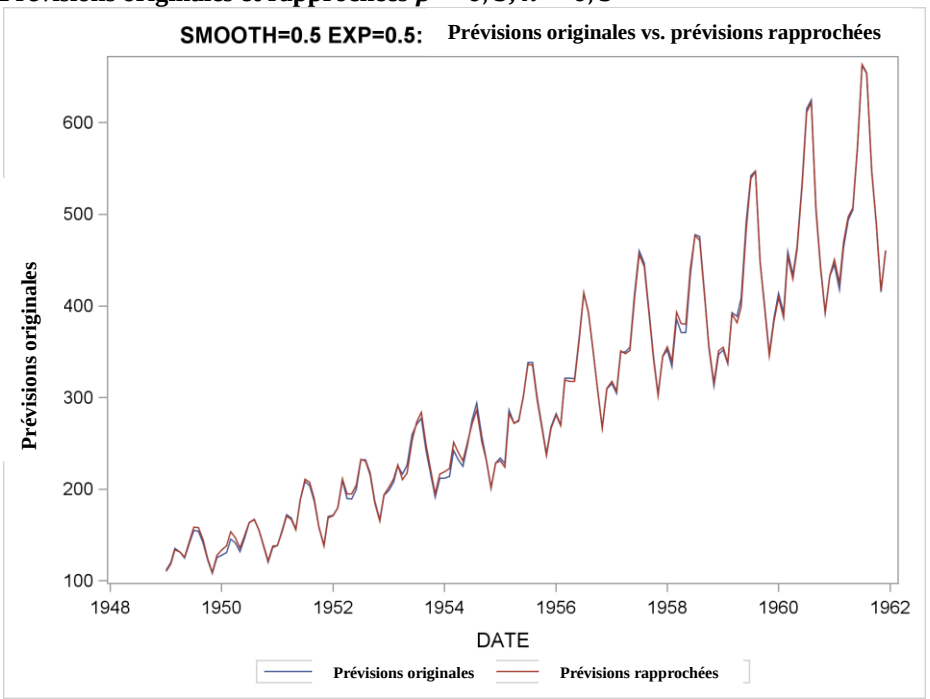
```
proc hpftemprecon
        data=OutForMon
        benchdata=OutForQtr
        outfor=BenFor
        outstat=BenStat
        exp=0.5
        smooth=0.5;
   id date interval=month;
   benchid date interval=qtr;
run;
```

Premièrement, notons que le jeu de données des prévisions mensuelles est l'argument de l'option DATA= dans l'instruction HPFTEMPRECON, et le jeu de données des prévisions trimestrielles est l'argument de l'option BENCHDATA=.

Deuxièmement, notons qu'il existe deux instructions pour spécifier la fréquence des données, une pour chaque jeu de données d'entrée qui contient les prédictions. L'instruction ID est associée au jeu de données DATA= et spécifie la variable qui contient l'indice temporel des prédictions indicateur et sa fréquence relative (intervalle). L'instruction BENCHID est associée au jeu de données BENCHDATA= et spécifie la variable qui contient l'indice temporel des prédictions de référence et sa fréquence relative. Rappelons que l'intervalle de la variable ID doit être entièrement emboîté dans l'intervalle de la variable BENCHID. Par exemple, les mois sont entièrement emboîtés dans les trimestres. Au contraire, les semaines ne sont pas entièrement emboîtées dans les mois, puisqu'une semaine peut s'étendre sur deux mois. Par conséquent, la fréquence de la série indicateur ne peut pas être hebdomadaire quand la fréquence de la série de référence est mensuelle.

Les valeurs des paramètres ρ et λ sont fixées par les options EXP= et SMOOTH=, respectivement, dans l'instruction HPFTEMPRECON. On peut faire varier les prévisions rapprochées en choisissant les valeurs des options SMOOTH= et EXP=. La figure 3-1 montre les prévisions originales comparativement aux prévisions rapprochées quand les deux paramètres sont égaux à 0,5.

Figure 3-1
Prévisions originales et rapprochées $\rho = 0,5, \lambda = 0,5$

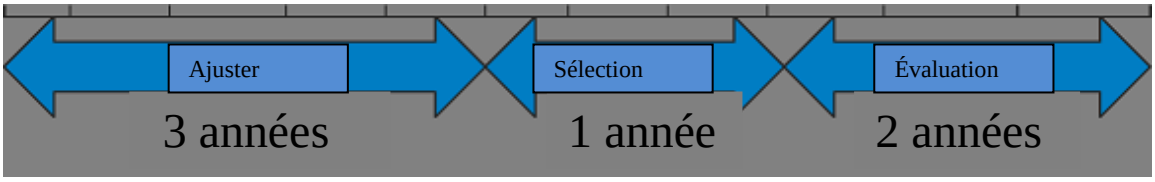


4. Analyse des données

À la présente section, nous appliquons la méthode décrite aux sections qui précèdent à un jeu de données réelles comprenant plusieurs séries chronologiques, dont la plupart présentent un comportement intermittent. Les données sont celles sur la demande mensuelle de la Royal Air Force (RAF) britannique pour 753 pièces qui couvrent une période de six ans allant de juillet 1992 à juin 1998, ce qui donne un total de 72 observations. La demande de pièces de rechange est un exemple type de situation où la demande est habituellement intermittente. Et, effectivement, une majorité des séries de cet ensemble de données présentent un comportement intermittent.

Premièrement, nous générons les prévisions indépendamment à intervalles d'un mois et d'un trimestre. Nous utilisons les données couvrant deux années pour ajuster le modèle. Les données d'une année sont utilisées pour choisir le modèle hors échantillon. Après la sélection du modèle, les paramètres de celui-ci sont de nouveau estimés en utilisant les données des trois années complètes. Cela nous laisse des données de deux années en vue d'évaluer la performance des prévisions. Nous avons utilisé SAS Forecast Server pour effectuer la sélection du modèle. La description détaillée de la procédure de sélection du modèle utilisée est décrite dans Leonard (2002).

Figure 4-1
Sélection et évaluation du modèle



La racine carrée de l'erreur quadratique moyenne (RMSE, *root mean square error*) est choisie comme critère de sélection, parce qu'elle peut être calculée sans équivoque quelle que soit la valeur de la série. L'erreur absolue moyenne en pourcentage (MAPE, *mean absolute percentage error*), qui est le critère de sélection utilisé le plus fréquemment en pratique par les prévisionnistes, est sans signification dans le cas de séries intermittentes.

Les figures 4-2 et 4-3 présentent les familles de modèles choisies pour les données mensuelles et trimestrielles, respectivement. L'examen de ces figures montre que, pour environ 50 % des séries mensuelles, un modèle pour données intermittentes est choisi. Cette proportion est considérablement plus faible pour les données trimestrielles.

Figure 4-2
Répartition de la famille de modèles pour les données mensuelles

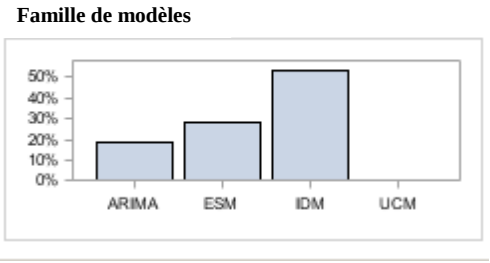
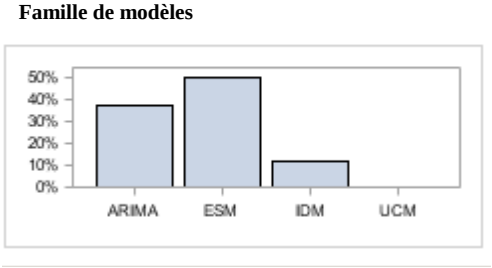


Figure 4-3
Répartition de la famille de modèles pour les données trimestrielles



Les prévisions mensuelles sont rapprochées des prévisions trimestrielles pour une grille de valeurs de ρ et de λ, avec $\rho \in (0, 0.1, 0.2, \ldots, 0.9, 1)$ et $\lambda \in (0, 0.5, 1)$. Pour chaque série, les ensembles de valeurs de ρ et λ sont sélectionnés comme étant ceux qui minimisent la RMSE hors échantillon dans l'intervalle de sélection. Enfin, la RMSE des prévisions rapprochées est comparée à celle des prévisions du modèle original pour la période d'évaluation de deux ans.

La RMSE des prévisions mensuelles rapprochées pour les valeurs choisies de ρ et λ est améliorée pour 562 des 753 séries comparativement à la RMSE du modèle original. L'amélioration moyenne pour ces 562 séries est de 52 %.

5. Conclusion

Le présent article présente une méthode de rapprochement des prévisions faites à fréquence élevée des prévisions faites à fréquence faible pour une série chronologique cumulée dans une hiérarchie d'intervalles de temps. La méthode est fondée sur la minimisation d'une fonction de perte quadratique sous la contrainte que les prévisions à fréquence faible rapprochées s'accumulent pour concorder avec les intervalles de prévision à fréquence élevée. Dans certaines circonstances, le problème peut également être interprété comme un problème de régression. Cette méthode est mise en œuvre en utilisant la procédure SAS HPFTEMPRECON. La fonction cible dépend de deux paramètres dont la sélection peut reposer sur les mêmes critères que ceux utilisés pour sélectionner les modèles pour les prévisions aux deux fréquences. L'application de cette méthode peut aboutir à des prévisions plus exactes quand les données recueillies à fréquence élevée sont principalement intermittentes et ne conviennent par conséquent pas pour les modèles qui comprennent des caractéristiques telles que des variables d'entrée, des événements et des composantes saisonnières.

Bibliographie

Box, G. et G.M. Jenkins (1976), *Time Series Analysis*: *Forecasting and Control*, édition révisée, San Francisco : Holden-Day.

Croston, J.D. (1972), « Forecasting and stock control for intermittent demands », *Operations Research Quarterly*, vol. 23, n° 3.

Dagum, E.B. et P.A. Cholette (2006), « Benchmarking, temporal distribution, and reconciliation methods for Time Series », *Lecture Notes in Statistics*, vol. 186, Springer.

Denton, F. (1971), « Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization ». *Journal of the American Statistical Association*, vol. 66, n° 333, p. 99 à 102.

Leonard, M.J. (2002), « Large Scale Automatic Forecasting: Millions of Forecasts », document présenté lors de l'International Symposium of Forecasting.

Quenneville, B., Fortier, S. Chen, A.-G. et E. Latendresse (2006), « Recent Developments in Benchmarking to Annual Totals in X-12-ARIMA and at Statistics Canada », dans *Proceedings of the Eurostat Conference on Seasonality, Seasonal Adjustment, and Their Implications for Short-Term Analysis and Forecasting*, Luxembourg.

Wei, W.S. (1990), *Time Series Analysis: Univariate and Multivariate Methods*, Redwood: Addison-Wesley.

Amélioration de la calendarisation en utilisant X-12-ARIMA : application aux données sur la TPS

Rossana Manríquez[1]

Résumé

L'Agence du revenu du Canada partage de l'information sur les entreprises avec Statistique Canada sous forme de déclaration de la taxe sur les produits et services. Les déclarations des entreprises peuvent être faites sur une base annuelle, trimestrielle, mensuelle ou plus fréquente. Un processus de calendarisation est nécessaire pour que les transactions soient sur une base mensuelle comparable. Une des entrées de la calendarisation est la série indicatrice qui donne le mouvement mensuel. Nous discutons ici de la construction des séries indicatrices, des améliorations apportées et d'études entreprises afin de valider les séries indicatrices.

Mots clés : Données administratives ; calendarisation ; séries indicatrices ; X-12-ARIMA; révisions.

1. Introduction

Les données administratives ont de plus en plus d'importance au sein des différents programmes d'enquêtes à Statistique Canada. L'Agence du revenu du Canada (ARC) recueille des renseignements sur la taxe sur les produits et services (TPS) auprès des entreprises constituées en société et des entreprises non constituées en personne morale. La TPS est une taxe sur la valeur ajoutée et touche la plupart des biens et services offerts au Canada. Les entreprises perçoivent la TPS et la déclarent à l'ARC; celle-ci partage cette information tous les mois avec Statistique Canada depuis 2003. L'information comprend des données sur le revenu de l'entreprise, le montant de taxe et le crédit de taxe sur les intrants.

Les entreprises doivent déclarer leur revenu à une certaine fréquence, selon la catégorie de revenu. Ainsi, les entreprises ayant un revenu annuel inférieur à 1,5 million de dollars doivent déclarer leur revenu minimalement à une fréquence annuelle; celles ayant un revenu annuel entre 1,5 million et 6 millions de dollars, à une fréquence trimestrielle; celles ayant un revenu supérieur à 6 millions de dollars, à une fréquence mensuelle. Chaque mois, l'ARC envoie à Statistique Canada de nouvelles transactions ainsi que des mises à jour au fichier du mois antérieur. Le fichier peut contenir des transactions remontant à quatre ans en arrière, mais la majorité des transactions sont récentes. Les transactions ont des longueurs et périodicités différentes et ne coïncident pas nécessairement avec un début ou une fin de mois. Statistique Canada applique certains processus aux données, dont la calendarisation qui permet d'obtenir des transactions mensuelles. Les données sur la TPS constituent une source mensuelle de données administratives offrant une solution de rechange au coût et au fardeau de réponse liés aux activités des enquêtes auprès des entreprises. Plusieurs enquêtes infra-annuelles auprès des entreprises et les Comptes nationaux les utilisent. Pour les besoins internes, les données doivent être disponibles rapidement et être de qualité, tout en réduisant le nombre de révisions.

La section 2 définit la calendarisation ainsi que la construction des séries indicatrices. Nous décrivons et comparons deux méthodes de construction de séries indicatrices. La section 3 décrit une étude évaluant si le mouvement trimestriel des séries indicatrices fondées sur des entreprises déclarant leur revenu sur une base mensuelle correspond au mouvement que l'on retrouve dans les entreprises déclarant leur revenu sur une base trimestrielle. La section 4 présente une mesure empirique de la stabilité des séries indicatrices. Une conclusion termine l'article.

[1]Rossana Manríquez, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (rossana.manriquez@statcan.gc.ca).

2. Calendarisation et séries indicatrices

2.1 Généralités sur la calendarisation

La calendarisation est un processus qui consiste à transformer les valeurs d'une série chronologique de flux observées sur différents intervalles de temps en valeurs qui couvrent des intervalles calendrier, tels que le jour, le mois, le trimestre et l'année. Par exemple, prenons le cas d'une entreprise œuvrant dans le commerce de détail qui déclare son revenu sur une base trimestrielle. En particulier, intéressons-nous à la transaction allant du 1er octobre au 31 décembre, pour un revenu total donné. On veut diviser ce revenu total en trois transactions couvrant respectivement les mois d'octobre, novembre et décembre. Diviser le revenu total en trois montants égaux n'est pas la meilleure approche puisque le commerce de détail génère plus de revenus en décembre. Nous avons donc besoin d'une distribution temporelle nous indiquant comment répartir les revenus selon les mois calendrier. Cette distribution temporelle sera ce que nous appellerons séries indicatrices pour la suite de l'article.

Nous employons deux méthodes. L'une qui utilise les techniques de régression (Dagum et Cholette, 2006) et l'autre, l'interpolation linéaire à l'aide d'un spline (Quenneville et coll., 2010). Voici l'équation qui définit la calendarisation en termes de problème de minimisation. Soit ρ et λ, donnés, a_m est le revenu de la me transaction, il faut trouver les valeurs $\hat{\theta}_t, t = 1, \ldots, T,$ qui minimisent cette fonction de θ :

$$\left(1-\rho^2\right)\left(\frac{s_1^* - \theta_1}{\left|s_1^*\right|^\lambda}\right)^2 + \sum_{t=2}^{T}\left\{\left(\frac{s_t^* - \theta_t}{\left|s_t^*\right|^\lambda}\right) - \rho\left(\frac{s_{t-1}^* - \theta_{t-1}}{\left|s_{t-1}^*\right|^\lambda}\right)\right\}^2. \qquad (1)$$

sous les contraintes, $\sum_{t\in m}\theta_t = a_m$. La série s_t^* est la série indicatrice donnant le mouvement que nous souhaitons pour la série calendarisée.

2.2 Construction des séries indicatrices

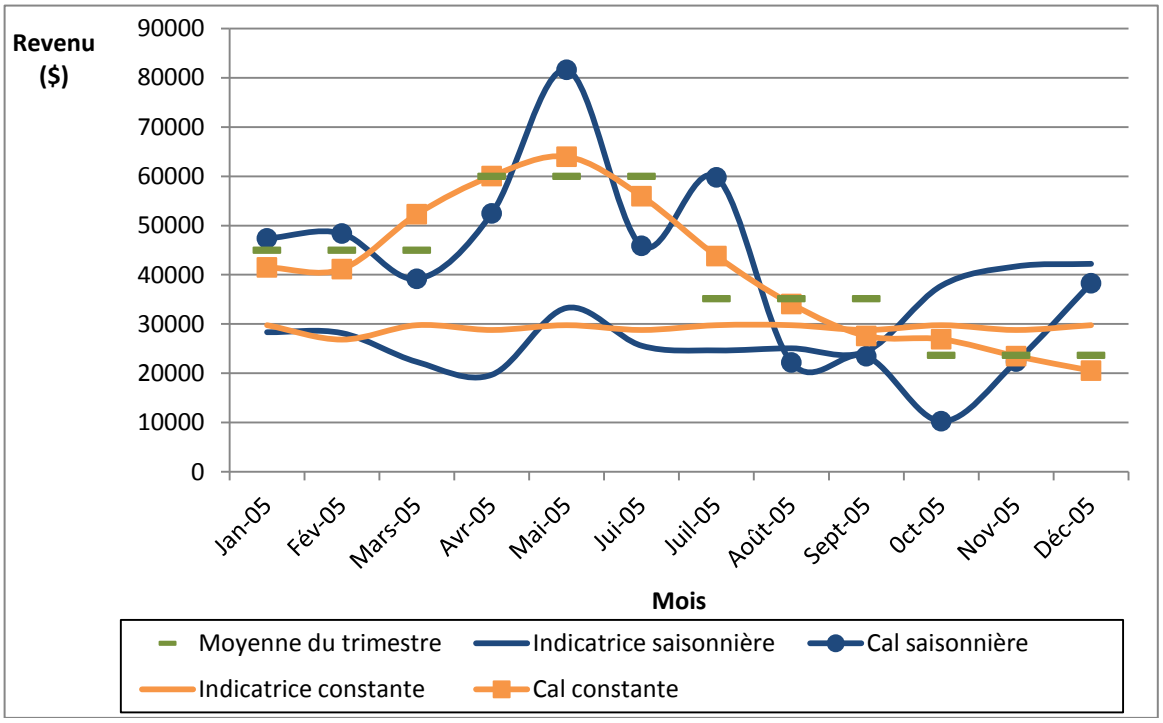
Une des entrées à la calendarisation est la distribution temporelle qui donne le mouvement mensuel de l'industrie, soit la série indicatrice. Le seul mouvement observable et disponible s'obtient comme fonction de certaines entreprises déclarant leur revenu mensuellement. Ces entreprises sont un sous-ensemble (environ 5 %) de la population totale. Il s'agit principalement des entreprises à revenu élevé. En moyenne, sur les quatre dernières années, elles ont eu un revenu annuel 25 fois plus élevé que celui de la population à calendariser. Ces entreprises ont un revenu élevé et sont utilisées pour calendariser les entreprises ayant un revenu moyen ou faible. Il semble peu probable que l'économie se comporte de la même manière pour les entreprises à revenu élevé que pour celles à faible ou moyen revenu. Cependant, la proportion du revenu sur la base qui est calendarisée n'est que de 20 %, donc l'influence du processus de calendarisation au niveau global n'est pas importante.

Il est aussi possible que pour une industrie donnée, le mouvement mensuel soit uniquement l'effet du nombre de jours dans le mois. Alors, une série indicatrice constante qui dépend uniquement de la longueur du mois sera utilisée. À la figure 2-1, nous présentons un exemple de calendarisation à l'aide d'une série indicatrice constante et avec une série indicatrice saisonnière pour une entreprise trimestrielle. Nous observons, par exemple, que le revenu de mars de la série calendarisée à l'aide de la série indicatrice constante est au-dessus de la moyenne du trimestre; le revenu pour le prochain trimestre étant plus élevé. Par contre, avec la série indicatrice saisonnière, nous voyons que la valeur de mars est en dessous de la moyenne, étant donné que la série indicatrice diminue entre février et mars.

La figure 2-1 fait ressortir l'importance de choisir judicieusement les séries indicatrices pour le processus de calendarisation. Nous utilisons le logiciel X-12-ARIMA afin de les explorer. Ce logiciel permet d'estimer les composantes des séries chronologiques : la tendance-cycle, la composante irrégulière, la saisonnalité, l'effet de Pâques et l'effet de jours ouvrables.

Dans le texte qui suit, nous présentons deux méthodes de construction des séries indicatrices : celle utilisée en production jusqu'en septembre 2010 et celle utilisée en ce moment en production. Nous y mentionnons les avantages et désavantages de chaque méthode. Ces séries sont produites au niveau national, pour chaque industrie, pour un total d'environ 1 000 industries.

Figure 2-1
Exemple de données fictives calendarisées à l'aide d'une série indicatrice constante et d'une série indicatrice saisonnière



2.2.1 Construction des séries indicatrices – ancienne méthode

Pour l'ancienne méthode, une étude a été menée afin de déterminer la présence de saisonnalité de chacune des industries. La décision d'utiliser une série indicatrice constante ou saisonnière a été prise, entre autres, en utilisant le logiciel X-12-ARIMA; ses nombreux diagnostics permettant d'identifier la présence ou l'absence de saisonnalité dans une série chronologique.

Dans le cas d'industries non saisonnières, une série constante est utilisée dans la calendarisation. Autrement, les séries indicatrices sont jugées saisonnières et reconstruites chaque mois. Les séries indicatrices ainsi obtenues sont à jour puisqu'elles tiennent compte de chaque changement. Cependant, elles occasionnent des révisions qui ne sont pas toujours justifiables.

Les séries indicatrices sont composées du revenu moyen par code d'industrie de certaines entreprises qui déclarent leur revenu mensuellement. Pour faire partie des contributeurs à la série indicatrice, l'entreprise doit exister au mois de référence le plus récent. Ainsi, d'un mois à l'autre, des entreprises disparaissent ou s'ajoutent aux contributeurs, contribuant à la variabilité de la série indicatrice. De plus, le code d'industrie le plus récent est utilisé, ce qui fait varier les séries indicatrices quand une entreprise change d'industrie à la suite d'une mise à jour de la classification. La mise à jour des déclarations passées des contributeurs est une autre source de révision des séries indicatrices.

Par construction, ces séries indicatrices comprennent toutes les composantes des séries chronologiques. Nous pouvons faire l'hypothèse que la composante irrégulière soit le fait d'une valeur atypique provenant d'une seule entreprise et non pas d'un cas réel d'une composante irrégulière de l'industrie. La tendance cycle permet de suivre l'économie en temps réel, tant que cela n'est pas uniquement dû à un contributeur unique.

Les entreprises qui contribuent à la série indicatrice ont un revenu élevé. Nous pouvons mettre en doute que l'économie aura une incidence sur les entreprises ayant un revenu moyen ou faible de la même façon. Les renseignements sur les entreprises qui sont calendarisées à l'aide de ces séries indicatrices sont révisés tous les mois et sur toute leur histoire, car la série indicatrice change tous les mois. Le passé est donc constamment révisé (Beaulieu et Quenneville, 2008).

2.2.2 Construction des séries indicatrices – méthode actuelle

Une fois de plus, nous dérivons la série indicatrice à l'aide de certaines entreprises qui déclarent leur revenu mensuellement. Cependant, sous la méthode actuelle, nous imposons le même revenu annuel, soit un étalon, à toutes ces entreprises. Plus précisément, nous calendarisons le revenu étalon à l'aide du revenu mensuel de l'entreprise. De cette façon, les mouvements mensuels des entreprises sont de niveaux comparables et ce n'est donc plus l'entreprise ayant le revenu le plus élevé qui dirige le mouvement. Une moyenne est par la suite calculée que nous appelons la « moyenne démocratique ».

La population d'entreprises contribuant aux séries indicatrices n'est pas limitée à celles actives au mois courant. Le nombre d'entreprises est alors accru, car les entreprises contribuent tous les mois où elles sont actives. Le choix de la classification par industrie est aussi différent. Il correspond à la classe au mois de décembre de l'année visée. Enfin, si la série est saisonnière, on extrait les composantes de la série obtenue à partir de la moyenne démocratique. Sinon, nous utilisons une série constante. Nous conservons les facteurs du calendrier, c'est-à-dire que nous excluons la tendance-cycle et la composante irrégulière, mais incluons avec les facteurs saisonniers la composante de Pâques et les jours ouvrables le cas échéant.

Sous la méthode actuelle, la révision des séries indicatrices est effectuée une fois l'an au lieu de mensuellement. Par conséquent, ce que nous perdons en actualité, nous le gagnons en stabilité. Pour les mois non couverts par des données, nous utilisons les prévisions générées par X-12-ARIMA. Au niveau des entreprises à calendariser, la révision due à la calendarisation est effectuée moins souvent; elle est maintenant annuelle au lieu d'être mensuelle.

3. Étude effectuée pour valider la calendarisation des entreprises trimestrielles

3.1 Calendarisation de séries trimestrielles

Nous utilisons le mouvement d'une série indicatrice provenant d'entreprises qui déclarent leur revenu mensuellement afin de calendariser, entre autres, des entreprises qui déclarent leur revenu trimestriellement. Nous voulons valider cette pratique. La distribution mensuelle des entreprises qui déclarent leur revenu trimestriellement n'est pas disponible, donc nous devons faire nos comparaisons au niveau trimestriel. Nous avons abordé le problème de deux façons. Nous avons choisi six codes de l'industrie correspondants aux services de restauration et débits de boissons (France, 2010).

3.1.1 Calendarisation avec des séries indicatrices trimestrielles

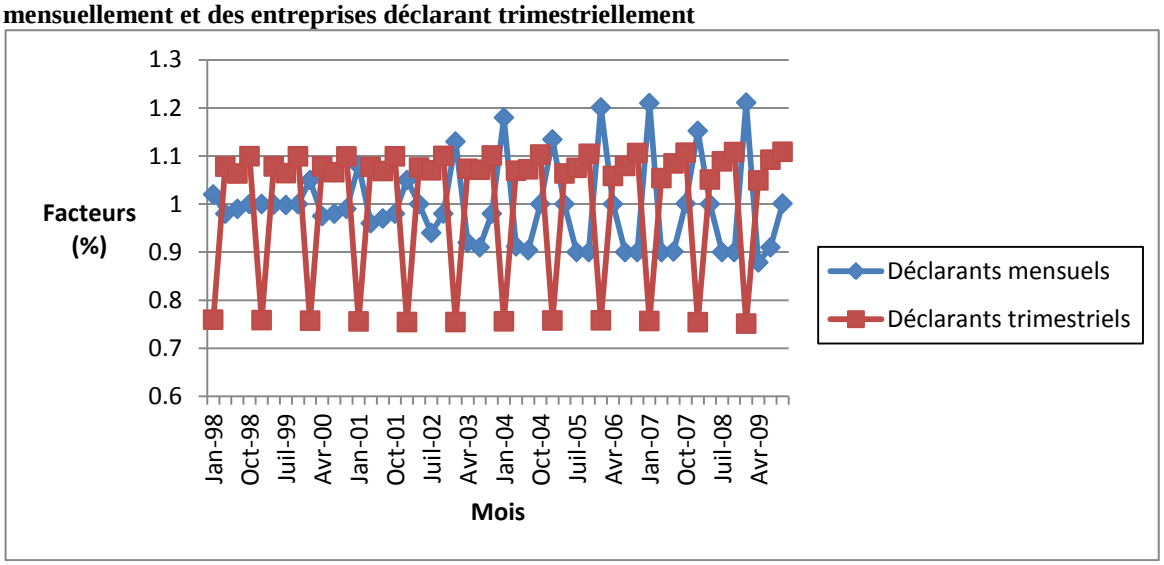
Nous avons construit de nouvelles séries indicatrices au niveau trimestriel. Puis, nous avons pris en compte les entreprises déclarantes selon les trimestres fiscaux. Nous avons extrait les facteurs de calendrier à l'aide de X-12-ARIMA. Et nous avons calendarisé des entreprises devant déclarer leur revenu trimestriellement. Ensuite, nous avons comparé ces résultats aux données de production de ces mêmes entreprises en sommant les revenus des mois pour recréer les trimestres correspondants. Les résultats étaient comparables. Les plus grandes différences ont été observées pour les entreprises ayant moins de huit trimestres déclarés.

3.1.2 Comparaison des facteurs saisonniers

Nous avons pris en compte les contributeurs mensuels de la série indicatrice utilisée en production. Nous avons cumulé les mois en trimestres fiscaux et nous en avons extrait les facteurs de calendrier. Puis, nous avons effectué la comparaison avec les facteurs de calendrier obtenus lors de l'étude précédente.

En général, nous avons obtenu de très bons résultats. Le seul résultat discutable concerne les séries des industries n'ayant pas été identifiées comme saisonnières à l'aide des contributeurs mensuels. Nous donnons un exemple à la figure 3-1. Nous voyons que de l'information auxiliaire ou la connaissance de l'expert en la matière pourrait être essentielle pour construire une série indicatrice prenant en compte la saisonnalité des entreprises à déclaration trimestrielle.

Figure 3-1
Exemple de comparaison de facteurs de calendrier trimestriels obtenus avec des entreprises déclarant mensuellement et des entreprises déclarant trimestriellement



4. Mesure empirique de la stabilité et de la représentativité des séries indicatrices

4.1 Les objectifs de la mesure

La série indicatrice est produite à partir d'un sous-ensemble de la population. Nous ne pouvons pas vérifier qu'elle représente bien le mouvement mensuel de la population entière. Toutefois, nous pouvons vérifier si elle représente la sous-population de laquelle elle est issue et si la saisonnalité est semblable pour chaque contributeur (Delavaquerie, 2011). Nous pouvons également vérifier si elle est influencée par certains éléments qui la composent. Pour ce faire, nous avons choisi de développer une mesure empirique fondée sur la méthode du jackknife (Girard, 2009).

Nous émettons l'hypothèse que plus le nombre de contributeurs est grand plus la série indicatrice sera stable, c'est-à-dire que la suppression d'un contributeur n'aura pas d'influence sur la série indicatrice qui en découle. Afin de valider cette hypothèse, nous ventilerons donc les résultats par nombre de contributeurs dans l'industrie. Nous utilisons déjà une mesure d'ajustement de la désaisonnalisation (statistique M7 de X-12-ARIMA), car nous voulons nous assurer que la mesure empirique de stabilité ne mesure pas la même chose.

Nous avons environ 1 000 industries à vérifier dans un laps de temps assez court, donc nous voulons identifier les industries ayant besoin d'analyses plus approfondies avant de décider d'utiliser une série indicatrice saisonnière ou constante. La mesure devra également répondre à ce besoin.

4.2 La méthode

Nous calculons la moyenne démocratique pour chaque réplique et faisons l'extraction des facteurs de calendrier, à l'aide de X-12-ARIMA, en imposant le modèle multiplicatif et en utilisant les modèles automatiques. Nous obtenons la série indicatrice s_r^* composée des facteurs du calendrier pour la réplique *r*. La série originale s_o^* est celle des facteurs de calendrier extraits de la série construite avec tous les contributeurs. Nous comparons ensuite toutes les

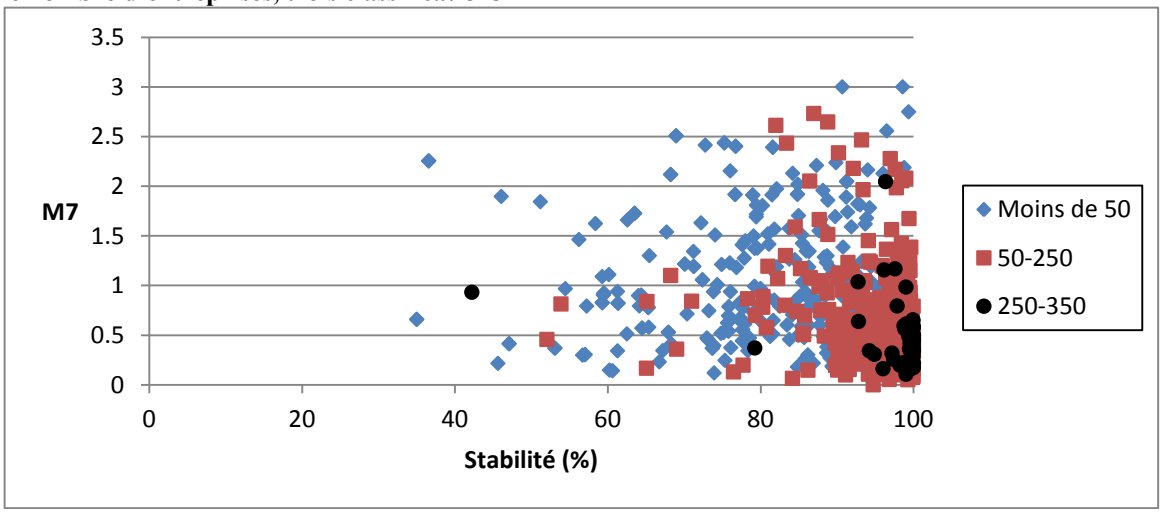
répliques à la série originale et présentons deux façons de résumer l'information. Certaines séries indicatrices pour des industries particulières ne couvrent pas les exigences de X-12-ARIMA. Ces industries ont été mises de côté. Pour des raisons de temps de calcul informatique, nous avons mis de côté les industries ayant plus de 350 contributeurs.

4.2.1 Mesure empirique inspirée du spec Sliding Span de X-12-ARIMA

Pour chaque réplique *r* et chaque mois de référence *t*, nous calculons la mesure suivante $\left|s^*_{t,r} - s^*_{t,o}\right| \Big/ s^*_{t,o}$ et nous utilisons les seuils suggérés par le spec Sliding Span de X-12-ARIMA (U.S Census Bureau, 2009). Si cette distance est supérieure à 3 %, nous dirons que la mesure est instable, autrement elle est stable. Si, pour une industrie, le rapport entre le nombre de mesures stables et le nombre total de mesures est supérieur à 85 %, nous dirons que la série de l'industrie est stable.

À la figure 4-1, nous voyons que la statistique M7 est liée à la stabilité, mais la relation n'est pas parfaite. Nous observons également que la stabilité dépend du nombre de contributeurs dans l'industrie et qu'une industrie avec plus de 250 contributeurs est presque toujours stable. Nous pouvons ici fixer des seuils en dessous desquels la stabilité est jugée insatisfaisante et examiner ces cas plus méticuleusement. Lors de la mise en oeuvre de cette mesure en production, nous pourrons facilement définir des seuils pour identifier les industries ayant besoin d'analyses plus soutenues, par exemple, dans la catégorie 250-350, les deux industries ayant une stabilité inférieure à 85 % devront faire l'objet d'un suivi.

Figure 4-1
Ajustement de la désaisonnalisation en fonction du pourcentage de mesures stables – Industries classées selon le nombre d'entreprises, trois classifications



4.2.2 Mesure empirique inspirée du coefficient de variation

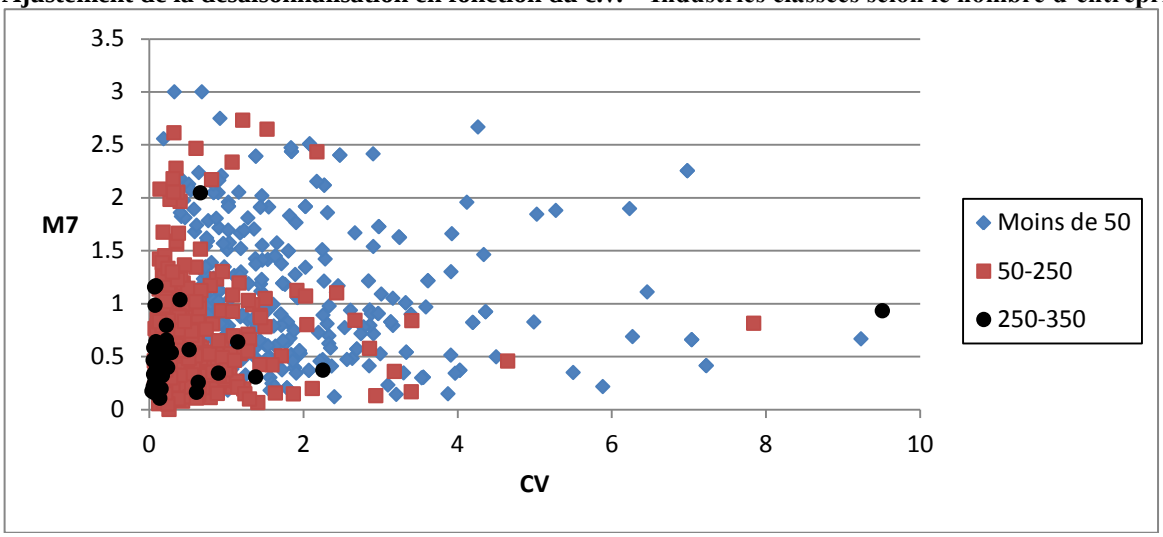
Nous définissons une mesure qui est, à toutes fins pratiques, un coefficient de variation (c.v.) :

$$CV_t = \frac{1,4286 * Médiane_r \left|s^*_{t,r} - s^*_{t,o}\right|}{s^*_{t,o}}. \qquad (2)$$

La correction de 1,4286 permet de dire que, dans un échantillon normalement distribué, la valeur retournée devrait être, en moyenne, approximativement égale à l'écart-type. Ainsi, la mesure peut être vue comme un estimateur non biaisé de l'écart-type dans la population. Nous obtenons une mesure par mois. Nous pouvons donc combiner ces mesures pour étudier une période en particulier, par exemple, le passé récent de la série indicatrice. Pour la suite de l'article, nous prenons en compte la moyenne sur tous les mois disponibles. Nous obtenons ainsi un c.v. par industrie.

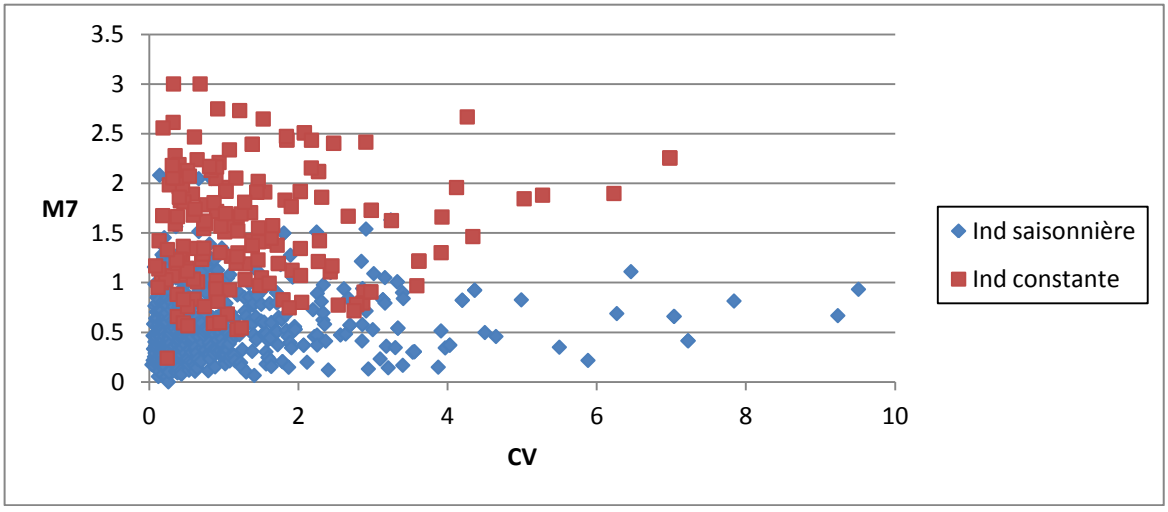
À la figure 4-2, nous voyons que le nombre de contributeurs joue sur la mesure du c.v. tout comme il avait une influence sur le pourcentage de mesures stables à la figure 4-1. Nous déterminerons des seuils différents selon le nombre de contributeurs afin d'identifier les industries à analyser de façon plus approfondie.

Figure 4-2
Ajustement de la désaisonnalisation en fonction du c.v. – Industries classées selon le nombre d'entreprises



À la figure 4-3, nous avons classé l'industrie selon qu'elle est saisonnière ou non. Nous voyons que le c.v. mesuré semble plus petit du côté des industries non saisonnières. Pour les séries indicatrices saisonnières, nous utilisons les valeurs M7 obtenues en production. La décision d'utiliser ou non une série indicatrice saisonnière pour une industrie dépend de plusieurs facteurs, dont la connaissance à priori de l'expert en la matière, le nombre de contributeurs et les révisions potentielles, et non seulement de la désaisonnalisation.

Figure 4-3
Ajustement final de la désaisonnalisation en fonction du c.v. – Classification des industries selon que la série indicatrice est saisonnière ou constante



5. Conclusion

Nous avons récemment amélioré la calendarisation en changeant la construction des séries indicatrices. Les nouvelles séries ont un plus grand nombre de contributeurs et la classification est plus stable lors de la révision des séries indicatrices. Elles sont moins influencées par les entreprises à revenu élevé, puisque le calcul est fait à l'aide de la moyenne démocratique. Ces séries sont maintenant révisées annuellement, ce qui permet de limiter les révisions dues aux séries indicatrices. Nous sommes conscients de la perte d'actualité du côté de la composante de la tendance-cycle. Des efforts devront être déployés pour améliorer cet aspect.

Nous avons vérifié que pour six industries le résultat de la calendarisation semblait valide. Les études sur la calendarisation des entreprises trimestrielles sont encourageantes. Il serait intéressant de reprendre l'étude pour toutes les industries. La mesure inspirée du jackknife semble très prometteuse pour mieux comprendre les contributeurs et gagner en efficacité dans nos efforts à consacrer pour chaque industrie. Nous voulons maintenant mettre en œuvre la mesure en production. Nous voulons également utiliser la mesure correspondant à un c.v. afin d'étudier la stabilité du passé récent des séries indicatrices.

Remerciements

L'auteure aimerait remercier Susie Fortier, Benoit Quenneville et Christian Wolfe pour leurs conseils prodigués tout au long de ces travaux. Des remerciements vont également à Martin Beaulieu, Johanne Boivin, François Brisebois, Pierre Cholette, Estela Bee Dagum, Hugo Delavaquerie, Louis-Baptiste France, Claude Girard, Joanne Leung, Marie-Eve Mainville, Frédéric Picard et les membres du Centre de recherche et d'analyse en séries chronologiques pour leur contribution au projet.

Bibliographie

Beaulieu, M. et B. Quenneville (2008), « Calendarization of the Goods and Services Tax (GST) Data: issues and solutions », *Proceedings of the Joint Statistical Meetings*, Section on Survey Research Methods, JSM 2008.

Dagum, E.B. et P.A. Cholette (2006), « Benchmarking, temporal distribution, and reconcialiation methods for time series », *Lecture Notes in Statistics*, New York: Springer, vol. 186.

Delavaquerie, H. (2011), « Une mesure de la volatilité des facteurs saisonniers », rapport non publié, Ottawa, Canada, Statistique Canada.

France, L.-B. (2010), « Rapport de stage effectué à Statistique Canada », rapport non publié, Ottawa, Canada, Statistique Canada.

Girard, C. (2009), « Un guide d'estimation de la variance pratiquement intelligible », document de travail, Ottawa, Canada, Statistique Canada.

Latendresse, E., Djona, M. et S. Fortier (2007), « Benchmarking sub-annual series to annual totals – from concepts to SAS®Procedure and SAS®EnterpriseGuide®Custom Task », *Proceedings of the 2007 SAS Global Forum.*

Quenneville, B., Picard, F. et S. Fortier (2010), « Interpolation, benchmarking and temporal distribution with natural splines », *Proceedings of the Joint Statistical Meetings*, Business and Economic Section, JSM 2010.

U.S. Census Bureau (2009), *X-12-ARIMA Reference Manuel*, Version 0.3, Washington, D.C.

L'erreur dans les estimations du cycle économique obtenues d'après des données désaisonnalisées

Tucker McElroy[1]

Résumé

Les économistes s'intéressent vivement à la mesure du cycle économique inhérent aux séries chronologiques, mais ils fondent habituellement leurs analyses sur des données désaisonnalisées publiées. Nous cherchons à savoir comment l'estimation du cycle est affectée par la désaisonnalisation fondée sur un modèle et comment il est possible de quantifier l'erreur supplémentaire d'extraction du signal avant la désaisonnalisation.

Nous donnons une description théorique précise des valeurs asymptotiques des estimations du maximum de vraisemblance obtenues en ajustant des modèles spécifiés incorrectement, et utilisons ces pseudo vraies valeurs des paramètres pour quantifier l'erreur quadratique moyenne asymptotique des estimations du cycle calculées d'après des données désaisonnalisées. Un exemple complet est fourni au moyen d'une série de données sur l'emploi.

[1]Tucker McElroy, U.S. Census Bureau, États-Unis.

SÉANCE 11A

CONCEPTION ET UTILISATION DE SYSTÈMES GÉNÉRALISÉS

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Systèmes généralisés : l'expérience de Statistique Canada

Yves Deguire, Laurie Reedman et Michael Wenzowski[1]

Résumé

Statistique Canada a une longue tradition en matière de conception de solutions logicielles généralisées. Notre série de systèmes généralisés couvre la totalité du cycle type de traitement des données d'enquête : de l'échantillonnage et l'estimation, à la vérification et l'imputation, en passant par la totalisation et la protection contre la divulgation.

L'ensemble actuel de systèmes généralisés de Statistique Canada comprend de nombreux systèmes dont l'évolution s'est étendue sur plusieurs générations et plusieurs dizaines d'années. Cela nous a donné de nombreuses occasions de « tirer des leçons » et, par le fait même, nous a permis de concevoir un ensemble de systèmes de plus en plus perfectionnés et efficaces.

Nous présentons un court historique de la motivation, de la spécification, de l'ingénierie et de l'utilisation des systèmes généralisés à Statistique Canada ; des méthodes et techniques utilisées ordinairement pour les définir et pour prescrire leur utilisation ; des défis touchant à l'architecture et à l'ingénierie inhérents au processus de construction ; et des mécanismes de rétroaction et de contrôle mis en œuvre en collaboration avec les utilisateurs respectifs de chaque produit.

Mots clés : Systèmes généralisés ; ingénierie ; architecture.

1. Contexte

Statistique Canada a une longue tradition en matière de création de logiciels de traitement statistique généralisés. Nos premiers systèmes ont vu le jour à la fin des années 1970, et ils étaient initialement axés sur la vérification, la préparation de tableaux et la gestion de bases de données statistiques. Depuis, nous avons produit des systèmes généralisés pour répondre à des exigences comme l'échantillonnage, l'estimation, le contrôle et l'imputation, le codage, l'analyse de séries chronologiques et la prévention de la divulgation.

On ne se trompe pas beaucoup en disant que le programme des systèmes généralisés à Statistique Canada fait partie des ensembles les plus exhaustifs de logiciels de traitement statistique généraux disponibles dans un organisme statistique national. Même si notre ensemble actuel de produits est le résultat de nombreuses décennies d'ajustement progressif et de nouveau développement, on doit souligner que l'élan initial qui a mené à leur création tient à la nature même de l'organisme (Kovar, Jeays et Poirier, 1999). Les systèmes statistiques du gouvernement du Canada sont très centralisés, et la majeure partie du traitement de ce vaste assortiment d'ensembles de données disparates est effectuée par Statistique Canada. Cela a donné lieu très tôt à la reconnaissance des possibilités importantes de réutilisation et de généralisation, les statisticiens étant constamment exposés à des exigences similaires liées à de nombreuses applications de traitement de données différentes.

Le premier ensemble de systèmes généralisés faisait partie d'une architecture de traitement unique (un ordinateur central IBM) et était de nature assez monolithique. Les systèmes généralisés dans leur forme actuelle comprennent des logiciels qui fonctionnent pour une large part dans un environnement de traitement Microsoft, mais aussi à l'intérieur d'une vaste gamme d'environnements de traitement Unix. En outre, nos systèmes actuels sont très modulaires, ce qui permet aux utilisateurs de sélectionner uniquement les fonctions et les modes d'exploitation requis pour une application particulière. Nombre de nos systèmes peuvent être utilisés à la fois en mode par lots et en mode interactif, en plus de pouvoir être intégrés dans un autre environnement logiciel hôte. Ce dernier mode d'exploitation permet de supprimer toute indication externe qu'un système généralisé particulier est utilisé, et offre à l'utilisateur uniquement l'interface présentée par l'application d'hébergement.

[1]Yves Deguire, Laurie Reedman et Michael Wenzowski, Statistique Canada, Ottawa (Ontario) K1A 0T6.

À ces décennies d'évolution technologique du logiciel s'est ajoutée une tendance importante vers l'accroissement et l'amélioration de la convivialité générale (Outrata et Chinnappa, 1989). Cela a pour résultat que nos systèmes n'ont pas besoin de l'intervention d'employés techniques pour pouvoir être utilisés dans une application donnée. Ils nécessitent plutôt une expertise spécialisée et méthodologique, pour veiller à ce que le traitement soit exécuté de façon appropriée.

Les avantages découlant de ces programmes comprennent la production d'applications logicielles très robustes. Cela vient de ce que le potentiel d'application d'une solution généralisée à de nombreuses applications de traitement permet de justifier des coûts de développement initiaux plus élevés. Cela a pour résultat que le logiciel comporte un plus grand nombre de fonctions et fait l'objet d'une meilleure mise à l'essai. En outre, comme le logiciel est utilisé dans les nombreuses applications, le code qu'il comprend est exécuté beaucoup plus fréquemment, et de façon beaucoup plus variée, que dans une application personnalisée type. Ainsi, les problèmes, les bogues et les limites sont décelés très tôt dans le cycle de vie du produit, ce qui profite en dernier ressort aux applications en aval.

Un autre avantage important est lié à l'« uniformisation » de la méthode utilisée dans un domaine posant un problème particulier. Cela a comme avantage de permettre le transfert plus facile d'employés d'un projet à l'autre, étant donné qu'ils sont déjà familiers avec la méthodologie et le logiciel utilisés.

Le tableau qui suit comporte une liste des systèmes généralisés qui constituent l'ensemble actuel de produits de Statistique Canada.

Figure 1-1
Ensemble actuel des systèmes généralisés de Statistique Canada

Nom du système	Fonction
BANFF	Contrôle et imputation
G-Code	Codage
G-Confid	Contrôle de la divulgation
SGE	Pondération et estimation
G-COUP	Couplage d'enregistrements
SGECH	Échantillonnage
G-Series	Séries chronologiques
G-Tab	Totalisation
LogiPlus	Vérification (tables de décision)

Même si, en général, il est simplement plus coûteux et plus long de développer un système généralisé qu'un système personnalisé, l'utilisation d'un tel système peut entraîner des économies importantes (Poirier, 2011). La première découle de l'utilisation du logiciel pour répondre à des besoins multiples de nombreuses applications différentes. La deuxième est liée à la réduction importante des coûts de soutien permanent, du fait du personnel réduit affecté au soutien des nombreuses applications différentes. (Comme de nombreuses applications utilisent le même logiciel, celui ci peut être appuyé par un seul groupe de soutien.) Parmi les autres économies figurent les délais de planification plus courts. Il peut être beaucoup moins long de planifier de nouvelles applications d'enquête si celles ci reposent sur des systèmes généralisés répandus et bien compris.

Pour réaliser un tel programme, il est essentiel que du financement et du soutien soient assurés à un niveau élevé, et qu'un certain degré d'utilisation obligatoire soit prescrit. La centralisation du financement fait en sorte qu'aucune application n'a à supporter seule le coût du développement du logiciel, et l'utilisation prescrite fait en sorte que l'organisme est en mesure de profiter des dépenses en immobilisations et des efforts de développement au niveau prévu. Évidemment, aucun système généralisé ne peut offrir absolument tout à toutes les applications dans lesquelles il est déployé. Le succès du programme dépend dans une large mesure de l'intégration de toutes les fonctions de traitement importantes dans une solution facile à utiliser, ainsi que de la volonté de l'utilisateur final du logiciel d'accepter certains compromis. Nous croyons avoir réussi dans une certaine mesure à établir un tel équilibre dans nos systèmes actuels, et nous continuons de collaborer étroitement avec tous les utilisateurs, tant actuels que potentiels, du logiciel pour faire en sorte que les lacunes et les possibilités d'amélioration soient déterminées tôt, fassent l'objet d'un ordre de priorité et soient incluses dans les plans à plus long terme de soutien de nos produits.

2. Rôles et responsabilités

2.1 Rôles en matière de développement

L'étape du développement repose sur un processus unifié (RUP) bien structuré (Kroll et Kruchten). Au cours de l'étape du développement, trois groupes distincts sont mis à contribution. Les chercheurs statistiques élaborent la méthodologie et construisent les prototypes, les développeurs de méthodologie généralisent la méthodologie et rédigent les spécifications détaillées, et les ingénieurs des systèmes déterminent l'architecture et rédigent le code de programmation.

Tout d'abord, les chercheurs statistiques déterminent la nécessité d'une méthodologie particulière dans une application statistique. Ils élaborent l'idée et l'expriment en termes de concepts, d'algorithmes et d'expressions algébriques. Les chercheurs statistiques construisent un prototype, comme preuve que le concept va fonctionner. Le prototype est mis à l'essai, afin de s'assurer qu'il produit les résultats attendus. Par exemple, on procède à des essais pour s'assurer que les résultats seront exacts lorsque le prototype est utilisé pour des données réelles. On vérifie aussi si le rendement sera acceptable dans des conditions réalistes, en utilisant à la fois des données types et des données extrêmes. Les chercheurs statistiques documentent les fonctions du prototype, la théorie sous jacente et le mode de fonctionnement.

Les développeurs de méthodologie examinent le travail effectué par les chercheurs statistiques, ainsi que les exigences opérationnelles de clients spécialisés, et recommandent à la direction les méthodes à inclure dans les systèmes généralisés. Certaines méthodes se prêtent de toute évidence bien à l'inclusion, comme l'échantillonnage aléatoire simple, la pondération par calage et l'imputation par donneur. Toutefois, dans des domaines comme le contrôle de la divulgation et la coordination de l'échantillon, la recherche se poursuit afin de trouver des méthodes optimales. Il est difficile de déterminer à quel moment une méthode peut être considérée comme suffisamment juste et développée ou comme pouvant s'appliquer à un nombre suffisant de programmes d'enquête différents pour pouvoir être intégrée dans un système généralisé. On trouve souvent un compromis, en établissant un équilibre entre les ressources et la demande (Poirier, 2004). En dernier ressort, la direction donne son appui à l'utilisation d'une méthodologie particulière et, par conséquent, à son développement comme module généralisé pour répondre aux besoins globaux (de plusieurs programmes), plutôt que locaux. Le prototype et sa documentation sont remis aux développeurs de méthodologie, qui les utilisent à l'étape de l'essai, et qui se servent de la documentation pour rédiger les spécifications détaillées.

Les développeurs de méthodologie documentent la façon dont les utilisateurs interagiront avec les modules, les paramètres qui sont nécessaires, les entrées qui sont requises et les produits qui en résulteront. Ils servent à assurer le lien entre les chercheurs statistiques et les ingénieurs des systèmes. Ces méthodologistes comprennent comment le prototype fonctionne et comment il sera utilisé. Ils comprennent les concepts généraux et sont en mesure de rédiger des spécifications détaillées. Parmi les produits livrables dont ils sont chargés figurent les exigences opérationnelles ainsi que les spécifications détaillées. Ces spécifications décrivent les entrées et les sorties en langage simple. Des formules mathématiques, ainsi que des descriptions en toutes lettres, servent à décrire les manipulations auxquelles les entrées doivent être soumises pour obtenir les sorties souhaitées. Les spécifications ne comprennent pas de pseudocodes.

Les ingénieurs des systèmes analysent les spécifications, afin de déterminer comment en optimiser la mise en œuvre. Ils tiennent des rencontres régulières avec les développeurs de méthodologie pour préciser les besoins et déterminer comment les différents modules interagiront les uns avec les autres. Ils examinent les options de mise en œuvre et déterminent l'architecture de système la plus appropriée. Ce n'est qu'une fois les spécifications pleinement analysées et la complexité de la tâche de programmation bien comprise que les ingénieurs des systèmes peuvent prédire avec précision combien de temps l'étape de la programmation durera.

2.2 Mise à l'essai et certification

Les ingénieurs des systèmes mettent à l'essai chaque module, afin de s'assurer qu'il fonctionne selon leur interprétation des spécifications. Les développeurs de méthodologie mettent à l'essai chaque fonction individuellement, ainsi que de façon intégrée avec les autres modules, pour s'assurer qu'elle fonctionne comme

prévu dans des situations typiques ainsi qu'extrêmes. Souvent, les modules sont mis à la disposition d'un ensemble d'utilisateurs d'expérience pour un essai bêta. Il s'agit souvent de méthodologistes qui prévoient utiliser les modules lorsqu'ils seront terminés et qui peuvent les mettre à l'essai dans un environnement réel. Les responsables des essais bêta fournissent de la rétroaction très utile aux développeurs de méthodologie et aux ingénieurs des systèmes. Les essais bêta fournissent aux architectes des systèmes des secteurs spécialisés une occasion très opportune de constater comment les modules s'intègrent dans leurs systèmes de production.

Les étapes du développement et de la mise à l'essai sont itératives. Une fois qu'un module est passé par les diverses étapes d'essai, les développeurs de méthodologie confirment qu'il est complet, et il est officiellement diffusé dans la collectivité des utilisateurs.

2.3 Soutien permanent des utilisateurs

Les développeurs de méthodologie sont responsables d'assurer des communications bilatérales avec la collectivité des utilisateurs, tout au long du cycle de développement. Avec l'appui de la direction, ils font la promotion de l'utilisation des systèmes généralisés, plutôt que d'autres systèmes de « solution locale ». Ils donnent des séminaires pour faire la promotion des méthodes disponibles, ainsi que des cours de formation et des ateliers permettant d'acquérir une expérience pratique. Les développeurs de méthodologie trouvent constamment des solutions aux problèmes des utilisateurs qui ont trait à la méthodologie proprement dite. Ils communiquent aux chercheurs statistiques les besoins supplémentaires qui sont exprimés par les utilisateurs. Cela peut donner lieu à une demande de changement ou à la poursuite du développement (Kozak, 2005). Ils produisent et mettent à jour un guide de l'utilisateur, un didacticiel et de la documentation méthodologique pour accompagner chaque module. L'utilisateur final représente le public cible du guide de l'utilisateur. Il peut s'agir des méthodologistes ou des spécialistes. Le guide de l'utilisateur décrit comment faire exécuter les diverses méthodes au moyen du module, par exemple, les paramètres qui doivent être établis et les entrées qui sont requises. Le didacticiel est aussi élaboré principalement à l'intention des utilisateurs finaux. Il permet un apprentissage en autonomie, grâce à des exemples. La documentation méthodologique décrit la théorie statistique et fournit plus de détails que le guide de l'utilisateur.

Les ingénieurs des systèmes sont responsables de résoudre les problèmes des utilisateurs qui sont liés au logiciel proprement dit. Ils sont aussi responsables du maintien du code de programmation et de la documentation connexe. Ils règlent les bogues et procèdent à des mises à niveau au besoin. Les ingénieurs des systèmes peuvent faire une demande de changement selon les considérations liées au logiciel, par exemple, rédiger à nouveau des parties du code pour profiter des améliorations des versions plus récentes du logiciel de base.

3. Considérations liées au génie logiciel

3.1 Génie logiciel et systèmes généralisés

Le développement de systèmes généralisés produit un logiciel qui permet la mise en œuvre d'algorithmes complexes et le traitement de volumes considérables de données d'enquête. Un tel logiciel doit être de grande qualité et peut uniquement être développé au moyen des pratiques de génie logiciel bien établies (Pressman, 2005).

Le génie logiciel comporte une approche disciplinée et systématique, que vient renforcer l'utilisation d'outils de développement modernes. Le génie logiciel est utilisé tout au long du cycle de vie du développement, qui comprend l'analyse, la conception et la mise en œuvre du logiciel. Parmi les pratiques utilisées à Statistique Canada dans le processus de développement figurent les suivantes : modélisation logicielle, lignes directrices de codage, contrôle des versions, examens des codes et essais unitaires.

3.2 Quatre caractéristiques importantes de logiciel

L'application de pratiques de génie logiciel rigoureuses doit s'accompagner d'une compréhension claire des résultats du processus. Dans le contexte des systèmes généralisés, cela peut être décrit au moyen de quatre caractéristiques que le logiciel élaboré doit posséder.

Adaptabilité

L'adaptabilité s'entend de la capacité à s'adapter à diverses exigences. Un système généralisé, comme son nom l'indique, doit être utilisé pour un grand nombre d'enquêtes et, de ce fait, doit être développé pour tenir compte des spécifications de traitement de la configuration. À cette fin, les systèmes généralisés doivent être développés de façon modulaire très cohérente, chaque module étant chargé d'une fonction statistique particulière (Veryard, 2001). Lorsqu'on ne surcharge pas un module, il est plus facile de développer le logiciel avec souplesse, afin que l'utilisateur puisse facilement modifier les contraintes et les hypothèses (autrement dit, le module doit « reposer sur des paramètres »).

Fiabilité

La fiabilité s'entend de la capacité de produire des résultats exacts au moment opportun. Du fait de leur statut normatif et de leur utilisation à grande échelle, les systèmes généralisés doivent comporter des méthodes statistiques justes, bien comprises et défendables. Ainsi, un système généralisé ne représente pas un terrain propice à la recherche. Les utilisateurs s'attendent aussi à ce que le logiciel soit robuste et à ce que son exécution produise des résultats fiables. Le développement du logiciel s'accompagne donc d'un processus continu d'assurance de la qualité.

Le logiciel doit aussi être efficace pour le traitement de volumes considérables de données et doit pouvoir produire des résultats dans un délai raisonnable. L'efficacité ne doit pas venir au second plan. Elle commence à l'étape de la spécification, par la remise en question des méthodes inefficaces. Les chercheurs statistiques jouent un rôle clé en élaborant des prototypes de ces méthodes, avant qu'on décide de leur mise en œuvre. Le logiciel est aussi conçu et élaboré en tenant compte de l'efficacité tout au long du processus de développement.

Maintenabilité

La maintenabilité s'entend de la capacité d'améliorer les fonctions existantes ou d'ajouter de nouvelles fonctions, ainsi que de s'adapter aux nouveaux environnements opérationnels. Le développement de systèmes généralisés est un processus long et coûteux. Le logiciel qui en découle doit être en production pendant de nombreuses années pour justifier un investissement aussi important de la part de l'organisme statistique. C'est pourquoi le logiciel doit être développé de façon à pouvoir survivre à de nombreux changements dans l'environnement opérationnel et permettre les mises à niveau de ses fonctions.

Comme il a été mentionné précédemment, le logiciel doit être développé sous forme d'un ensemble de modules très cohérents. Chacun de ces modules peut être amélioré, à condition que les améliorations aient trait à la fonction statistique qu'il sert à mettre en œuvre. Les nouvelles fonctions statistiques devraient être développées sous forme de nouveaux modules et venir s'ajouter à l'ensemble.

Le logiciel devrait aussi être développé selon une approche par couches, afin d'isoler les composantes de base de l'environnement opérationnel. L'adoption d'une machine virtuelle d'application, comme SAS® ou Microsoft.Net, joue un rôle essentiel à cet égard. La section qui suit explique de façon plus détaillée les couches de logiciel.

Interopérabilité

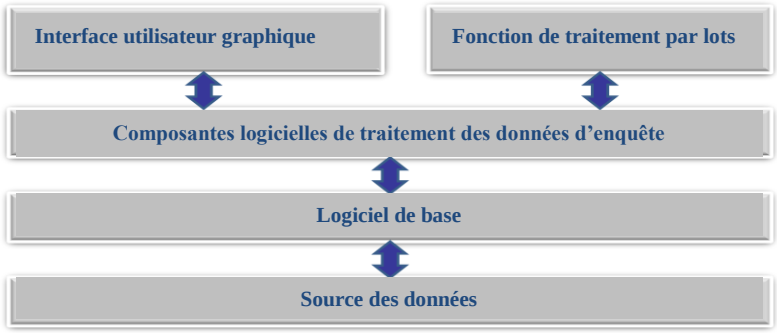
L'interopérabilité s'entend de la capacité d'interagir avec d'autres systèmes et logiciels. Plusieurs modules de systèmes généralisés sont assemblés pour mettre en œuvre une application propre à une enquête. Par ailleurs, les logiciels personnalisés et commerciaux viennent généralement compléter des modules de systèmes généralisés. Cette réalité fait en sorte que ces derniers doivent non seulement être cohérents, mais doivent aussi être relativement autonomes par rapport aux autres modules. Les modules doivent comporter une interface de programmation nette et bien définie. Cette interface permet d'utiliser les sorties d'un module comme entrées dans un autre module, sans avoir à connaître leurs modalités respectives de mise en œuvre interne.

Lorsque les modules sont utilisés à l'intérieur d'un réseau, comme c'est le cas pour une architecture axée sur le service (AAS), il est souhaitable d'avoir recours au langage de balisage extensible (XML). Ce langage constitue une norme dans l'industrie et permet la messagerie entre les modules. Parallèlement, du fait de sa capacité de représenter des structures de données arbitraires, il s'agit d'un bon outil pour définir des instructions de traitement complexes.

3.3 Architecture de logiciel proposée pour les systèmes généralisés

Jusqu'à maintenant, nous nous sommes penchés sur les pratiques de développement permettant de produire des logiciels comportant quatre caractéristiques importantes. Nous proposons maintenant une architecture de logiciel : il s'agit essentiellement des plans détaillés qui organisent les différents éléments de logiciel qui composent les systèmes généralisés. De façon plus particulière, nous présentons l'architecture modulaire à plusieurs niveaux qui a été adoptée par Statistique Canada pour le développement des systèmes généralisés. L'architecture est représentée graphiquement dans le diagramme qui suit.

Figure 3.3-1
Représentation graphique de l'architecture modulaire à plusieurs niveaux des systèmes généralisés



Comme son nom l'indique, cette architecture repose sur des composantes logicielles et est organisée en couches appelées niveaux (Heineman et Councill, 2001). Nous avons déjà présenté la notion de composantes logicielles en décrivant les modules très cohérents couplés de façon lâche comme un moyen d'obtenir un logiciel adaptable, fiable, modifiable et interexploitable. Dans le contexte de l'architecture logicielle, on les appelle composantes de logiciel de traitement d'enquête. Il s'agit de modules de haut niveau qui servent à mettre en œuvre les méthodes statistiques. De ce fait, elles sont au centre de cette architecture.

La notion de couches de logiciel a déjà été présentée brièvement. Nous avons abordé l'importance d'isoler les composantes logicielles au moyen d'une machine virtuelle d'application. Une architecture à plusieurs niveaux va beaucoup plus loin, étant donné qu'elle permet de distinguer la présentation, le traitement d'application et la gestion des données. En répartissant un système en niveaux, on peut déployer le logiciel dans plusieurs ordinateurs, afin de tirer parti de la puissance de calcul des serveurs et de permettre la décentralisation des systèmes. Le logiciel peut aussi être développé un niveau à la fois. Qui plus est, un niveau particulier peut faire l'objet d'un nouveau développement sans affecter les autres niveaux, ce qui réduit le coût d'entretien du système global au fil du temps. L'architecture proposée dans ce cas comprend cinq niveaux.

Sources des données

Les systèmes généralisés doivent permettre d'accéder aux données en divers formats dans plusieurs systèmes disparates. Cette couche représente simplement les divers outils d'entreposage des données, comme les bases de données et les fichiers séquentiels.

Logiciel de base

C'est à cette étape que la machine virtuelle d'application se matérialise. Il s'agit d'un ensemble riche de fonctions et de modules pour le traitement et l'analyse statistique, qu'un système généralisé permet simplement d'enrichir. Il s'agit aussi d'un environnement d'exécution très évolutif, qui convient bien au traitement de volumes élevés de données par lots, un scénario typique dans le traitement des données d'enquête. Le logiciel de base permet la création de « chaînes de montage », qui permettent d'exécuter un certain nombre d'étapes, en série et/ou en parallèle. SAS® représente un bon exemple de logiciel de base comportant ces caractéristiques.

Composantes logicielles de traitement des données d'enquête

Les composantes logicielles sont habituellement mises en œuvre au moyen d'un langage de programmation bien connu et ayant un bon soutien, comme C et SAS®.

Interface utilisateur graphique

Il s'agit d'une couche de présentation qui permet l'utilisation visuelle et interactive des composantes logicielles. Ce niveau est optionnel et est habituellement développé à la deuxième étape du projet de développement parce qu'il exige que les composantes logicielles de base soient en place avant de pouvoir être mis en œuvre. Le développement et le maintien d'interfaces utilisateur nécessitent un effort important. Les interfaces sont sujettes à des mises à jour fréquentes, parce qu'elles dépendent dans une large mesure d'une technologie qui évolue rapidement et qui est aussi très subjective. À cet égard, les systèmes généralisés les plus récents ont permis le développement d'interfaces graphiques (GUI) sous forme de modules d'extension d'interfaces existantes. Cela réduit la portée de l'interface et le coût de son développement et de mise à jour. Le développement de tâches personnalisées SAS® Enterprise Guide®, au moyen du langage de programmation VB.Net ou C#, en est un bon exemple.

Fonction de traitement par lots

Cette couche de présentation optionnelle vise à faciliter la création des tâches de production, au moyen de composantes logicielles préfabriquées (généralisées, personnalisées, commerciales et ouvertes) et grâce à des étapes de traitement propres aux enquêtes. Une approche consiste à définir une série de tâches faisant appel aux diverses composantes et étapes du traitement au moyen d'un répertoire de métadonnées. Les tâches proprement dites existent uniquement pendant la durée de l'exécution. Elles sont produites au moment de l'exécution à partir de métadonnées. Cette approche permet à l'utilisateur de créer et de modifier les tâches de production avec un minimum d'interventions de la part des TI.

Dans l'ensemble, l'architecture proposée met l'accent sur la souplesse, la réutilisation, le faible coût et la facilité d'utilisation. Elle maximise l'utilisation des systèmes généralisés, en permettant d'adapter les composantes aux besoins de chaque enquête et d'insérer d'autres composantes au besoin. Le coût est relativement faible parce que la majeure partie des efforts est consacrée au développement du logiciel propre aux systèmes généralisés, tout en permettant le déploiement du logiciel qui en résulte dans un environnement de traitement décentralisé.

4. Conclusions

Il ne fait aucun doute que le développement de la méthodologie, la rédaction des spécifications, la rédaction du code de programmation et la mise à l'essai constituent un processus long et complexe pour les systèmes généralisés. Toutefois, cela a pour avantage une utilisation plus efficace des ressources à long terme et des outils plus robustes, qui peuvent être mis à jour et s'ajouter au fil du temps. Un partenariat étroit entre la Méthodologie, l'Informatique et les secteurs spécialisés joue un rôle clé pour la réussite du développement de systèmes généralisés, en vue de l'exécution de fonctions statistiques complexes. La gouvernance est essentielle pour garantir l'adoption de ces systèmes.

Bibliographie

Heineman, G.T. et W.T. Councill (2001), *Component-Based Software Engineering: Putting the Pieces Together*, 1e édition, Addison-Wesley Professional.

Kovar, J., Jeays, M. et C. Poirier (1999), « Generalized Systems: Where are we at and where are we going », document interne présenté au Comité consultatif des méthodes statistiques, réunion no 28, avril 1999.

Kozak, R. (2005), « Le système Banff pour l'éditage et l'imputation automatique », *Actes du Groupe des méthodes d'enquête*, Congrès annuel de la Société statistique du Canada, juin 2005.

Kroll, P. et P. Kruchten (2003), *The Rational Unified Process Made Easy*, 1e édition, Addison Wesley Professional.

Outrata, E. et B.N. Chinnappa (1989), « General survey function design at Statistics Canada », *Bulletin of the International Statistical Institute*, vol. 53, no 2, p. 219 à 238.

Poirier, C. (2004), « The Processing Environment Behind a Statistical Program », *Actes du Groupe des méthodes d'enquête*, Congrès annuel de la Société statistique du Canada, juin 2004.

Poirier, C. (2011), « The Impact of a Changing Business Architecture on Editing », Séance de travail de la Commission économique des Nations Unies pour l'Europe sur la révision des données statistiques, Slovénie, mai 2011.

Pressman, R.S. (2004), *Software Engineering: A Practitioner's Approach*, 6e édition, McGraw-Hill Science.

Veryard, R. (2001), *The Component-based business: Plug and Play*, 1e édition, London : Springer.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Triton : un outil général de collecte et de microvérification des données

Johan Erikson[1]

Résumé

Triton est un projet en cours dont l'objectif est d'établir un environnement de production général mais souple pour la collecte et la microvérification des données. Le but est de couvrir la plupart des types d'enquête, mais en guise de première étape, le projet est axé sur des enquêtes dans lesquelles les données sont recueillies directement au moyen de questionnaires (en ligne ou papier). Bien qu'une version de la plateforme soit déjà utilisée, une nouvelle version, considérablement améliorée, est en cours d'élaboration. Celle ci sera lancée à la fin de juin 2011. L'adoption de la nouvelle plateforme a pour but de remplacer bon nombre des anciens systèmes de TI particuliers aux enquêtes, de l'utiliser pour la majorité des enquêtes réalisées à Statistics Sweden, d'intégrer les outils communs déjà en place et d'éliminer autant de travail manuel que possible. Certains des avantages attendus les plus importants de la plateforme sont que les métadonnées auront un effet réel sur le processus de production, que l'assurance de la qualité sera intégrée dans ce processus et que la production de nombreuses enquêtes sera normalisée, ce qui facilitera le regroupement des ressources. Outre l'intégration des outils communs existants, tels que l'outil de collecte en ligne et le système de balayage électronique, la plateforme comprendra trois nouveaux éléments principaux : un outil d'administration ou de conception pour l'établissement des paramètres d'une enquête particulière et la surveillance de la progression de l'enquête, un outil pour le travail avec des objets individuels et une plateforme de communication reliant tous les éléments de la plateforme. Dans le cadre du présent article, divers éléments de la plateforme et la façon dont ils sont utilisés dans le travail quotidien seront présentés, en prenant une enquête particulière comme exemple.

Mots clés : Systèmes généralisés ; standardisation.

1. Introduction

Le résumé ci dessus a été rédigé au printemps 2011, alors que le présent article l'a été à la fin de 2011. Par conséquent, certains énoncés du résumé ne sont plus d'actualité. La nouvelle version de Triton a été diffusée à la fin de juin 2011 comme prévu et a été mise en œuvre dans six enquêtes pendant la deuxième moitié de l'année. En 2012, il est prévu qu'environ 20 programmes d'enquête commencent à utiliser la plateforme.

2. Contexte

Le projet Triton comprend plusieurs facteurs qui, ensemble, constituaient un bon point de départ pour les travaux.

1. Statistics Sweden est en train de passer à un système de production statistique axé sur les processus, en utilisant dans la mesure du possible des outils, des méthodes et des routines communs. Ces outils, méthodes et routines sont décrits dans le système de soutien des processus (SSP) qui a été créé en 2008 et a, jusqu'à présent, été principalement une banque d'information sur ce qu'il convient de faire dans diverses situations. L'objectif à long terme en ce qui concerne le SSP est qu'il devienne un outil plus interactif effectivement utilisé pour guider des processus de production.
2. Statistics Sweden vise aussi à évoluer vers un système de production statistique guidé par les métadonnées, dans lequel les choix en matière de conception auront une incidence directe sur les outils de TI utilisés et sur les étapes du système de production proprement dit dans lesquelles sera intégrée une plus grande assurance de la qualité.
3. Le processus de collecte des données s'appuie déjà sur un certain nombre d'outils communs, dont un outil de collecte des données en ligne, un outil de collecte des données par interview téléphonique et un outil de balayage optique des questionnaires papier. Cependant, avant le lancement du projet Triton, ces outils étaient – du moins pour les enquêtes sur les entreprises et sur les administrations publiques – utilisés comme ajouts aux

[1]Johan Erikson, Statistics Sweden, Process Department, Örebro, SE-701 89, Suède, johan.erikson@scb.se.

systèmes de production propres aux enquêtes. Les passerelles entre les différents outils étaient mauvaises, de sorte que beaucoup de manipulations et de transformations manuelles des données étaient nécessaires. En outre, les systèmes de production propres aux enquêtes vieillissent, ont souvent été construits selon des techniques dépassées aujourd'hui et, dans de nombreux cas, dépendent pour leur maintenance, de personnes particulières, souvent celles qui ont construit le système il y a un certain nombre d'années.

4. La collecte des données (y compris la microvérification) des enquêtes sur les entreprises et les administrations publiques n'a été centralisée que dans les dernières années et était effectuée auparavant par les divers secteurs spécialisés. La centralisation a, en soi, rendu nécessaire l'utilisation d'un plus grand nombre d'outils communs afin de réaliser les gains qui doivent en découler, c'est à dire une production plus efficace et plus rationnelle, regroupant les ressources, *etc*.

Les troisième et quatrième facteurs ont été les principales raisons du lancement d'un projet en vue de créer un environnement de production généralisé pour la collecte et la microvérification des données – le projet Triton. À mesure que le projet a progressé, il est devenu évident qu'il s'agissait aussi d'une occasion de faire un grand pas vers la réalisation des objectifs spécifiés aux deux premiers points susmentionnés. La dernière partie du projet, qui consiste à faire évoluer la plateforme Triton d'une version prototype à une version plus facilement accessible qui pourrait être utilisée par un grand nombre de programmes d'enquête, a par conséquent abouti à la décision d'installer la nouvelle version de la plateforme dans une version mise à jour du système SSP, et de faire ainsi un grand pas vers tous les objectifs décrits plus haut.

La présentation de l'article est la suivante. La section 3 décrit le SSP à Statistics Sweden et les travaux qui ont été effectués dans le cadre du projet Triton pour l'élargir. La section 4 décrit les différents éléments de la plateforme Triton et la façon dont ils fonctionnent aujourd'hui, et offre certaines réflexions quant à l'avenir et les prochaines étapes. La section 5 résume les résultats.

3. Le système de soutien des processus à Statistics Sweden

En plus d'évoluer vers un système de production statistique axé sur les processus, Statistics Sweden a, comme la plupart des autres pays, sélectionné un modèle de processus qui décrit le processus de production statistique. Comme celui de nombreux autres pays, le modèle de processus est similaire, mais pas identique, au Modèle générique du processus de production statistique élaboré par les Nations Unies. Le modèle de processus adopté à Statistics Sweden subdivise le processus de production statistique en huit sous processus :

1. Déterminer les besoins
2. Concevoir et planifier
3. Construire et tester
4. Recueillir les données
5. Traiter
6. Analyser
7. Diffuser et communiquer
8. Évaluer et commenter en retour

Un neuvième sous processus, « Soutien et infrastructure », est également défini, mais est un peu en dehors du processus de production statistique.

Cinq propriétaires de processus ont été désignés (pour les processus 1+7, 2+3, 4, 5+6 et 8+9). Il leur incombe de fournir aux programmes d'enquête des méthodes, des outils et des routines de travail appliqués à l'échelle de l'organisme en se fondant sur les divers besoins, et d'entrer l'information dans le système de soutien des processus (SSP). L'information enregistrée dans le SSP couvre les méthodes, les routines et les outils communs qu'il convient d'utiliser pour exécuter le processus de production statistique aussi efficacement que possible. La description est faite de la façon suivante : chaque sous processus est à son tour subdivisé en sous processus à un niveau plus fin de détail (le niveau de ventilation varie de un à cinq niveaux supplémentaires). Chaque sous processus est décrit dans un modèle comprenant quatre éléments principaux : une brève description de la raison d'être du processus, les intrants nécessaires, la partie principale qui décrit comment exécuter les sous processus, et le produit qui émane du sous processus. La partie principale comprend des renseignements détaillés sur les outils et les méthodes à utiliser et le moment où il faut les utiliser – au besoin, des renseignements encore plus détaillés, des instructions, des guides de

l'utilisateur, *etc.*, sont placés dans des documents supplémentaires qui peuvent être consultés à partir de la page du sous processus – et en outre comprend souvent une description étape par étape des phases par lesquelles il faut passer pour exécuter le processus. Elle peut aussi inclure des modèles ou des listes de vérification lorsque ces documents sont pertinents. Si plusieurs méthodes et routines peuvent ou doivent être utilisées dans diverses situations, elles sont chacune décrites en mentionnant pour le moment où il convient de l'utiliser. Globalement, on peut dire que le SSP contient les normes recommandées par Statistics Sweden.

Le SSP contient beaucoup d'information, si bien qu'un grand nombre d'employés de Statistics Sweden ont de la difficulté à trouver tous les renseignements pertinents et nécessaires, et une question fréquemment posée par des programmes d'enquête particuliers est celle de savoir quelles parties du système sont pertinentes pour eux. Les progrès réalisés dans le cadre du projet Triton en collaboration avec d'autres spécialistes d'un projet de SSP ont permis de résoudre ce problème. Une nouvelle partie du SSP, appelée domaines de processus, a été créée. Ces domaines de processus, qui correspondent à des domaines particuliers d'enquête, permettent aux diverses enquêtes concernées d'avoir accès à l'information, aux outils ou aux documents qui les intéressent. Les domaines de processus sont créés en Microsoft Sharepoint, ce qui donne aussi la possibilité d'utiliser les fonctions intégrées dans ce logiciel. Le domaine de processus est le tableau de bord à partir duquel le programme d'enquête peut exécuter la production. Pour le moment, les domaines de processus ne couvrent que la collecte et la microvérification des données, parce que ce sont les processus couverts par le projet Triton, mais dans l'avenir, le concept des domaines de processus et de la fonction de tableau de bord seront étendus à d'autres éléments du processus de production statistique. Afin de pouvoir exécuter des enquêtes permanentes de divers types et de permettre que les choix de conception varient au cours du temps pour une enquête particulière, les domaines de processus ont une configuration hiérarchique à trois niveaux, à savoir l'enquête, le cycle d'enquête et le cycle de collecte. Chaque cycle de collecte peut être surveillé individuellement, mais il existe également certaines possibilités de réutiliser l'information d'un cycle à l'autre (accroître les possibilités de réutiliser l'information est l'une des grandes priorités des futurs travaux de développement). Lorsque nous étendrons le concept du domaine de processus à d'autres processus, nous envisageons de créer des domaines tels que « cycle de traitement » et « cycle de publication ».

Pour ce qui est de fournir à chaque programme d'enquête les renseignements dont il a besoin, en fonction des propriétés de l'enquête, voici ce qui a été fait :

1. Sur la base des renseignements figurant dans le SSP « ordinaire », nous avons créé un certain nombre d'activités. Ces dernières sont les tâches qui doivent être exécutées pour que le processus de production (collecte et microvérification des données) soit mené à bien. Chaque activité est dotée d'une description (la petite partie particulière du SSP ordinaire qui couvre l'activité en question). Au besoin, un document (tel qu'un modèle, une liste de contrôle ou une instruction plus détaillée) est annexé à l'activité. Pour les cas où différents documents existent pour différents types d'enquête (par exemple des modèles différents à utiliser pour une enquête-entreprise et une enquête-ménage), deux activités différentes, mais similaires, ont été créées, en annexant le modèle pertinent à chacune.
2. Un formulaire (qui peut également être décrit comme un questionnaire) devant être rempli par le programme d'enquête qui souhaite utiliser un domaine de processus a été créé. Le programme d'enquête y inscrit ses propriétés et ses choix de conception (à un niveau assez élevé).
3. Chaque activité a été reliée au formulaire dans un format simple (« si la réponse à la question x est y, afficher cette activité »). L'énoncé peut être rendu plus compliqué en utilisant des énoncés « et »/ « ou ».
4. Le programme d'enquête crée un domaine de processus en partant de ses cycles d'enquête et de ses cycles de collecte, et remplit le formulaire concernant la conception et les propriétés. Le formulaire est interprété par le moteur sous jacent des domaines de processus et de Triton, et les activités pertinentes sont copiées dans le nouveau domaine de cycle de collecte. Cela signifie que le programme d'enquête dispose alors de tous les renseignements et documents nécessaires pour ses activités de collecte et de microvérification des données.
5. Les activités peuvent être assignées à une personne particulière, et elles peuvent également être cochées quand elles sont terminées. Cela donne au gestionnaire de la production la possibilité d'orienter et de surveiller les activités. Les activités sont une caractéristique importante de la fonctionnalité tableau de bord des domaines de processus.

La fonctionnalité tableau de bord des domaines de processus comprend aussi un certain nombre d'autres fonctions importantes pour l'exécution de l'enquête (elles sont décrites plus en détail à la section 4) :

- Possibilité d'affecter les personnes à divers rôles dans le processus de production et de leur donner accès à l'enquête et à ses outils.
- Définition et description des variables de collecte.
- Accès aux rapports sur les paradonnées du processus de collecte, *etc*.
- Accès aux outils nécessaires pour configurer la collecte et la microvérification des données.

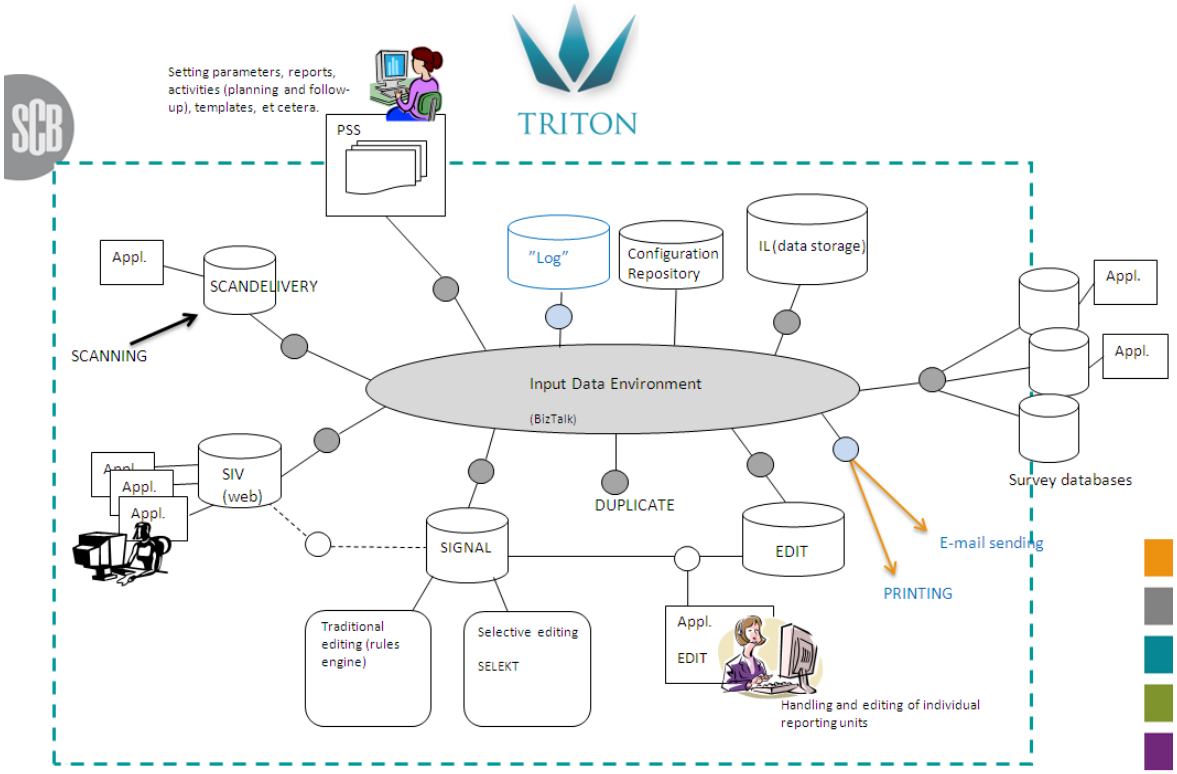
4. La plateforme Triton

Comme nous l'avons montré à la section 3, les domaines de processus intégrés dans le SSP sont le point de départ, le tableau de bord, pour les programmes d'enquête qui se servent de la plateforme Triton. Cependant, cette plateforme est loin de se limiter aux domaines de processus. On pourrait dire qu'elle représente plus qu'un simple système de production ou un certain nombre d'outils, et qu'il s'agit du concept global de travail dans un environnement intégré de production axée sur les processus que l'on peut appeler Triton. Toutefois, pour les employés de Statistics Sweden, qui se soucient davantage d'avoir accès à un système de production fonctionnel que de principes architecturaux ou de grands concepts, il s'agit spécifiquement d'un système de production qui intègre certains outils déjà existants à de nouveaux outils dans un ensemble intelligent.

4.1 Les différentes parties de Triton : un aperçu

L'illustration qui suit montre les diverses parties incluses dans la plateforme et la façon dont elles sont reliées.

Figure 4.1.1
La plateforme Triton : un aperçu



- Le centre (ou cœur) de la plateforme est la plateforme de communication proprement dite, c'est-à-dire l'environnement des données d'entrée (Input Data Environment), qui est fondé sur Microsoft BizTalk. Cette partie gère la circulation de toute l'information et de toutes les données vers les autres parties de la plateforme et en provenance de celles-ci.
- La partie du système de soutien des processus (PSS en anglais) qui figure en haut de l'illustration correspond aux domaines de processus décrits à la section 3.
- L'entrepôt des configurations (Configuration Repository) est la partie où est stockée toute l'information au sujet des règles de transport des données à l'intérieur de la plateforme en fonction des choix effectués.
- La partie IL (stockage des données) est la base de données où les données effectivement collectées sont stockées aussi longtemps qu'elles sont traitées à l'intérieur de la plateforme.
- SIV et SCANDELIVERY sont les systèmes pour les deux outils de collecte reliés à la plateforme à l'heure actuelle, c'est-à-dire la collecte de données en ligne et la lecture optique des questionnaires papier. C'est là que sont définis le questionnaire en ligne et le format électronique du questionnaire papier. Les systèmes sont dotés de bases de données dans lesquelles ils stockent les données qui sont transmises par les répondants. Les données sont transportées de ces bases de données dans la plateforme pour le stockage dans la base de données IL.
- SIGNAL et DUPLICATE sont deux services qui traitent les erreurs signalées et les enregistrements en double. SIGNAL peut utiliser la vérification classique ainsi que la vérification sélective au moyen de l'outil SELEKT.
- EDIT est l'interface où les employés responsables de la collecte des données traitent les unités déclarantes, envoient des rappels et étudient les erreurs possibles.
- Les bases de données d'enquête sont celles où les programmes d'enquête effectuent leurs activités avant et après la collecte et la microvérification des données. Tant que la plateforme ne couvrira que ces deux processus, les bases de données d'enquête resteront nécessaires.

4.2 Travailler dans la plateforme Triton : étape par étape

Voici un bref aperçu de ce que fait un programme d'enquête quand il procède à la collecte et à la microvérification des données en se servant de Triton.

1. Définir l'enquête, ses cycles d'enquête et ses cycles de collecte. Cette tâche est effectuée dans le domaine de processus du SSP et est décrite plus haut à la section 3.
2. Définir les personnes qui travaillent sur l'enquête et leurs rôles. Cette tâche est exécutée dans le domaine de processus du SSP. À l'heure actuelle, Statistics Sweden procède à un examen approfondi du traitement de l'autorisation si bien que cette fonctionnalité est encore très élémentaire.
3. Charger les sous-systèmes. Lorsque le domaine de processus de l'enquête transmet les paramètres établis, les outils pertinents reçoivent l'information qu'un cycle de collecte va avoir lieu et quels outils il conviendra d'utiliser. Les flux de données fondés sur les choix de conception sont mémorisés dans l'entrepôt des configurations.
4. Définir les variables. Le domaine de processus a accès à un outil spécifique pour effectuer cette tâche. Le programme d'enquête définit, décrit et nomme les variables qui doivent être utilisées pour la collecte. Si l'enquête est permanente et qu'il existe déjà un questionnaire en ligne, les variables de ce dernier peuvent être importées dans l'outil de définition des variables et utilisées comme point de départ. L'information est envoyée à la plateforme et mémorisée dans l'entrepôt des configurations.
5. Construire l'instrument de collecte. Cette tâche est exécutée dans le système SIV (pour les questionnaires en ligne) et dans Microsoft Word ou Crystal Reports pour les questionnaires papier. Les questionnaires sont également définis électroniquement dans le système de balayage optique. À cette étape, chaque cellule définie dans l'un ou l'autre de ces systèmes est reliée à une variable définie à l'étape 3. Aucun autre nom de variable que ceux définis ne peut être utilisé.
6. Définir les contrôles. La vérification peut être effectuée à plusieurs étapes du processus de production, tant du côté du répondant, pendant qu'il répond à un questionnaire (applicable aux questionnaires en ligne mais non aux questionnaires papier), que par après, quand les données arrivent à Statistics Sweden. Pour définir les contrôles du côté du répondant, un « créateur de contrôles » était déjà intégré dans le système SIV. Il a été élargi afin que tous les contrôles utilisés dans le processus de vérification classique y soient définis, et un

paramètre spécifique est configuré pour chaque contrôle, afin d'indiquer s'il doit être exécuté du côté du répondant, par après, ou dans les deux cas. Si un programme d'enquête veut utiliser la vérification sélective, tous les paramètres de vérification pour cette dernière sont configurés dans SELEKT. Toute l'information au sujet des contrôles est envoyée à la plateforme et mémorisée dans l'entrepôt des configurations.

7. Définir les ensembles de valeurs. Pour les questions qui comprennent un ensemble fixe d'options de réponse, les options sont définies et communiquées à la plateforme. À l'heure actuelle, cette étape requiert beaucoup de travail manuel (couper et coller dans Microsoft Excel) et nous envisageons dans l'avenir de réutiliser les options de réponse définies dans le questionnaire en ligne (pour les petits ensembles de valeurs) et d'effectuer un raccordement à l'entrepôt central de métadonnées pour les classifications plus importantes.
8. Configurer les paramètres d'EDIT. L'interface d'examen manuel des erreurs possibles et des activités de collecte des données est configurée en fonction des besoins de l'enquête. On se sert pour cela de l'outil EDIT, en utilisant les variables définies à l'étape 3.
9. Charger les renseignements sur l'échantillon et les renseignements contextuels. Le domaine de processus contient une interface qui permet de définir à partir de la base de données d'enquête, où la configuration est faite à l'heure actuelle, quand la collecte doit être effectuée auprès de l'échantillon. Les futurs développements comprendront l'accès direct à la plateforme à partir des outils d'échantillonnage. Les renseignements contextuels (tels que l'information contenue dans les registres, l'information sur les personnes à contacter, les données des cycles précédents, *etc*.) sont également définis et chargés dans la plateforme.
10. Envoyer le questionnaire. Pour le moment, cette tâche est effectuée en dehors de la plateforme, mais dans l'avenir, un lien sera créé avec les fonctions d'impression et d'envoi; la conception graphique de la façon dont cela fonctionnera est faite, mais la fonctionnalité n'est pas encore mise en œuvre dans la plateforme.
11. Transmission des données par les répondants. L'architecture de la plateforme est fondée sur le traitement individuel de chaque unité déclarante, si bien que, chaque fois qu'un répondant envoie des données à Statistics Sweden, la plateforme le constate et prend les données de la collecte en ligne ou de la base de données du balayage optique et les entreposent dans la base de données IL.
12. Traiter les enregistrements en double. La règle quant à la façon de traiter les enregistrements en double est établie par le programme d'enquête dans le formulaire de l'étape 1 et elle est appliquée automatiquement chaque fois qu'un ensemble de données est reçu là où il en existe déjà un. Pour le moment, trois règles simples sont disponibles : choisir le premier, choisir le dernier ou décider manuellement. Une interface graphique pour le choix manuel est prévue, mais n'est pas encore mise en œuvre.
13. Vérifier les données. Si l'enquête comporte une vérification, les données entrantes sont envoyées au service SIGNAL qui exécute les contrôles définis au préalable et présentent un résultat. S'il n'y a aucune erreur, les données sont transférées à la base de données de l'enquête, ce qui signifie que cette dernière est remplie progressivement par les données provenant des unités déclarantes entrantes, à mesure qu'elles sont acceptées à l'étape de la vérification. Si des erreurs éventuelles doivent être traitées manuellement, les données sont envoyées à l'interface EDIT, où les employés de Statistics Sweden peuvent examiner les erreurs possibles. Les unités déclarantes apparaissent aussi progressivement dans EDIT, à mesure que les données sont envoyées et qu'elles sont signalées par SIGNAL comme contenant éventuellement des erreurs.
14. Rassembler les rapports sur l'état d'avancement de l'enquête, le nombre d'unités déclarantes entrantes par strate, etc., dans les rapports de paradonnées et envoyer ceux-ci quotidiennement au domaine de processus. À l'heure actuelle, quatre rapports d'étape prédéfinis ont été préparés et, après le transfert à Sharepoint 2010 en février 2012, cette fonctionnalité sera mise en œuvre. Le nombre de rapports de paradonnées et leur couverture devraient augmenter rapidement au fil du temps.

4.3 L'avenir : prochaines étapes?

La plateforme Triton, y compris les domaines de processus du SSP, a été lancée dans la version décrite aux sections 3 et 4 à la fin de juin 2011. Après l'été, le premier stade de mise en œuvre des enquêtes a démarré. À l'automne 2011, six programmes d'enquête utilisaient la nouvelle version du système en obtenant de bons résultats. Comme d'habitude, il a fallu régler certains problèmes qui n'ont été décelés que quand le système a été utilisé en pratique. Certaines améliorations mineures ont été apportées aux fonctions existantes, mais aucune nouvelle partie importante n'a été ajoutée. En 2012, il est prévu de mettre en œuvre environ 20 enquêtes supplémentaires dans le système. Pour y arriver, des équipes de mise en œuvre particulières seront créées pour aider les programmes d'enquête quand ils commenceront à utiliser le système. Ces équipes seront formées de représentants des services de soutien TI et de soutien de la collecte des données de Statistics Sweden. En outre,

nous avons mis sur pied un groupe de maintenance et créé des routines pour traiter la déclaration des incidents (parties de la plateforme qui ne fonctionnent pas) et les suggestions d'améliorations.

En plus de la maintenance et de la mise en œuvre, il est prévu d'accroître la portée et la profondeur de la plateforme. Cela se fera en plusieurs étapes, mais deux projets ont déjà débuté. Le premier a pour objectif d'améliorer les parties existantes de la plateforme et d'étendre leur fonctionnalité comme il est décrit à la section 4. L'autre portera sur la plateforme proprement dite et les domaines de processus, ainsi que la façon de les étendre pour couvrir d'autres sous-processus du processus de production statistique. Ces deux projets seront exécutés en 2012.

5. Conclusion

Le projet Triton a permis à Statistics Sweden de faire le premier pas vers un environnement de production axé sur les processus. D'importants progrès ont été réalisés grâce au développement d'un outil généralisé de collecte et de microvérification des données, ainsi qu'à l'élaboration du concept des domaines de processus pour la diffusion de l'information et des normes aux programmes d'enquête qui utilisent la plateforme. La plateforme de collecte et de microvérification des données a permis d'intégrer les outils nouveaux et existants en un ensemble cohérent et intelligent. Six programmes d'enquête ont mis en œuvre la version actuelle et 20 autres le feront en 2012. De nombreuses étapes devront encore être accomplies avant qu'une plateforme complète de production axée sur les processus soit établie, et certaines fonctions manquent encore dans la version existante. Toutefois, le projet Triton s'avère très prometteur. D'autres projets en vue d'étendre et d'approfondir le contenu de la plateforme sont déjà à l'étape de la planification.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

La boîte à outils méthodologique standard de Statistics New Zealand

John Lopdell et Gary Dunnet[1]

Résumé

Statistics New Zealand a élaboré récemment un « carnet de route pour la normalisation » qui trace la voie vers une normalisation accrue de nos méthodes, processus, modes de gestion des données et technologies. En réponse à ce carnet de route, nous avons créé une boîte à outils méthodologique standard, qui fournit une liste définitive et validée des outils méthodologiques utilisés dans le processus statistique. La boîte à outils a pour but de réduire les coûts et d'accroître l'efficacité en favorisant l'utilisation plus fréquente d'outils standard dans les diverses activités de collecte, infrastructures et plateformes. On peut s'attendre à ce qu'un ensemble simplifié et complémentaire d'outils réduise le niveau requis de maintenance, de soutien et de formation, tout en augmentant la capacité. Il devrait aussi favoriser le recours à des méthodes correspondant aux meilleures pratiques. Par ailleurs, la boîte à outils renferme de l'information sur la propriété et la version des outils, et offre un mécanisme de gestion du statut de chaque outil (par exemple, nouveau, actuel ou historique), en plus d'être un point de référence centralisé pour la documentation sur les outils.

Cette communication donne un aperçu de la boîte à outils méthodologique standard et de la façon dont elle contribue au modèle des processus opérationnels génériques normalisés de Statistics New Zealand. Elle décrit aussi les travaux à venir en vue de poursuivre le développement de la boîte à outils.

[1]John Lopdell et Gary Dunnet, Statistics New Zealand, Nouvelle Zélande.

SÉANCE 11B

MODÉLISATION ET ESTIMATION

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Estimation semi-paramétrique fondée sur un modèle des composantes de séries chronologiques et l'erreur quadratique moyenne des estimateurs

Michail Sverchkov, Richard Tiller et Danny Pfeffermann[1]

Résumé

Cette communication porte sur l'analyse des séries chronologiques, plus précisément sur l'estimation de la saisonnalité ajustée ainsi qu'aux composantes de la tendance et à l'erreur quadratique moyenne (EQM) des estimateurs. Nous comparons les estimateurs de composantes obtenus en utilisant la méthode X-11 ARIMA avec les estimateurs obtenus en ajustant les modèles d'espaces d'états qui prennent en compte plus directement les erreurs d'échantillonnage corrélées. Les estimateurs de composantes et les estimateurs de l'EQM sont obtenus selon une définition différente des composantes visées. Par cette définition, les composantes inconnues sont définies comme les estimations X-11 en l'absence d'erreurs d'échantillonnage et si les séries chronologiques en questions sont assez longues pour l'application de filtres symétriques intercalés dans cette procédure. Nous proposons de nouveaux estimateurs de l'EQM relativement à cette définition. La performance de ces estimateurs est évaluée au moyen de séries simulées qui évaluent approximativement une vraie série produite par le Bureau of Labor Statistics des États-Unis.

[1]Michail Sverchkov et Richard Tiller, Bureau of Labor Statistics, États Unis ; Danny Pfeffermann, Hebrew University of Jerusalem, Israël, et University of Southampton, Royaume-Uni.

Défis et enjeux de la pondération de l'Enquête sur les voyages des résidents du Canada

Félix Labrecque-Synnott[1]

Résumé

Cet article est axé sur la pondération de l'Enquête sur les voyages des résidents du Canada (EVRC), qui vient d'être remanié. Nous présentons d'abord un aperçu de l'enquête et des concepts à l'origine de la méthode de pondération, puis nous examinons de façon plus approfondie les différents facteurs de correction de la pondération. La plupart de ces facteurs figurent aussi dans les méthodes de pondération utilisées dans de nombreuses enquêtes auprès des ménages à Statistique Canada. L'EVRC comporte des complexités qui lui sont propres, comme, des multiples unités d'analyse (voyages, personnes et voyage-personne), le listage et le sous échantillonnage des voyages, ainsi que le rappel sur deux mois pour les voyages de plus d'une journée.

Mots clés : Pondération ; non-réponse ; enquête-ménage ; propension de réponse ; calage.

1. Introduction

Cet article présente le système de pondération de l'Enquête sur les voyages des résidents du Canada (EVRC). Ce système a été revu à la suite d'un remaniement de l'enquête effectué en 2011. Comme l'indique le titre de l'enquête, une grande importance est accordée aux voyages ; d'ailleurs, les voyages tout comme les répondants doivent être pondérés. Après un survol et une brève mise en contexte de l'enquête, nous décrirons ces deux volets de la pondération en détail. Enfin, nous reprenons les différents défis rencontrés tout au long du processus de pondération, qui découlent, pour la plupart, des particularités de l'EVRC et nécessitent une certaine flexibilité dans le traitement de la pondération.

2. Survol de l'EVRC

2.1 Contexte et aperçu de l'enquête

L'EVRC est un supplément mensuel à l'Enquête sur la population active (EPA) portant sur les voyages et les frais de voyages effectués par les résidents du Canada à l'intérieur du pays. L'EPA, enquête-ménage phare de Statistique Canada, comporte environ 55 000 ménages répondants par mois (Statistique Canada, 1998). Chaque ménage sélectionné est interviewé à six reprises, à raison d'une entrevue par mois pendant six mois. L'échantillon de l'EPA est divisé en six groupes de renouvellement : chaque mois, un groupe de renouvellement quitte l'échantillon et est remplacé par un nouveau. Le plan de sondage de l'EPA est complexe, stratifié et comporte de multiples degrés (Statistique Canada, 1998).

Chaque mois, un adulte est sélectionné au hasard dans chaque ménage dans le groupe de renouvellement qui a tout juste complété la deuxième des six entrevues de l'EPA pour répondre à l'EVRC. Environ 9 100 personnes sont sollicitées chaque mois pour participer à cette enquête.

Comme l'indique le titre de l'enquête, les utilisateurs s'intéressent surtout aux voyageurs, aux voyages et aux dépenses liées aux voyages. Les répondants non voyageurs sont nécessaires afin d'obtenir des taux d'incidence de voyage, mais ils apportent comparativement peu d'information. Malheureusement, ceux-ci forment la majorité de

[1]Félix Labrecque-Synnott, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6 (felix.labrecque-synnott@statcan.gc.ca).

l'échantillon. En effet, bien que la proportion de voyageurs varie d'un mois à l'autre (sans grande surprise, elle atteint son maximum en juillet et en août), elle se situe pratiquement toujours entre 25 % et 35 %.

En outre, historiquement, la taille de l'échantillon de l'EVRC était approximativement deux fois plus importante. Jusqu'en 2009, les ménages étaient sollicités pour participer à l'EVRC après avoir complété la deuxième et la sixième entrevue de l'EPA. Cette coupe échantillonnale a été particulièrement problématique puisque les utilisateurs souhaitent obtenir des estimations des dépenses et des volumes de voyages ainsi que des activités auxquelles prennent part les voyageurs, au niveau des régions touristiques. Il était donc nécessaire de produire des estimations à ce niveau en n'utilisant qu'un sous-ensemble de l'échantillon, dans un contexte de coupe échantillonnale.

Afin de répondre à ce défi et d'augmenter le nombre de voyages et de voyageurs dans l'échantillon, l'enquête a été remaniée ; les données sont recueillies sous la nouvelle version de l'enquête depuis janvier 2011.

2.2 Remaniement de 2011

Le remaniement de 2011 avait deux objectifs principaux : augmenter le nombre de voyages recueillis et limiter la longueur des entrevues. Augmenter le nombre de voyages recueillis permet d'améliorer la précision des estimations relatives aux volumes et aux frais de voyages. De plus, il serait intéressant d'augmenter la proportion des répondants qui déclarent un ou plusieurs voyages afin de rentabiliser le plus possible les efforts requis pour contacter les répondants. Puisque l'EVRC est un supplément de l'EPA, elle donne lieu à une entrevue d'au plus 15 minutes. Il est aussi souhaitable de limiter la durée de l'entrevue afin d'éviter d'alourdir le fardeau de réponse : bien que la majorité des répondants ne voyage pas, et que la majorité des voyageurs ne déclare qu'un seul voyage, une minorité importante voyage fréquemment, et est donc en mesure de nous apporter des données particulièrement intéressantes. Il serait dommage de perdre ces répondants en cours d'entrevue si celle-ci est trop longue.

L'un des principaux changements apportés par le remaniement est l'ajout d'un deuxième mois de rappel pour les voyages de plus d'un jour. Ainsi, lors de la collecte de novembre, on demande aux répondants d'énumérer les voyages de plus d'un jour qu'ils ont fait et qui se sont terminé en octobre ou en septembre, ainsi que les voyages d'un seul jour s'étant terminé en octobre. Ce remaniement permet de recueillir un plus grand nombre de voyages et augmente aussi la proportion d'entrevues « utiles », c'est-à-dire pour lesquelles au moins un voyage est déclaré.

Afin de limiter la longueur de l'entrevue, plutôt que de recueillir des renseignements détaillés au sujet de tous les voyages déclarés, nous demandons maintenant au répondant de faire la liste des voyages ayant pris fin dans les mois de référence et de nous en donner l'information de base (raison principale, principal moyen de transport utilisé, durée du voyage et nombre d'adultes du ménage y ayant participé) à leur sujet. Cette information sera utilisée afin de déterminer si les voyages sont admissibles à l'enquête et, le cas échéant, d'en faire la pondération. Pour chaque répondant, cette liste de voyage sert de base échantillonnale afin de sous-échantillonner un ou plusieurs voyages pour des questions plus détaillées (municipalités visitées lors du voyage, hébergement, détail des dépenses et activités). Ainsi, alors que l'ajout d'un deuxième mois de rappel permet d'augmenter le nombre de voyages recueillis et la proportion de répondants voyageurs, le listage et l'échantillonnage des voyages aident à contrôler la durée de l'entrevue.

Les utilisateurs de l'EVRC reçoivent deux fichiers : un au niveau des personnes, utilisé pour obtenir des taux de voyage et établir le portrait démographique des voyageurs/non voyageurs, et un au niveau des voyages, afin d'obtenir des estimations relatives aux dépenses et activités. Puisque des informations détaillées ne sont pas recueillies pour tous les voyages listés, deux options sont présentement étudiées pour le fichier voyage : pondérer les voyages sélectionnés pour tenir compte du sous-échantillonnage et n'inclure que ceux-ci, ou inclure tous les voyages listés et imputer les détails des voyages non sélectionnés.

3. Pondération

3.1 Survol de la pondération

En raison de la structure de l'EVRC, de multiples ensembles de poids doivent être produits afin d'accompagner les fichiers remis aux utilisateurs. Outre les poids-personne, utilisés pour calculer les taux d'incidence de voyages, deux ensembles de poids doivent accompagner les données au niveau des voyages : les poids-personne-voyages, nécessaires pour estimer les volumes et caractéristiques des voyages, et les poids-ménage-voyages, afin d'estimer les dépenses liées aux voyages.

L'ajout d'un second mois de rappel pour les voyages de plus d'un jour a également un impact important sur le système de pondération. Sous la version remaniée de l'EVRC, les données recueillies lors d'un mois donné correspondent à deux mois de référence. Ainsi, lorsque les répondants sont sollicités en novembre, ils sont sondés sur les voyages ayant pris fin en septembre et en octobre. Réciproquement, chaque mois de référence correspondent deux mois de collecte : afin d'obtenir des estimations relatives au tourisme en septembre, les données recueillies en octobre (1er mois de rappel) et en novembre (2^{e} mois de rappel) seront requises. Cependant, puisque le second mois de rappel ne s'applique qu'aux voyages de plus d'un jour, la définition d'un « voyageur » diffère pour les deux mois de rappel.

Ainsi, deux fichiers au niveau « personne » (et deux ensembles de poids) sont nécessaires. Le premier, contenant uniquement les répondants du premier mois de rappel, sera utilisé pour calculer des taux de voyages « globaux » (y compris les voyages d'un seul jour et les voyages de plus d'un jour). Le second contient les deux mois de rappel et sera utilisé lorsque l'usager ne s'intéresse qu'aux voyages de plus d'un jour. Ces deux jeux de poids sont nécessaires, puisque les poids-voyages pour les voyages d'un seul jour et de plus d'un jour seront fondés sur des poids-personne différents.

3.2 Pondération des personnes

Lors du processus de pondération pour un mois de référence donné, les deux mois de collecte sont d'abord traités indépendamment. Ces deux mois seront donc pondérés l comme décrit dans cette section, et calés afin d'être tous deux représentatifs de la population canadienne au mois de référence. Afin de produire un fichier complet, contenant l'ensemble des répondants interviewés au sujet du mois de référence, ces deux fichiers sont combinés, et les poids-personne sont divisés par deux. Le fichier complet représente donc lui aussi la population canadienne au mois de référence. Le fichier complet et le fichier du premier mois de collecte seront tous les deux utilisés afin de produire des estimations et afin d'établir les poids-voyage.

Le point de départ du système de pondération de l'EVRC, et sa première unité d'analyse, est le répondant : un adulte choisi au hasard parmi le ménage répondant à l'EPA. Le poids de base utilisé par l'EVRC est le sous-poids de l'EPA, tenant compte du plan d'échantillonnage et de la non-réponse de l'EPA. Nous appliquons ensuite des facteurs d'ajustement pour la sélection d'un des six groupes de renouvellement de l'EPA, et pour la sélection aléatoire d'un adulte du ménage.

Un facteur d'ajustement est ensuite appliqué afin de traiter la non-réponse de l'EVRC. À cette fin, la propension de réponse de chaque individu dans l'échantillon est modélisée à l'aide d'un modèle de régression logistique fondé sur des paradonnées et des informations démographiques de l'EPA. Des classes à propension de réponse homogène sont ensuite créées. À l'intérieur de chaque classe, le poids des répondants est ajusté par l'inverse du taux de réponse observé pondéré. Ce type d'ajustement pour la non-réponse est fréquemment mentionné dans la littérature scientifique et est utilisé dans plusieurs enquêtes (Haziza et Beaumont, 2007 ; Little, 1986 ; Eltinge et Yansaneh, 1997).

Une particularité de l'EVRC est la présence de voyageurs non répondants : des individus ayant déclaré au moins un voyage visé par l'enquête, mais dont les données sont incomplètes et ne pouvant donc pas être considérés comme des répondants. Ceux-ci font l'objet d'un ajustement spécifique : leur poids est redistribué parmi les voyageurs répondants à l'intérieur de classes fondées sur l'âge, le sexe et la province de résidence.

Finalement, un calage par groupe d'âge, sexe et région métropolitaine de recensement est appliqué, afin d'assurer la concordance entre les totaux des poids et des estimations démographiques fondées sur le recensement. Encore une fois, il s'agit d'une méthode éprouvée (Deville et Särndal, 1992).

3.3 Pondération des voyages

Afin de pondérer les voyages listés, des facteurs d'ajustement sont appliqués au poids-personne du répondant. Le poids de base est le poids-personne du premier mois de collecte pour les voyages d'un seul jour, et le poids-personne du fichier complet pour les voyages de plus d'un jour. Cette distinction est nécessaire afin d'éviter que les voyages d'un seul jour soient systématiquement sous-représentés dans la pondération finale (et donc sous-estimés dans les analyses). En effet, les répondants ne sont sollicités au sujet des voyages d'un seul jour que dans un des deux mois de collecte. Puisque les poids dans le fichier complet correspondent aux poids des deux mois de collecte divisés par deux, cette distinction permet de corriger le fait que les voyages de plus d'un jour ont l'occasion d'être déclarés par deux fois plus de répondants que les voyages d'un seul jour.

Le premier facteur consiste à multiplier le poids par le nombre de voyages identiques au voyage listé. Lors du listage, pour chaque voyage déclaré, le répondant a la possibilité de spécifier un nombre de voyages identiques pour chaque voyage listé. Ces voyages identiques ne sont pas listés séparément, ce qui permet de limiter la durée de l'entrevue et de limiter le fardeau de réponse pour les répondants voyageant fréquemment (par exemple, ceux qui se rendent à un chalet durant les fins de semaine). La multiplication du poids du voyage représente donc le fait que l'unique voyage listé représente en fait d'autres voyages non listés.

Un autre facteur d'ajustement est appliqué pour corriger des disparités entre le nombre de voyages listés et le nombre de voyages initialement déclarés en début d'entrevue.

Certains voyages seraient normalement admissibles à l'enquête, mais ont des valeurs manquantes pour certaines variables non imputées. Par exemple, un voyage touristique dont la destination principale est à l'intérieur du pays, mais inconnue se situerait dans cette catégorie. Si le répondant a listé au moins un voyage de ce type et aucun voyage valide, il devient un voyageur non répondant. Si un voyage de ce type ainsi qu'au moins un voyage valide sont listés, alors le poids des voyages valides est ajusté pour tenir compte du voyage admissible à données manquantes.

La non-réponse peut également être présente au niveau des voyages : certains voyages sélectionnés pour les questions plus détaillées sont en effet absents des fichiers détaillés. Ces voyages ne sont pas imputés et n'apparaissent pas sur les fichiers remis aux clients, et sont retirés de la liste de voyages. Si tous les voyages sélectionnés d'un répondant sont manquants, ce répondant devient un voyageur non répondant. Sinon, un ajustement est appliqué aux poids des voyages listés (sélectionnés et non sélectionnés) à l'intérieur d'une même classe de non-réponse. Ces classes sont fondées sur les provinces d'origine et de destination des voyages, ainsi que sur leur durée.

À partir des poids des voyages listés, nous pouvons obtenir des poids pour les voyages sélectionnés. Indépendamment de l'option retenue pour les fichiers remis aux clients (uniquement les voyages sélectionnés, ou voyages sélectionnés et voyages imputés), ceux-ci seront nécessaires : dans un cas comme poids accompagnant les voyages sélectionnés ; dans l'autre, afin de valider le système d'imputation.

Ces poids sont obtenus en multipliant les poids listés par un facteur d'ajustement pour le sous-échantillonnage des voyages. Les voyages sont sélectionnés par échantillonnage non équiprobable, et les probabilités de sélection dépendent du nombre de voyages identiques, du statut intra ou interprovincial et de la durée du voyage. Les poids ainsi obtenus sont ensuite calés sur les totaux des poids des voyages listés en utilisant la province de destination, le statut interprovincial et le nombre de nuits comme variables de calage.

Finalement, des poids (ménage-) voyage (listés et sélectionnés) sont obtenus en divisant, pour chaque voyage, les poids-personne-voyage (listés et sélectionnés) par le nombre d'adultes du ménage ayant pris part au voyage. Les poids-personne-voyage sont utilisés pour estimer les volumes et caractéristiques des voyages, alors que les poids (ménage-) voyage sont utilisées pour estimer les dépenses liées aux voyages.

4. Conclusion

En résumé, plusieurs caractéristiques de l'EVRC ont complexifié l'élaboration du système de pondération. Cette enquête comporte différentes unités d'analyse, et nécessite la production de plusieurs ensembles de poids : poids-personne pour le premier mois de collecte, poids-personne pour l'échantillon complet, poids-personne-voyage pour les voyages listés et sélectionnés, et poids (ménage-) voyage pour les voyages listés et sélectionnés. Le tableau 4-1 récapitule l'ensemble des facteurs d'ajustement appliqués lors de la pondération.

Tableau 4-1
Facteurs d'ajustement requis pour produire les poids de l'EVRC

Poids…	facteurs d'ajustement
personne	groupe de renouvellement, adultes du ménage, non-réponse (personnes), non-réponse (voyageurs), calage
personne-voyage (listés)	voyages identiques, nombre originellement déclaré, variables non imputées manquantes, non-réponse (voyages)
personne-voyage (sélectionnés)	probabilité de sélection, calage
voyage	adultes du ménage présents

La non-réponse est présente sous plusieurs formes dans cette enquête : refus et non-contacts, répondants avec une liste de voyage incomplète, voyages sélectionnés manquants et voyageurs non répondants. Ces derniers constituent d'ailleurs un statut de réponse au niveau « personne » déterminé par les données recueillies au niveau « voyage ». Il y a donc une relation entre les différentes unités d'analyse présente dans l'EVRC.

Le deuxième mois de rappel pour certains voyages seulement nécessite de combiner les échantillons et les poids de deux collectes successives et mène à un dédoublement des statuts de réponse – un individu peut être répondant pour un mois de référence et non-répondant pour le suivant.

Afin de surmonter ces défis, des méthodes communes sont mises en œuvre tout au long du processus de pondération. Par exemple, la régression logistique est utilisée afin de modéliser la propension de réponse, la non-réponse est traitée en utilisant des classes de réponse homogène et le calage est utilisé afin d'assurer la cohérence des poids des voyages listés et sélectionnés ainsi que la concordance entre les poids-personne et les estimations démographiques.

Bibliographie

Deville, J.-C. et C.-E., Särndal (1992), « Calibration estimators in survey sampling », *Journal of the American statistical association*, 87, p. 376 à 382.

Eltinge, J.L. et I.S., Yansaneh (1997), « Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey », *Techniques d'enquête*, 23, p. 37 à 45.

Haziza, D. et J.-F., Beaumont (2007), « On the construction of imputation classes in surveys », *International Statistical Review*, 75, p. 25 à 43.

Little, R.J.A. (1986), « Survey nonresponse adjustments for estimates of means », *International Statistical Review*, 54, p. 139-157.

Statistique Canada (1998), « *Méthodologie de l'Enquête sur la population active du Canada* », N° 71-526-X au catalogue.

Recueil du Symposium 2011 de Statistique Canada
Stratégies de normalisation des méthodes et des outils – Comment y arriver

Méthodes statistiques d'évaluation des tendances séculaires en utilisant des estimations provenant d'enquêtes annuelles à échantillons probabilistes complexes transversaux indépendants

Philip J. Smith et Zhen Zhao[1]

Résumé

Lorsque l'on dispose de données provenant d'enquêtes à échantillons probabilistes complexes transversaux réalisées annuellement, on analyse souvent les tendances en se servant de méthodes de régression qui regroupent les données recueillies pour les diverses années de référence des enquêtes. Le cas échéant, on se sert de la pente estimée de la courbe de régression pour résumer la tendance au fil des ans et l'on estime l'erreur type de la pente estimée d'après toutes les observations pour toutes les années d'enquête. Cette méthode omet de tenir compte du fait que i) seules les estimations annuelles sont pertinentes pour estimer la tendance et ii) l'échantillon complet sur les diverses années de l'enquête est sans pertinence pour estimer la précision (c'est-à-dire l'erreur type) de la tendance estimée. En outre, l'échantillon utilisé pour une année particulière de l'enquête ne contribue qu'à l'estimation de la précision des résultats de l'enquête et non à celle de la précision de la pente de la régression. D'autres méthodes s'appuient sur la régression en fonction de l'estimation annuelle des résultats de l'enquête. Ces approches ignorent l'incertitude des estimations calculées d'après les données d'enquête dans l'évaluation de la signification statistique de la pente.

La méthode que nous proposons tient compte de l'incertitude des estimations annuelles des résultats de l'enquête et reconnaît que l'échantillon complet sur les diverses années de l'enquête n'a aucune pertinence pour l'estimation de la précision de la tendance estimée. Notre méthode comprend trois étapes : i) bootstrap de la régression des estimations par répliques bootstrap des résultats de l'enquête annuelle sur les valeurs permutées des diverses années d'enquête pour obtenir la distribution de la pente estimée sous l'hypothèse nulle de l'absence d'une relation entre le résultat de l'enquête et l'année d'enquête ; ii) bootstrap de la régression des estimations par répliques bootstrap sur les valeurs pour les diverses années de l'enquête afin d'obtenir la distribution de la pente sous l'hypothèse alternative d'une relation non nulle entre le résultat et l'année d'enquête, et iii) utilisation de la statistique de Wilcoxon pour tester si les distributions de la pente estimées sous les hypothèses nulle et alternative diffèrent de manière significative. Dans ce cadre, les estimations par répliques bootstrap tiennent compte de l'incertitude dans le résultat estimé de l'enquête, des poids de sondage et du plan d'échantillonnage probabiliste complexe, l'utilisation du bootstrap pour effectuer les tests de permutation génère la distribution de la pente sous les hypothèses nulle et alternative, et la statistique de Wilcoxon fournit une méthode pour déceler les écarts significatifs entre ces deux distributions. Cette méthode peut être étendue facilement en vue de tenir compte de points de renversement multiples de la tendance séculaire sur l'intervalle des années d'enquête.

[1]Philip J. Smith et Zhen Zhao, Centers for Disease Control and Prevention, États-Unis.